Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Integrating Deep Learning and Genetic Approaches to Uncover Molecular Mechanisms of Cellular Organization

Sophia Clara Mädler

aus

Leipzig

2024

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. rer. nat. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Sophia Clara Mädler, München, 13.08.2024

Dissertation eingereicht am	13. August 2024
Erster Gutachter	Prof. Dr. rer. nat. Matthias Mann
Zweiter Gutachter	Prof. Dr. med. Veit Hornung
Mündliche Prüfung am	04. Oktober 2024

Contents

1 Summary

Organisation is a fundamental principle of life. Matter needs to be arranged in space in such a way that it enables reproduction. In biology, this organisation occurs at different scales: from whole ecosystems to multicellular organisms and their tissues, to individual cells and subcellular compartments with defined functions.

The blueprint for the spatial organisation of individual organisms is outlined in their genomes. Understanding these blueprints to define what differentiates individual organisms from one another is a fundamental task in biology. Since an organism's identity is defined by its genome, relating its structural composition directly to its genome provides deeper biological insights into how life is organised. In cell biology, we focus on understanding these relationships from the level of individual cells to tissues.

One biochemical method to analyse the spatial composition of individual cells builds on subcellular fractionation. In this approach cells are split into their distinct compartments, for example by sequential centrifugation steps. The composition of each compartment can then be investigated separately. By coupling cellular fractionation to mass spectrometry (MS), this in principle allows for the unbiased identification of the subcellular localisation of all components in a cell. This technique can further be combined with perturbing a cell's genome to directly link specific genes to their effect on subcellular composition. I demonstrated the strengths of this approach in my characterisation of the molecular mechanisms underlying activation of the immune sensor NLRP3. Using subcellular fractionation coupled to mass spectrometry, we identified the recruitment of NLRP3 to the trans-Golgi network as a key mechanism governing inflammasome activation.

While this approach can generate deep biological insights, it is restricted to a comparatively low number of genes that can be investigated and provides limited spatial resolution. Light microscopy delivers much higher spatial resolution while also allowing for high-throughput analysis of composition and architecture of millions of cells. However, gaining biological insights from microscopy images is not trivial. In recent

years, a new method has emerged from computer vision research that uses machine learning powered by deep neural networks to identify and compress complex patterns into a representative feature space. This approach, called deep learning, shows promise for extracting meaningful biological information from microscopy data.

Another technology that allows for the investigation of spatial composition at the level of tissues is deep visual proteomics (DVP). In DVP, we use microscopy images to identify cells within the larger spatial context of tissues and analyse them further using mass spectrometry. This allows us to collect unbiased information on the molecular composition of these cells while preserving spatial information. By increasing MS sensitivity, we can even break this down to investigate the molecular composition of single cells. Using this approach, I was able to delineate key markers defining hepatocyte zonation in the liver.

Taking it a step further, I used deep learning models to unbiasedly phenotype hepatocytes on the basis of their subcellular distribution of Alpha-1 antitrypsin (AAT) in the fibrogenic liver disease AAT deficiency (AATD), which is characterised by the misfolding and accumulation of AAT . Combining this deep learning-driven phenotyping of cellular morphology with DVP, resulted in the identification of a terminal hepatocyte state marked by globular protein aggregates with a distinct proteomic signature, that holds promise for understanding and ultimately counteracting the molecular mechanisms underlying AATD disease progression.

The above-described approaches are observational, linking distinct cellular compositions assayed using microscopy and MS to their functional implications. However, the high throughput facilitated by modern microscopes allows for the assessment of various aspects of cellular composition over millions of cells, which is compatible with a perturbational approach that looks at the effect of all coding genes on specific subcellular phenotypes.

To enable this type of analysis, I developed spatially resolved CRISPR screening (SPARCS). SPARCS uses automated high-speed laser microdissection to physically isolate phenotypic variants *in situ* for subsequent genotyping. This enables robust, genome-wide, high-throughput screening for spatial cellular phenotypes. Using SPARCS, I was able to identify most known regulators of the cellular process of macroautophagy in a single experiment, and even identified a gene with a previously undescribed cellular phenotype. SPARCS opens up a new paradigm for investigating the genetic basis of subcellular phenotypes that can be applied to a variety of biological contexts.

Finally, to facilitate the types of spatial analysis performed throughout this thesis, I developed a software platform called *scPortrait* that generates single-cell images from raw microscopy data. These single-cell images can be used for deep learning-based cell phenotyping, as demonstrated throughout this thesis, but also for the development of new deep learning models that generate even deeper biological insights. Completely open source and building on available open data formats, *scPortrait* is maximally compatible and provides a framework for the routine implementation of deep learning-based investigation of cellular composition across various areas of biology.

2 Introduction

2.1 From Observational to Perturbational Genetics

2.1.1 Decoding Phenotypic Traits: The Role of Genes in Protein Biosynthesis and Deciphering the Genetic Code

Understanding the fundamental principles governing life has been a topic of profound interest for centuries. Up until the mid-1800s, beliefs about the origins and diversity of life were predominantly influenced by religious and philosophical doctrines; the most prevalent view at the time was that humans were created by a divine entity, a perspective deeply rooted in cultural and religious traditions and buttressed by the belief that the complexity and diversity of life could only be the result of a higher power.

Later in the 19^{th} century, scientists began to challenge these views. In 1859, Charles Darwin published his seminal work *On the Origin of Species*, where he proposed that the diversity and complexity observed across species is the result of an evolution that is given a direction through the process of natural selection (Darwin 1951). According to Darwin, individuals with traits better suited to their environment, i.e. those with a higher fitness, are more likely to survive and reproduce. These advantageous traits are then passed on to subsequent generations, leading to gradual changes - or evolution - in the species over time. This theory was based on the observation of the natural variability of traits observed within populations and how certain traits conferred a survival advantage in specific environments. Darwin's theory provided a framework for understanding the vast variety of observable traits found across species. From the beak variations in finches to the shape of tortoise shells, this diversity results from an adaptation to different environments.

Only seven years after Darwin's publication, Gregor Mendel laid the foundation for understanding how these highly varied traits are passed from one individual to the next. In meticulous experiments in pea plants, he observed that phenotypic traits did not blend, but instead appeared in specific ratios in the resulting offspring. For example, crossing a plant with purple flowers and one with white flowers did not result in a plant with pink flowers, but rather a fixed ratio of offspring with either white or purple flowers. Based on these and other observations, Mendel proposed that parents contribute "invisible factors", discrete units which carry phenotypic information for a particular trait, to their offspring (Mendel 1866). Each individual carries two "factors" for each trait, one from each parent. These factors can either be "dominant" or "recessive", whereby the presence of a "dominant" factor always masks the presence of a "recessive" factor in the resulting phenotype. Furthermore, Mendel stipulated that the inheritance of these traits occurs independently of each other, allowing for the generation of unique combinations of traits in offspring.

Combining Darwin's theories on evolution by means of natural selection and Mendel's theories on inheritance, a new framework emerged in the first half of the 20$^{\text{th}}$ century termed the "modern synthesis", which sought to create a unified theory of evolution and genetics (Fisher 1930; Wright 1931; Haldane 1932; Dobzhansky 1941). According to modern synthesis, the main drivers of evolution are small genetic changes (mutations) that occur within populations. These genetic changes are inherited according to Mendelian principles, where traits are passed from parents to offspring as discrete units. Natural selection then acts on these variations, favouring those that enhance an organism's fitness. Over time, this process leads to the evolution of new species. Today, we understand that the "invisible factors" Mendel described are genes, segments of DNA that encode the instructions for building and maintaining an organism, but in the early 20$^{\text{th}}$ century, the field of genetics was still in its infancy. The concept of a *gene* as the fundamental unit of heredity was first introduced in 1909 by the Danish botanist and geneticist Wilhelm Johannsen (Johannsen 1909). Through the introduction of the term, he sought to distinguish between the physical basis of heredity (genes or genotype) and their observable effects (phenotype). However, at the time it remained unclear how this genetic information was stored so that it could be passed from one individual to the next.

Only a few years after Mendel published his work on inheritance, Friedrich Miescher, who was studying the composition of white blood cells, isolated a new substance from the cell nuclei which he called *nuclein*. He found that *nuclein* was distinct from proteins, and he noted that it was acidic, high in phosphorus, and contained nitrogen

(Miescher 1871). Eventually, this substance was identified as deoxyribonucleic acid (DNA), a molecule consisting of a sugar-phosphate backbone and the four nucleotide bases adenine (A), cytosine (C), guanine (G), and thymine (T) (Kossel 1911; P. Levene and London 1929; P. A. Levene and Jacobs 1909). Despite Miescher's initial tentative suggestion that nuclein could be the sought after carrier of genetic information (Dahm 2008), experimental evidence conclusively identifying DNA as the molecule encoding genes was only obtained much later, in the middle of the 20$^{\text{th}}$ century.

Meanwhile, exciting developments were taking place in the study of chromosomes, thread-like structures that could be observed during cell division. Theories brought forward by Edmund Wilson, Theodor Boveri and Walter Sutton postulated that chromosomes were the carriers of genetic information and that their segregation during cell division aligns with the Mendelian principles of inheritance (Schäfer 1897; Boveri 1902; Sutton 1902; Sutton 1903).

A key contribution to validating these theories was made by American geneticist Nettie Stevens, who, for the first time, showed that the presence of specific chromosomes directly correlates with the manifestation of specific physical traits. Stevens's research focused on the process of spermatogenesis (the production of sperm) in a variety of insect species. During her research she made the key observation that somatic cells from male mealworms had one chromosome that was smaller than the others, while those of somatic cells from females were all the same size (Stevens 1905). Based on this observation, she reasoned that this smaller chromosome determines the sex of the organism. In a second work published a year later, she further showed that during the process of spermatogenesis, the smaller and larger chromosomes, which later became known as the X and Y chromosomes, exhibit Mendelian behaviour in how they segregate during meiosis I and II (Stevens 1906).

Shortly thereafter, Thomas Hunt Morgan published the first experimental evidence that genes reside on chromosomes. After performing random mutagenesis on the fruit fly *Drosophila melanogaster*, Morgan discovered a mutant which had white eyes instead of red ones. After cross-breeding experiments with red-eyed flies, he observed that only male offspring inherited the white-eye trait. Through further experiments he confirmed that this white-eye trait fit the pattern of sex-linked recessive inheritance (Morgan 1910). In subsequent studies, he and his team observed and statistically analysed the patterns of inheritance of multiple different traits (Morgan 1911; Morgan 1919; Bridges and Morgan 1923). Their research revealed a phenomenon known as "genetic linkage", where certain traits appear to be inherited together more often than

would be expected if all traits were inherited independently of one another, suggesting that the genes governing these traits reside in close proximity to one another on the same chromosome.

Together, their work established that certain regions of chromosomes are associated with specific traits. This built a bridge between the abstract concept of heredity and physical structures within cells. Yet, even with these advancements, the precise biochemical nature of the hereditary substance remained a mystery.

This mystery was solved a few years later. In a seminal experiment conducted by Frederick Griffith in 1928, he discovered that non-virulent bacteria could become virulent when exposed to material from dead virulent bacteria, a concept he termed the "transforming principle" (Griffith 1928). Then, in 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty were able to identify the specific substance responsible for this transformation. Using enzymes to selectively degrade specific components of the bacterial lysates, they demonstrated that receiving bacteria were only transformed if the DNA was left intact (Avery et al. 1944). This proved that DNA is the carrier of genetic information, a finding that was later confirmed by Alfred Hershey and Martha Chase through their work with bacteriophages (Hershey and Chase 1952). This constituted a paradigm shift, as proteins, due to their higher complexity and diversity, were previously considered the likely candidates for genetic material.

Even with DNA identified as the carrier of genetic information, central questions remained: How does DNA encode information? And what is the molecular mechanism of its inheritance? In a further pivotal step in unraveling the genetic code, Edwin Chargaff, through meticulous analysis of the chemical constitution of DNA, discovered two key regularities in the base composition of DNA, now known as Chargaff's rules. First, he observed that the amount of adenine (A) always equals the amount of thymine (T), and the amount of guanine (G) always equals the amount of cytosine (C) (Chargaff 1950). This regularity suggested a specific pairing mechanism in the DNA structure. Second, he noted that, while the proportions of the bases vary between species, the ratio of A to T and G to C remains constant across species (Chargaff et al. 1951; Chargaff 1951). This suggested that this common pairing mechanism is universally relevant even across different species.

At the time, it was already understood that biological entities like DNA or proteins exist in three-dimensional shapes, although the structure of DNA remained elusive. In the quest to uncover this unknown structure, Rosalind Franklin, together with

Raymond Gosling, used X-ray crystallography to produce several high-resolution photographs of DNA fibers, in particular the famous "Photo 51" (Franklin and Gosling 1953) in 1953.

Building on Franklin and Goslings' X-ray diffraction data and Chargaff's rules, James Watson and Francis Crick proposed that DNA took on the shape of a double-helix. Their model suggested that DNA is composed of two opposing strands that form a helical shape, with each strand consisting of a sugar-phosphate backbone and paired bases (`A` with `T` and `G` with `C`) interacting with their counterpart through hydrogen bonds (Watson and Crick 1953b). This structure not only explained the regularities in base composition observed by Chargaff, but also suggested a mechanism for DNA replication: the two strands could separate and serve as templates for new complementary strands (Watson and Crick 1953a). Watson and Crick's groundbreaking papers marked the beginning of a new era in molecular biology, as they provided a clear framework for understanding how genetic information is not only stored but also replicated.

However, despite understanding the structure of DNA, scientists were still unsure about two key aspects: how the order of the DNA building blocks could be converted into proteins (the genetic code), and how DNA specifies the chains of amino acids that fold into three dimensional proteins. It was known that proteins are composed of twenty different amino acids, and the challenge was to understand how the linear sequence of four different nucleotides in DNA specified the amino acid sequence in proteins, and, thereby, their function. Moreover, given that in eukaryotes DNA resided in the nucleus and protein biosynthesis occurred outside of the nucleus, it was evident that DNA could not directly serve as a template for protein synthesis, but that there needed to be some type of intermediate substance that could move between cellular compartments.

Through a variety of experiments looking at the infectious potential of viruses (Fraenkel-Conrat and Williams 1955; Gierer and Schramm 1956), another nucleic acid called ribonucleic acid (RNA) became the most likely candidate for this intermediate substance. Like DNA, RNA consists of four nucleotide bases in combination with a sugar-phosphate backbone. In contrast to DNA, RNA contains the nucleotide base uracil (`U`) instead of thymine (`T`) and ribose, a sugar with an additional oxygen atom compared to the deoxyribose contained in DNA.

Building on theoretical postulations by the physicist George Gamow on a four-digit system underlying the genetic code (Gamow 1954; Gamow and Yčas 1955) and under

the assumption that RNA was the missing intermediate substance, Crick, Griffith and Orgel proposed the triplet codon hypothesis in 1957 to explain how a nucleic acid sequence could be translated into an amino acid sequence. This hypothesis suggested that each amino acid in a protein is specified by a set of three adjacent nucleotides, called a codon, in the DNA sequence (Crick et al. 1957). Central to their theory were the assumptions that all possible sequences of amino acids may occur and that at every point the sequence can only be read in one correct way. To accommodate these assumptions, they postulated that the genetic code needed to limit the number of amino acids it could encode. Furthermore, the codons needed to be selected in such a way that, during protein synthesis, the probability of accidentally reading frame-shifted codons was minimised.

The first experimental validation of the role of RNA in carrying genetic information came through experiments conducted by Sydney Brenner, François Jacob, and Matthew Meselson in 1961. Using pulse-chase labelling, they showed that newly synthesized RNA in *Escherichia coli* quickly became associated with ribosomes, serving as a template for protein synthesis (Brenner et al. 1961). This specific type of RNA, whose role it is to carry genetic information from the DNA to the ribosomes in the cytoplasm, is referred to as messenger RNA (mRNA) as proposed by Jacob and Jacques Monod in a second publication in 1961 (F. Jacob and Monod 1961).

Concurrently, Marshall Nirenberg and Heinrich Matthaei provided the first experimental validation for the triplet codon hypothesis. Using a cell-free system they demonstrated that synthetic RNA sequences could direct the synthesis of specific polypeptides (M. W. Nirenberg and Matthaei 1961). Their pioneering work showed that the RNA sequence `UUU` encoded the amino acid phenylalanine, confirming the existence of triplet codons. Through further experiments using the same system, Nirenberg continued to decipher more codons (M. Nirenberg and Leder 1964). In parallel, Har Gobind Khorana developed methods for synthesising defined sequences of repeating RNA (T. M. Jacob and Khorana 1965). By repeating a specific nucleotide sequences and then observing the polypeptides produced, he determined the corresponding amino acids – such as `AAG` specifying lysine (Nishimura et al. 1965; Söll et al. 1965).

By 1966, through the combined effort of Nirenberg, Khorana and other researchers, the codons for all 20 amino acids had been identified, establishing the rules of the genetic code specifying how triplets of nucleotides correspond to specific amino acids, start or stop signals during protein synthesis — thus arguably cracking the code of life (M. Nirenberg, Caskey, et al. 1966; Khorana et al. 1966). By working in a variety of

different organisms they determined how each amino acid was specified and demonstrated that this code of life is universally applicable, constant across organisms and must have existed for billions of years.

This breakthrough, paved the way for a deeper exploration of how genetic information is processed and utilised within cells. In a concept first published in 1958 (Crick 1958) and later expanded on in a second publication in 1970 (Crick 1970), Francis Crick proposed the "central dogma" of biology, in which he described how he imagined the flow of biological information, i.e. how the "the amino acid sequence of the protein" was specified. He envisioned a system in which genes are encoded within the DNA, which is then transcribed into RNA, which in turn is translated into proteins (Crick 1970). In his dogma the flow of information is unidirectional: once information has been transformed into an polypeptide chain it can no longer reverse to become a nucleic acid again (Crick 1958). This "central dogma" provided a framework to understand how genetic information directs the synthesis of proteins, thereby linking the genetic material to a functional output in living organisms. As James Watson, co-discoverer of the double helical structure of DNA, once noted, "We used to think our fate was in the stars. Now we know, in large measure, our fate is in our genes".

2.1.2 Phenotypes are a Manifestation of a Multitude of Elements Working in Concert

The identification of DNA as the carrier of genetic information and the subsequent deciphering of the genetic code, marked a critical step in furthering our understanding of the molecular mechanisms underpinning life. It became clear, that the biological processes defining an organism are fundamentally encoded within its DNA. In essence, DNA can be regarded as a sophisticated source code, providing a comprehensive blueprint required for the development of an organism. If DNA is the source code, then proteins should be considered the "molecular workhorses" of the cell. Proteins make up the majority of a cell's dry mass (Alberts et al. 2008). They serve not just as the fundamental units that construct a cell, but also as the key players that carry out nearly all cellular operations.

In the simplest view, proteins consist of a linear sequence of amino acids, a polypeptide chain, that directly corresponds to the DNA sequence. Upon synthesis, they undergo intricate three-dimensional structural changes as they assemble into their final three-dimensional shape or structural conformation (Anfinsen 1973). This process is driven

by non-covalent interactions between the different amino acids in the polypeptide chain (Alberts et al. 2008). These interactions are a consequence of the different chemical properties of the amino acids contained in the chain: polarity, charge, hydrophobicity, and the ability to form hydrogen or disulphide bonds, all fundamentally influence a proteins' final conformation and thus its biological function.

The same weak bonds that allow a polypeptide chain to fold, also allow proteins to bind to one another to form larger structures in the cell. In fact, most biochemical processes underlying cellular functions are orchestrated by large complexes of proteins working in concert. Recent data suggest that a protein in yeast cells may interact with at least 16 different partners on average (Michaelis et al. 2023).

As an added layer of complexity, proteins also frequently undergo post-translational modifications (PTMs) after their initial synthesis. These modifications entail the covalent addition of functional groups like phosphate, acetyl- or ubiquitin groups to a specific amino acid side chain, altering the chemical composition of the protein (Doll and Burlingame 2015). PTMs can significantly affect both a protein's conformation as well as its activity, localisation or its interactions with other molecules (Alberts et al. 2008).

Moreover, protein expression is highly regulated and subject to multiple layers of control. This regulation determines when and where a protein is synthesised, it's localisation and it's stability and eventual degradation (Preissler and Deuerling 2012; Dikic 2016). While proteins constitute the main executors of all biological processes, only about 3% of the human genome codes for proteins (Dunham et al. 2012). Some other parts of the genome contain instructions for non-protein coding RNAs like ribosomal (rRNA) that have a catalytic or structural function (Zhang et al. 2022), or species like long-non-coding RNAs (lncRNAs), which appear to predominantly have a regulatory function, for example during cell differentiation or development (Mattick et al. 2023). But, the largest part of the genome is not transcribed at all, instead fulfillling a regulatory role, for example by shaping the three-dimensional structure of the genome (Rowley and Corces 2018) or by modulating the binding of proteins regulating RNA transcription (Spitz and Furlong 2012). In this way, while individual gene products are the functional units that execute the processes of life, their creation, composition and interaction is tightly regulated through additional DNA sequences. If "phenotype" is defined as a set of observable characteristics or traits, then, in the context of individual cells, this would encompass all cellular components, as well as their spatial orientation. Considering what we now know about the regulation of these

individual components, it is clear that not a single gene or DNA sequence, but instead a multitude of sequences working in concert is the genetic basis for almost any given cellular phenotype.

2.1.3 Perturbational Genetics to Map Genes to their Biological Function

Since groundbreaking biological discoveries, like Stevens's identification of the Y chromosome and its role in sex determination, are built on observational studies, they can initially only generate correlative insights and not determine causation. In contrast, deliberately altering – or "perturbing" – the genomic sequence, also allows for the exploration of causative relations by enabling comparative experiments between perturbed and unchanged wild type organisms. This paradigm is called "perturbational genetics" and encompasses two fundamentally different approaches: reverse and forward genetics. In reverse genetics, we start with a known alteration in a gene sequence, and seek to determine the phenotypic effects resulting from this alteration. For example, we can disrupt or "knock out" a specific gene, and, by observing what happens to the organisms' phenotype, infer the gene's function. In contrast, in forward genetics, we start with a phenotype and then identify the genetic basis underlying that phenotype. For example, we can randomly induce mutations in a number of organisms and then select those individuals displaying a desired trait. By finding the location in the genome that has been mutated, one can then identify the region driving the chosen phenotype. Both techniques represent powerful tools to map genes to their biological function.

2.2 Key Techniques of Molecular Genetics: from sequencing to programmable gene editing

2.2.1 Nucleotide Sequencing

To better understand how a DNA sequence relates to molecular function, it is important to be able to efficiently read the order of nucleotides within a DNA molecule or "sequence" it.

The chain-termination method, introduced by Frederick Sanger and his team in 1977, marked a significant advancement in this area. Their method uses modified nucleotides called dideoxynucleotides (ddNTPs) that lack the 3' hydroxyl group found in regular nucleotides, preventing the addition of further nucleotides during DNA strand elongation. Consequently, DNA strand elongation terminates during synthesis whenever ddNTPs are incorporated. By performing four separate reactions, each incorporating a different radioactively labeled ddNTP (ddATP, ddTTP, ddCTP, or ddGTP), a mixture of DNA fragments of varying lengths is produced. Analysing the lengths of these fragments through gel electrophoresis allows the reconstruction of the original DNA sequence; Each fragment ends in the ddNTP used in that specific reaction, with the fragment length corresponding to the position of that nucleotide (Sanger et al. 1977).

Later advancements, generally referred to as sequencing by synthesis (SBS), greatly improved sequencing efficiency and throughput. Key innovations included transitioning from a four-reaction process to a single-reaction process by employing fluorescently labeled dideoxynucleotides instead of radioactive ones, which streamlined the detection process (Prober et al. 1987), and shifting from post-separation detection of fragment lengths to real-time detection, which further enhanced the speed and accuracy of sequencing (Hunkapiller et al. 1991; Luckey et al. 1990).

These improvements enabled large-scale sequencing projects, beginning with simpler model organisms like *Caenorhabditis elegans* (Waterston and Sulston 1995), *Saccharaomyces cerevisiae* (Goffeau et al. 1996) and moving to more complex organisms like *Drosophila melanogaster* (Adams et al. 2000) or *Mus musculus* (Chinwalla et al. 2002), and ultimately culminating in the completion of the human genome sequence in 2001 (Consortium et al. 2001; Venter et al. 2001; Nurk et al. 2022). Ongoing efforts aim to expand this to genomes of over 70 000 vertebrate species by 2030 (Rhie et al. 2021).

As the technologies improved, and specifically with the advent of next generation sequencing (NGS), which allows for massively parallel sequencing of millions of DNA fragments simultaneously, sequencing costs per base have exponentially decreased (Mardis 2011). While sequencing a whole human genome in 2008 cost approximately 1 million dollars (Wheeler et al. 2008), generating this same information in 2014 cost less than 1000 dollars (Hayden 2014) and has further decreased since.

Recently, novel sequencing technologies that focus on generating long reads have emerged (Kasianowicz et al. 1996; Jain, Olsen, et al. 2016). With the ability to

generate reads of tens to thousands of kilobases in length, these platforms have significantly improved the quality of genome assemblies by reducing gaps and ambiguities (Jain, Koren, et al. 2018), as well as allowed for the detection of structural variants, large-scale alterations in the genome that involve significant rearrangements of DNA (Amarasinghe et al. 2020).

The ability to sequence nucleotides has facilitated the emergence of centralised efforts, like ENCODE, to annotate all bases with their biological function (Dunham et al. 2012) across a multitude of species.

While a central application of nucleotide sequencing is to understand genomic DNA sequences, it can also be applied to the sequencing of RNA, thereby providing a snapshot of the expressed genes within cells at a given time. Sequencing RNA through SBS involves generating a cDNA template through a reverse transcription reaction (Baltimore 1970; Temin and Mizutani 1970). Novel long-read technologies like Nanopore now also facilitate the direct reading of RNA molecules (Jain, Abu-Shumays, et al. 2022; Workman et al. 2019).

2.2.2 Polymerase Chain Reaction

Another critical development was the invention of the polymerase chain reaction (PCR) by Kary Mullis in 1983. PCR allows for the targeted amplification and modification of specific DNA sequences, with minimal knowledge of the targeted sequence itself (Mullis and Faloona 1987).

A PCR consists of repeated cycles of DNA double strand denaturation at a high temperature, the annealing of short complementary oligonucleotides ("primers") flanking the targeted DNA sequence to the template DNA, and the elongation of these primers by a DNA polymerase enzyme. Because primers and building blocks for new DNA strands (dNTPs) are provided in a large excess, PCR exponentially amplifies the targeted sequence by a factor of 2^n, with n indicating the number of cycles. The targeted sequence can be modified by including sequences at the 5' end of the primers, which will be fused to the ends of the PCR product by the incorporation of the primers. In this capacity, PCR serves as the basis for most single-cell transcriptomic technologies as well as NGS methods.

2.2.3 Molecular Cloning

The invention of PCR along with the discovery of additional molecular tools such as restriction enzymes, which cut DNA at specific sequences called restriction sites, (Arber 1978) and DNA ligases, which join DNA fragments (Lehman 1974), enabled scientists to combine arbitrary DNA sequences together. The process of assembling these blocks and subsequently directing their replication within a host organism is referred to as molecular cloning. By using defined DNA constructs to transfer genetic information to new organisms, called a vector, DNA sequences assembled via cloning can be subsequently expressed in a compatible host organism. This technology can be used to investigate the function of naturally existing or newly created genes in controlled or biologically interesting contexts, or to drive the production of "recombinant" proteins.

For example, by deliberately employing a mismatched primer during a PCR reaction, a point mutation in a gene of interest can be introduced. In a subsequent PCR reaction, compatible restriction sites can be brought into the DNA sequence using a primer with overhanging 5' ends. This mutated gene PCR product can then be cloned into a plasmid, a short, circular piece of DNA, for example, by restriction and ligation. This plasmid can then be amplified in bacteria such as *Escherichia coli* before being used to express the mutated gene in a relevant model organism to study the effect of the mutation. Indeed, plasmids conferred the virulence trait observed by Griffith in his experiments on the nature of the genetic material (Griffith 1928).

This is only one example of the manifold applications possible using molecular cloning, coupled with protein expression, in perturbational genetics. While this is a very powerful strategy, synthetically introducing a gene construct into an organism generally does not accurately reflect physiological conditions. For instance, proteins might not undergo the same post-translational modifications as their native counterparts or might be shuttled to different cellular compartments, both of which could significantly alter their function. Furthermore, synthetically expressed constructs do not follow the same regulatory mechanisms as the native gene, which makes it difficult to study dynamic processes or look at temporal regulation.

2.2.4 Precise Genome Editing

Another strategy to manipulate an organism's DNA is through genome editing. Here, an organism's genomic DNA is directly modified in a defined manner. For example, the function of individual genes can be disrupted ("knock-out") or short DNA fragments can be inserted at random or defined positions ("knock-in"). This allows for the targeted perturbation of genetic sequences under physiological conditions. Modern genome editing technologies rely on endonucleases to create double-stranded breaks (DSBs) at specific locations in the genome. These breaks are then repaired by the cell's own machinery, either through the error-prone non-homologous end joining (NHEJ) (J. K. Moore and Haber 1996; Chang et al. 2017) or mostly error-free homologous recombination also called homology-directed repair (HDR) (Liang et al. 1998) pathways. NHEJ can occur throughout the cell cycle, whereas HDR is mainly limited to the S/G2 phases (Hendrickson 1997; Saleh-Gohari and Helleday 2004). If the repair pathway introduces a mutation in the targeted locus, the endonuclease no longer recognises its target site on the genomic DNA, completing the editing process. Otherwise, the endonuclease cuts the DNA again, and the process restarts.

If the resulting mutation is not an insertion or deletion of a number of nucleotides that is a multiple of three this disrupts the reading frame of the encoded protein, introducing a so-called frameshift mutation. This shift results in a completely altered protein sequence and, given that three of 64 possible codons encode a translation stop, likely the premature termination of the polypeptide sequence. These types of prematurely terminated mRNAs, where the stop codon is found in an unusual context, are usually degraded through specific cellular mechanisms such as nonsense-mediated mRNA decay (Lykke-Andersen and Jensen 2015), resulting in a "knock-out". If, during the gene-editing process, a repair sequence is provided which is homologous to the targeted locus, the HDR pathway can use this template to accurately repair the break (J. Y. Wang and Doudna 2023), allowing for precise insertion of the desired genetic sequence. While this approach usually has a lower efficiency than the generation of knock-outs, it can be used to introduce specific mutations, correct genetic defects, or insert new genes at targeted locations in the genome. Zinc finger nucleases (ZFNs) were among the first tools developed for precise genome editing. ZFNs are engineered nucleases created by combining a custom zinc finger DNA-binding domain with an non-specific DNA cleavage domain from the FokI restriction enzyme (Y. G. Kim et al. 1996). The DNA-binding domain consists of several zinc finger (ZF) motifs, each recognising a specific 3-base pair DNA sequence. Typically, ZFNs have three to six

ZFs, allowing them to target a 9 to 18-base pair sequence. For the FokI nuclease to be able to cut double-stranded DNA, it needs to dimerise, thus to create a DSB, a pair of ZFNs need to be designed that flank the desired locus (Miller, Holmes, et al. 2007). Transcription activator-like effector nucleases (TALENs) were developed more recently, and work similarly to ZFNs, but use different DNA-binding domains. In TALENs, the transcription activator-like effector (TALE) DNA-binding domain is fused to the same FokI cleavage domain (Miller, S. Tan, et al. 2011). A TALE domain consists of repeated units of 33-34 amino acids where the 12^{th} and 13^{th} positions determine the nucleotide recognised by that repeat. By linking several of these repeats together, the TALE domain can be designed to match the targeted DNA sequence (R. Moore et al. 2014). Compared to ZFNs, TALENs offer a modular design, which makes their application more straightforward and flexible, but they nevertheless requires that a specific set of proteins needs to be engineered for each genetic locus that is to be targeted. This is a time-consuming process. When a system called clustered regularly interspaced short palindromic repeats (CRISPR) was discovered, which facilitates RNA-guided precise genome editing, the field of genome engineering underwent a revolution. To target a new genetic locus using CRISPR, only a single, comparatively short sgRNA molecule needed to be provided instead of a new engineered protein.

Originally described in bacteria, CRISPR is a bacterial defence system against bacteria-targeting viruses called bacteriophages (Barrangou et al. 2007; Brouns et al. 2008; Gasiunas et al. 2012; Jinek et al. 2012). After becoming infected by a phage, bacteria capture short sequences of DNA from the invading phage and integrate them into a region called the CRISPR array in their own genome. The integrated sequences serve as a genetic memory of the infection. Upon reinfection, the CRISPR array is transcribed and processed into individual CRISPR RNAs (crRNAs). These crRNAs assemble with trans-activating crRNA (tracrRNA) and CRISPR associated (Cas) proteins to form an RNA-protein complex, which binds to the complementary DNA sequence in the phage genome. Through the nuclease activity of the Cas protein a DSB is then introduced to incapacitate the phage.

A crucial component of the CRISPR/Cas system is the Protospacer Adjacent Motif (PAM). The PAM is a short, conserved DNA sequence on the target DNA that is essential for the Cas protein to bind (Wiedenheft et al. 2012). Sequence specificity in CRISPR/Cas systems is therefore generated through the combination of two factors: complementarity to the spacer sequence and the presence of a PAM sequence directly adjacent to the spacer sequence. The requirement for a PAM ensures, that in bacteria, the CRISPR/Cas system as part of the prokaryotic immune system, only targets

foreign DNA and not the host's own genome, as the PAM sequence is not present in the CRISPR array itself.

Building on this system, scientists were able to design synthetic single guide RNAs (sgRNA) that allow the targeted introduction of a DSB at chosen genomic loci with a suitable PAM sequence (Ran et al. 2013). The application of the CRISPR/Cas system to genome editing has dramatically simplified the generation of genetic knockouts. In recognition of this breakthrough development, Emmanuelle Charpentier and Jennifer Doudna were awarded the Nobel Prize in Chemistry for their development of the CRISPR/Cas gene editing technology in 2020 (*Press release: The Nobel Prize in Chemistry 2020 - NobelPrize.org* 2024), less than 10 years after its discovery.

2.3 Proteomics as a Tool for the Unbiased Investigation of Cellular Composition

Proteomics is the large-scale study of protein abundances, modifications, localisations and interactions. As the final product in the gene expression cascade, proteins are not just the fundamental units that build a cell, but also the key players that execute nearly all cellular processes. Because proteins carry out the actual functions of the cell, they most directly determine cellular phenotype and function. Therefore, understanding the complete proteome, the entire set of proteins expressed by a cell, tissue, or organism, is crucial for comprehensively and mechanistically understanding biology at a molecular level.

2.3.1 Basic Principles of Mass Spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio (m/z) of charged particles (ions). It works by ionising chemical compounds to generate charged molecules or molecule fragments, which are then separated based on their m/z ratio using electric or magnetic fields (Sinha and Mann 2020).

This principle goes back to the English physicist, J.J. Thomson, who, while investigating the nature of cathode rays (streams of electrons observed in vacuum tubes), observed that these rays were deflected by electric and magnetic fields (Thomson 1897). The degree of their deflection depended on their m/z. As a result, ions with

different m/z will either travel at different speeds or follow different flight paths in the same electromagnetic field (Thomson 1921). This fundamental principle is used by MS to separate incoming ions according to their m/z. By measuring the intensity of the separated ions, a mass spectrum can be generated, which can be used to infer the composition and structure of the measured sample (Sinha and Mann 2020).

While all mass analysers rely on the same fundamental principles of how ions behave in electric and magnetic fields, the specific method through which they deduce an ion's m/z depends on the build of the mass analyser.

The earliest type of mass analyser, the magnetic sector, uses a magnetic field to bend the path of ions according to their momentum (Aston 1920; Thomson 1921). The degree of bending (deflection) of the ion's path in the magnetic field is directly related to its m/z ratio given the same prior acceleration. By precisely controlling the magnetic field strength, ions with specific m/z ratios can be directed to the detector.

Following the magnetic sector, the quadrupole mass analyser and the ion trap were developed (Paul and Steinwedel 1953). In a quadrupole, ions are passed through four parallel rods with alternating radiofrequency (RF) and direct current (DC) voltages. Only ions with a specific m/z have a stable trajectory and can pass through the quadrupole at a given time. By varying the RF and DC voltages, ions of different m/z ratios are selectively filtered and detected, allowing for the sequential scanning of a range of m/z values. In an ion trap, ions are confined within a three-dimensional quadrupole field generated by a ring electrode and two endcap electrodes with applied RF and DC voltages. Only ions with a specific m/z maintain a stable trajectory within the trap. By varying the RF voltage, ions of different m/z values become unstable and are sequentially ejected toward the detector, enabling the scanning of a mass spectrum.

Around the same time as quadrupole analysers, time-of-flight (TOF) mass analysers were introduced (Wolff and Stephens 1953). They function by accelerating ions with the same kinetic energy and allowing them to drift through a field-free region (Ligon 1979). Since all ions were accelerated with the same kinetic energy, but the velocity of an ion is dependent on its weight, lighter ions (with a lower m/z) travel faster than heavier ions (with a higher m/z) (Sinha and Mann 2020). The time it takes for ions to reach the detector (time of flight) is measured with great accuracy (ps) and used to calculate the m/z ratio.

Finally, the Orbitrap mass analyser represents a more recent advance in MS (Makarov 2000). The Orbitrap traps ions in an electrostatic field, where they orbit around a central electrode with a circular and axial component. The axial frequency of these oscillations only depends on the m/z ratio, and this frequency is detected, with the m/z ratio being directly proportional to the square root of the oscillation frequency. A Fourier transform converts these frequencies into a mass spectrum, providing high-resolution measurements of m/z.

In modern mass spectrometry (MS) devices, multiple mass analysers are often used in sequence to accurately quantify and identify even complex analytes. For instance, it is common for the initial mass analyser to be a quadrupole, which is then followed by a collision cell and a TOF or Orbitrap analyser. This configuration lends itself to a process known as tandem mass spectrometry (MS/MS). In MS/MS, ions generated in the initial mass spectrometry step, known as precursor ions, are selectively isolated and then fragmented within the collision cell generating so called fragment ions, which undergo a second round of mass analysis. The mass patterns of these fragment ions offer additional structural information, allowing for more precise identification of compounds and detailed determination of molecular structures. Often, signal intensities at the MS^1 level are used for quantification, while m/z spectra at the MS^2 level are used for identification (Sinha and Mann 2020).

In MS/MS devices, a central step is how precursor ions are selected for subsequent analysis. Until recently, this was ubiquitously performed through data-dependent acquisition (DDA) schemes, where the user defines a set of rules based on which the mass spectrometer selected as many precursor ions as possible for further characterisation (Sinha and Mann 2020). Recently, an alternative approach termed data-independent acquisition (DIA) has emerged and is set to become the new standard (Venable et al. 2004; Gillet et al. 2012). In DIA methods, the mass spectrometer continuously cycles across the entire mass range to select precursor ions in predefined fragmentation windows. This generates unbiased insights into the composition of the precursor ions, with much higher data completeness than DDA methods. However, it also generates much more complex spectra which require advanced analysis techniques to map back to individual proteins (Ludwig et al. 2018; Wallmann et al. 2024).

2.3.2 Application to Proteins

The application of mass spectrometry (MS) to analyse the entire protein composition of samples, known as proteomics, presents unique challenges due to the complex nature of proteins. Proteins exhibit a wide range of sizes, structures, and modifications, making their analysis far more intricate compared to small molecules or oligonucleotides.

Furthermore, in proteomics, it is not only crucial to identify the molecular weight of peptides derived from a sample, but also to gain insight into their molecular structures. For example, two peptides with identical molecular weights but different sequences can have completely distinct biological roles. Therefore, simply measuring the mass is not sufficient; the exact peptide sequence must be determined to accurately identify a protein. For peptides, the MS/MS step provides vital information on the amino acid sequence (Steen and Mann 2004). This complexity is further compounded by the presence of post-translational modifications (PTMs), which can significantly alter a protein's structure and function.

A technique to help resolve these types of complex mixtures is liquid chromatography coupled mass spectrometry (LC-MS), which first emerged in the late 1970s and early 1980s. Liquid chromatography (LC) is a technique that separates the components of a mixture based on their differential interactions with a stationary phase and a mobile phase, allowing for the isolation and analysis of individual compounds (Mondello et al. 2023). Combining LC with MS, allows for the separation of complex mixtures of analytes (like those in biological samples) followed by their precise identification and quantification through MS or MS/MS (Arnott et al. 1993; Eng et al. 1994).

A breakthrough in the applicability of this approach was achieved with the development of Electrospray Ionization (ESI) in the late 1980s (Whitehouse et al. 1985; Fenn et al. 1989; Mann et al. 1989). A discovery for which John Fenn was awarded the Nobel Prize in Chemistry in 2002 (*Press release: The Nobel Prize in Chemistry 2002 - NobelPrize.org* 2024).

ESI allows for the conversion of liquid-phase analyses (like those emerging from LC) directly into gas-phase ions (Whitehouse et al. 1985). This is achieved by introducing the sample, which is dissolved in a liquid solvent (often a mixture of water, organic solvents like methanol or acetonitrile, and a small amount of acid to enhance ionisation) into the ESI source through a narrow metal capillary or needle. A high voltage is applied to the metal capillary, creating a strong electric field at the tip. When the liquid emerges from the capillary, this electric field causes it to form a fine mist of charged

droplets, which are carried toward the mass spectrometer inlet by a combination of the electric field and a stream of heated gas (nebuliser or drying gas), often nitrogen. As the droplets travel, the solvent rapidly evaporates, causing them to shrink and the charge density on their surface to increase, until the repulsive forces between the like charges becomes greater than the surface tension of the droplet. At this point, the droplet undergoes a "Coulomb explosion", where it breaks apart into smaller droplets and free ions. This process can repeat multiple times, further reducing the size of the droplets, and eventually producing individual gas-phase ions from the analyte molecules. These gas-phase ions are then directed into the mass spectrometer through an inlet orifice, where they are analysed based on their m/z ratio. Through ESI, the analyte molecules typically acquire multiple charges resulting in multiply charged ions. This is advantageous because it effectively reduces the m/z ratio, allowing even large biomolecules to be detected within the mass range of most mass spectrometers (Fenn et al. 1989), but it also requires specialised software to extract molecular mass information from the resulting spectra (Mann et al. 1989).

In principle, this approach can be applied to intact proteins or to protein fragments called peptides as already alluded to above. So called top-down approaches analyse intact proteins directly, in principle providing information about the entire protein molecule, including its post-translational modifications and sequence variations. In contrast, bottom-up approaches analyse peptide mixtures, which are generated through the enzymatic digestion of complete proteins. While top-down approaches offer a more comprehensive view of protein structure and function, they can be challenging due to the larger size and complexity of intact proteins and limited sensitivity resulting from poor ionisation of full-length proteins (Tran et al. 2011). Much more common, is the application of bottom up proteomics (Nesvizhskii and Aebersold 2005), which provides a high degree of sensitivity and throughput, while retaining a high level of proteome coverage (Aebersold and Mann 2016).

Since the advent of MS based proteomics, many further advances both in LC performance as well as in MS sensitivity (Bian, Bayer, et al. 2021; Bian, Zheng, et al. 2020; Brunner et al. 2022; Thielert et al. 2023; Stewart et al. 2023) have greatly enhanced our ability to accurately identify and quantify the proteins composing a sample. In tissues and cell lines, 10 000 proteins can now be routinely quantified in a single LC-MS run (Meier et al. 2018), even with sample runtimes of only 30 min (Guzman et al. 2024). This has also facilitated the practical applicability of MS to the analysis of single-cells (Brunner et al. 2022). Using novel multiplexing strategies 3×60 samples

with a median depth of 2370 proteins per cell and per day have recently been analysed in our group (Thielert et al. 2023)

2.4 Forward Genetic Screens as a Tool for Uncovering Biology

Forward genetic screening is a powerful technique to map cellular phenotypes to their underlying genetic basis. By generating a pool of mutagenised individuals and selecting and genotyping those individuals which show a particular phenotype of interest, all of the genes relevant for a specific biological process can, in principle, be identified in an unbiased manner.

The process through which Thomas Hunt Morgan was able to map specific traits to regions on the chromosomes of *Drosophila melanogaster* was one of the earliest applications of a genetic screen. While he did not yet have the tools available to determine the underlying gene sequence, the general screening principles have remained the same. Since then, forward genetic screens have been used to understand a wide range of biological processes.

One of the first seminal genetic screens was performed by Leland Hartwell in 1970. After inducing random mutations in yeast cells, he identified temperature-sensitive mutants with replication defects. Because yeast cells exhibit distinct morphological changes at different stages of the cell cycle, Hartwell was able to determine the specific stage where each of the mutants arrested using microscopy. This allowed him to arrange them according to their time-resolved relevance during the cell cycle. Using this approach, he was able to identify key regulators of cell cycle progression (Hartwell et al. 1970), a breakthrough for which he was awarded the Nobel Prize in physiology and medicine in 2001 (*The Nobel Prize in Physiology or Medicine 2001 - NobelPrize.org* 2024).

Another pivotal forward genetic screen was carried out by Sydney Brenner in 1974 using the roundworm *Caenorhabditis elegans*. By employing random mutagenesis, Brenner generated a pool of approximately 300 mutants with altered morphology (Brenner 1974). Since *Caenorhabditis elegans* has a relatively simple body plan with a fixed number of cells and well-defined developmental stages, Brenner could easily identify mutants with aberrant developmental patterns or morphology. Through this

approach, he characterised the function of around 100 genes, establishing *Caenorhabditis elegans* as a vital model system for genetic studies and identifying key genes involved in its development.

In the 1980s, H. Robert Horvitz and his colleagues conducted another influential screen in *Caenorhabditis elegans*. Similar to Brenner and building on his work, they made use of the fact that *Caenorhabditis elegans* has a clearly defined body plan. *Caenorhabditis elegans* is one of the few organisms that has an invariant cell lineage, meaning that the exact number and position of cells in an adult worm are consistent across individuals. This allowed the scientists to readily identify mutants affecting regulated cell death pathways, as they showed an abnormal accumulation of cells that normally would die and disappear. Using this approach, he was able to characterise the process of apoptosis and identify regulators underlying this programmed cell death pathway (Ellis and Horvitz 1986).

Brenner and Horwitz both received the Nobel Prize for their discoveries in "genetic regulation and programmed cell death" in 2002 (*The Nobel Prize in Physiology or Medicine 2002 - NobelPrize.org* 2024).

Finally, another pivotal forward genetic screen was performed in 1980 by Christiane Nüsslein-Volhard and Eric Wieschaus. Using *Drosophila melanogaster* as a model system, they performed a large-scale mutagenesis screen to identify genes relevant to control segment formation during development (Nüsslein-Volhard and Wieschaus 1980). *Drosophila melanogaster* has a well-documented embryonic development process, including the formation of distinct segmentation patterns. This ensured that mutations affecting this process were readily visible under the microscope. This screen led to the discovery of many key developmental regulators such as "toll" or "knüppel", and again was awarded with a Nobel prize in physiology and medicine in 1995 (*The Nobel Prize in Physiology or Medicine 1995 - NobelPrize.org* 2024).

A commonality of all of these landmark screens is that a model organism was chosen in which the phenotype of interest was readily visible. Indeed, this marks a core principle of a successful screen: individuals with a positive phenotype need to be easily distinguishable from those without.

The advent of modern genome-editing technologies with the ability to disrupt both alleles of a given genetic locus, such as CRISPR/Cas, allows us to miniaturise genetic screens to the level of individual human cells. We can now generate large mutant

libraries, where each cell carries a different genetic knockout whereby all protein-coding genes in the genome can be targeted at once. While this significantly increases throughput, it also requires phenotypes to be read out at the single-cell level. As in the previously described screens, how interesting phenotypes or "hits" are identified and isolated is of critical importance. Considering the number of protein coding genes in the genome, these libraries typically consist of millions of cells whose phenotype needs to be assessed. This has largely limited cell-based genome-wide screens to one of three types of easily selectable phenotypes: a difference in proliferation rate (Shalem et al. 2014), an inhibition of cell death (T. Wang et al. 2014), or a change in fluorescence intensity compatible with fluorescence-activated cell sorting (FACS) (Parnas et al. 2015).

An example of an important discovery made by these modern pooled forward genetic screens is the protein Gasdermin D (Shi et al. 2015). This protein controls a strongly pro-inflammatory type of cell death called "pyroptosis". When Gasdermin D is activated by cleavage, its N-terminal portion translocates to the plasma membrane, where it forms large pores that lead to water influx. Eventually the plasma membrane ruptures and the cell bursts open, spreading its cytosolic contents. This stands in contrast to apoptosis, which is a cell death pathway during which no cytosolic content leaks. The discovery of Gasdermin D via a cell-based screen was corroborated by a concurrent forward genetic screen in mice (Kayagaki et al. 2015).

2.5 Light Microscopy as a Technique to Assay Cellular Composition

Light microscopy uses visible light focused through a series of lenses to magnify small objects, thereby enabling the exploration of the spatial composition of cells down to subcellular resolution. Especially following the invention of digital cameras, light microscopy has been automated to a high degree, making it possible to assay millions of cells at high throughput. However, translating the resulting images into quantifiable metrics that facilitate conclusions about biological phenotypes is not straightforward.

In principle, images can be considered as arrays of numbers, that represent the intensity of light at different regions – pixels, in case of digital image acquisition – of the imaged area. But, unlike most other datatypes, the values of individual pixels only obtain

their relevance through their neighbours. In contrast, in a gene expression dataset describing a cell, the order in which we consider the genes, which are labelled with an interpretable identifier, i.e. the gene name, is irrelevant and has no effect on the represented biology. In an image this is not the case. If we randomly permute the order of the pixels in an image, the information contained in the image will change. Furthermore, cellular images do not have an inherent orientation. If we take an image of a cell and rotate it by 180 degrees, then the cellular phenotype observed will not have changed, i.e. the image still captures the same underlying biology, but if we are directly comparing the matrix representation of these images, they will have almost no overlap. Thus, to be able to extract meaningful information from these images, one must be able to condense the extremely high dimensional pixel space into meaningful features that capture the underlying biology of interest.

In recent years, the field of computer vision has emerged which, focuses on obtaining detailed information and insights from digital images and videos through the development of algorithms and models. Coupled with advances in computer hardware this goal has recently become a reality.

2.5.1 A Brief History of Microscopy

Remarkably, light microscopy has already been used for over 300 years. One of the early pioneers of microscopy was Antonie van Leeuwenhoek, who developed new techniques for the crafting of microscopy lenses which allowed for clearer, more detailed observations at higher degrees of magnification (Zuylen 1981). His work led to the first descriptions of various microorganisms, including bacteria, protozoa, and spermatozoa (Leeuwenhoek 1997), and provided some of the earliest insights into human tissues, blood cells, and reproductive biology (Robertson 2023).

Since van Leeuwenhoek's time, microscopy has undergone tremendous and continuing advancements, particularly with the development of high-resolution techniques. The introduction of Charge-Coupled Devices (CCDs) in the middle of the 20^{th} century was a major breakthrough, enabling the conversion of optical images into digital signals (Boyle and Smith 1970). This innovation facilitated high-resolution, reproducible, and quantifiable imaging, allowing for real-time visualisation, simplified data storage and automated acquisition.

Additionally, the discovery and application of fluorescent proteins, such as Green Fluorescent Protein (GFP) (Shimomura et al. 1962), and fluorescently labeled antibodies have revolutionised the cell microscopy subfield by enabling researchers to tag and visualise specific proteins on and within cells (Miyawaki 2011).

Further developments in microscope architecture have greatly enhanced imaging capabilities, offering unprecedented detail and clarity in cellular analysis. Notable examples are confocal microscopy, which uses a pinhole to block out out-of-focus light and allow for the acquisition of high-resolution images also in the z-dimension (Marvin 1961; Minsky 1988), and super resolution techniques like STED (Stimulated Emission Depletion) (Hell and Wichmann 1994) and PALM (Photo-Activated Localisation Microscopy) (Betzig et al. 2006), that circumvent the diffraction limit of light microscopy, achieving nanometer-scale resolution.

2.5.2 A Brief History of Computer Vision

In 1943, Warren McCulloch and Walter Pitts introduced the first computational model of a neuron that mimicked the functionality of animal brains (McCulloch and Pitts 1943). Their model consisted of an input layer which receives multiple binary inputs, a logical layer which applies weights to these inputs and sums them up and an output layer which fires if the sums exceed a specific threshold.

While this model, also called perceptron, provided the starting point for the development of artificial neural networks, its application was limited to linear problems. Linear problems are classification tasks where the data can be separated or classified using a straight line (in two dimensions) or a hyperplane (in higher dimensions). Furthermore, the model did not include any type of learning mechanism, which meant that weights and threshold needed to be set manually.

A notable improvement on this model is Rosenblatt's perceptron, which also operates as a single-layer perceptron, but unlike McCulloch and Pitts' neuron, incorporated a learning algorithm to adjust weights to minimise classification errors based on previous results(Rosenblatt 1957; Rosenblatt 1958). This meant, that the weights could be iteratively adjusted through successively passed inputs, minimising the difference between desired and actual output. While still limited to linear problems, Rosenblatt's perceptron demonstrated the practical application of 'learning' in neural networks (Rosenblatt 1960).

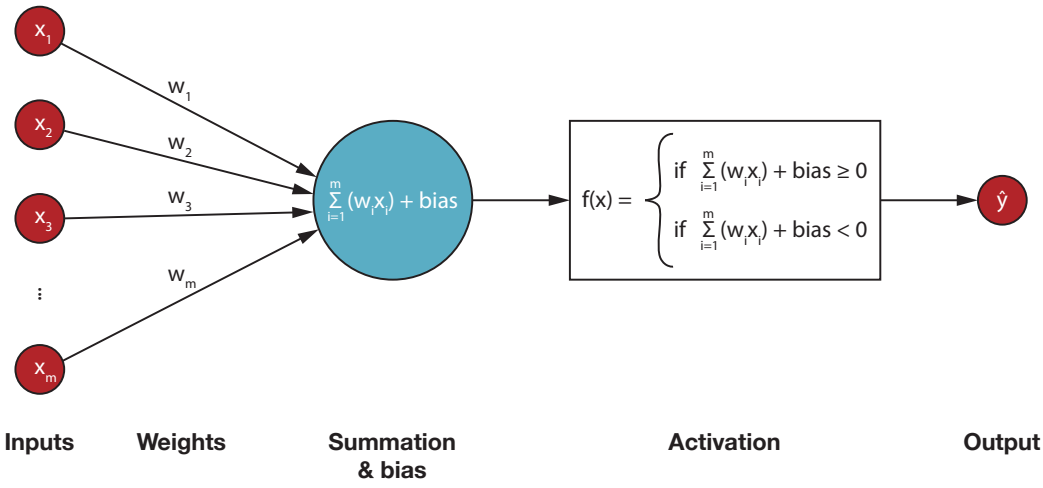


Figure 2.1 | Schematic Overview of a Perceptron. The basic structure of a neural network unit, or perceptron. Each input node has a weight associated with it, and the perceptron sums the weighted inputs and passes the result through an activation function, typically a step function, to produce the output.

The first implementation of a neural network which was able to solve a nonlinear classification problem was the multi-layer Perceptron (MLP). Nonlinear problems require more complex functions to separate the input space, which can no longer be described by a single straight line (Fig 2.2). In contrast to a single-layer perceptron, an MLP consists of multiple layers of nodes, with at least one input layer, one or more hidden layers with non-linear activation functions and a final output layer (Rosenblatt 1961). This concept presented a fundamental advance in the design of neural networks, which allowed for the approximation of complex functions and decision boundaries.

Early MLPs struggled with effective learning due to a lack of efficient algorithms to update weights in the model during training. The introduction of backpropagation by Geoffrey Hinton, David Rumelhart, and Ronald Williams in the mid-1980s made the efficient training of multilayer neural networks feasible (Rumelhart et al. 1986). Backpropagation involves computing the gradient of the loss function, the metric used to quantify the difference between the predicted values and the actual values during training, with respect to each weight by applying the chain rule of calculus. This gradient is then used to adjust the weights in the network to minimise the loss function. Through backpropagation the training of deep neural networks finally became feasible, and the field of neural networks experienced a resurgence of interest. By increasing the number of layers, or the depth of the network, neural networks can learn more intricate relationships between inputs and outputs, making it possible to model more

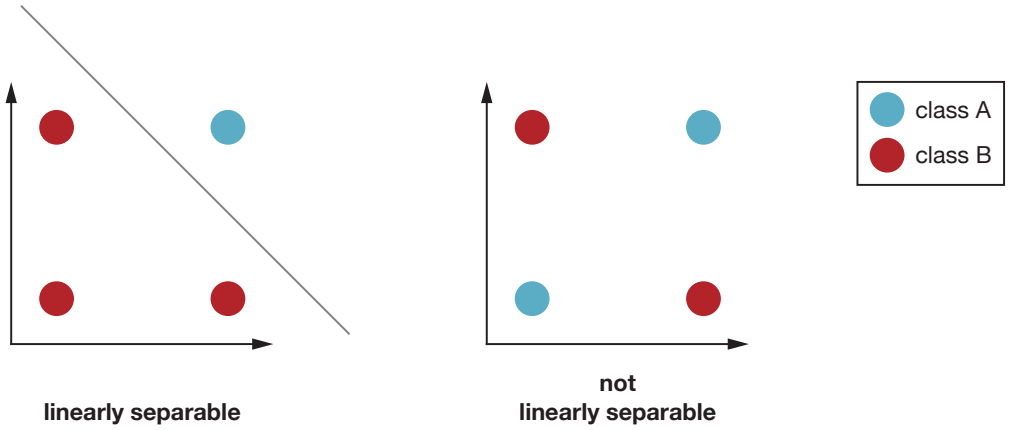


Figure 2.2 | A linear vs a nonlinear Problem. In linear problems the two classes can be separated by a single straight line. In nonlinear problems this is not possible.

complex data distributions. The training paradigm established by Hinton, Rumelhart, and Williams remains the basis for training deep neural networks to this day.

The next fundamental advance in neural networks was the introduction of convolutional kernels, so-called convolutional neural networks (CNN). CNNs combine two special types of layers: Convolutional layers with trainable kernel weights learn spatial hierarchies of features from input data. Pooling layers then reduce the dimensionality of their input data by summarising features in local regions.

In simple terms, a convolution is akin to placing a small window (the filter) over an image, multiplying the numbers under the window by the values in the filter, and summing them up to obtain a single number that goes into a new image (the feature map) (see Figure 2.3a). This process is repeated as the window slides across the entire image, helping the network identify patterns like edges, textures, or shapes.

By convolving the original image with the convolutional kernels, and applying a nonlinear activation function new feature mappings are obtained and each feature mapping can be used as a class of extracted image features. To extract higher-level and more complete feature representations, multiple convolution layers can be stacked within the network model.

The resulting features are then compressed by the pooling layer, which in essence divides the feature map into many distinct regions and aggregates or "pools" all values in that region together (see Figure 2.3b). This not only effectively compresses the

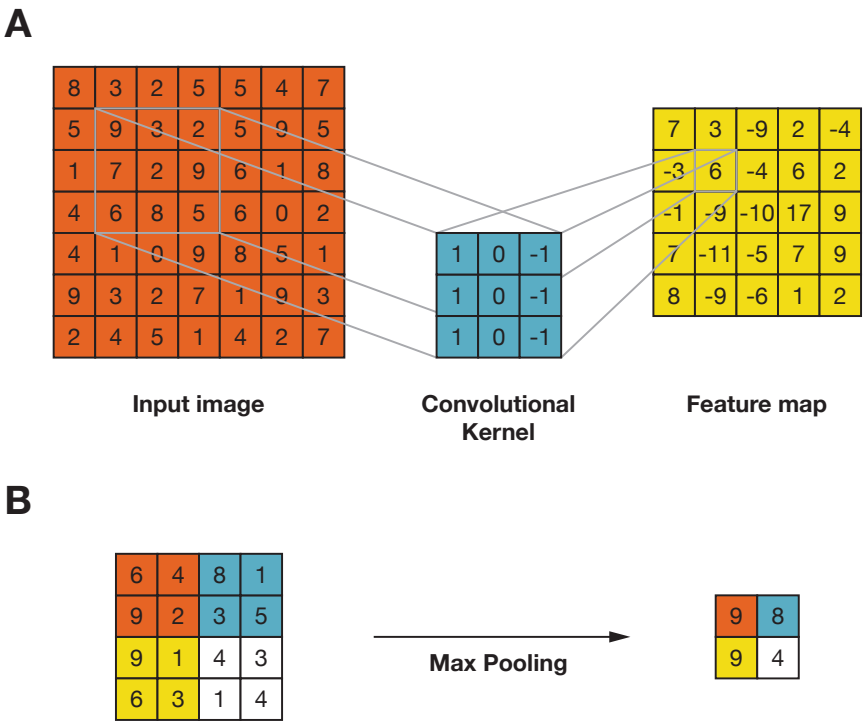


Figure 2.3 | Schematic Representation of a Convolution and a Max Pooling Operation. (**A**) A convolution is a mathematical operation where a kernel slides over an input image and is multiplied with the input values. The products are summed at each position to produce a feature map. (**B**) Max pooling is a mathematical operation where the input image is divided into regions and the values from each region are aggregated together by taking the maximum value.

amount of data and parameters, but also makes the network invariant to some small local morphological changes while retaining a larger perceptual field.

With this combination, CNNs are much better suited to generating meaningful low-dimensional features from input images than MLPs, even for larger input images.

The original concept of CNNs was first introduced by Kunihiko Fukushima in 1980 with his *Neocognitron* model, which was designed to recognise visual patterns using a hierarchical, multi-layered approach Fukushima.1980, but CNNs only gained practical applications through work by Yann LeCun and colleagues much later. Through their work on *LeNet-5*, a CNN designed for digit recognition tasks, they demonstrated the effectiveness of CNNs for image recognition tasks for the first time (Lecun et al. 1998).

In the 2010s, significant advancements in Graphics Processing Units (GPUs) dramatically accelerated the training of neural networks (Sun et al. 2019). GPUs, originally mostly applied in the gaming industry, are specialised hardware designed for parallel processing where they handle multiple computations simultaneously. This makes them ideal for the matrix operations that form the basis of deep learning (C. Song et al. 2024). With drastic speed improvements over previous hardware, GPUs enabled the efficient training of deeper and more complex neural networks.

The next iteration of CNNs - and revolution in deep learning - was marked by the introduction of AlexNet by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton in 2012. AlexNet introduced several innovations, including the Rectified Linear Unit (ReLU) activation function, which speeds up training, dropout for regularisation to prevent overfitting, and the concept of data augmentation to enhance model generalisation (Krizhevsky et al. 2012):

1. ReLU is an activation function that replaces negative values with zero and linearly scales positive values. This non-saturating behaviour reduces the likelihood of vanishing gradients, a problem that occurs when the gradients of the loss function become very small with respect to the model weights, so that weights are updated very slowly or not at all during training. Vanishing gradients are a problem more common in deeper networks, and preventing their occurrence is key for the training of deeper, more performant networks.

2. Dropout regularisation is a technique where, during training, random neurons are “dropped out” or deactivated in each training iteration. This forces the

network to learn more robust features and prevent overfitting because it cannot rely on the information from individual neurons. This method not only prevents overfitting, but also makes models more resilient to variations in training data.

3. The performance of a neural network critically depends on the amount of input data available for training. More extensive and diverse datasets typically enable models to learn more robust features and generalise better to new, unseen examples. The process of data augmentation, where additional training samples are synthetically generated by applying transformations such as rotations, translations or scaling to the original training dataset, increase variability in the training data, helping models to generalise better to new unseen data without requiring the acquisition of additional training data.

The development of *AlexNet* marked a pivotal moment in computer vision by achieving a substantial improvement in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. ImageNet is a visual database used for training and evaluating image classification algorithms. *AlexNet*'s performance was a breakthrough, as it reduced the top-5 error rate in withheld test data by more than 10 percentage points compared to the previous best results (Krizhevsky et al. 2012), showcasing a dramatic leap in accuracy and setting a new standard for performance in image recognition tasks.

Since the development of *AlexNet*, a variety of further CNNs have been developed, each improving upon previous models and achieving notable performance increases (see Table 2.1).

These advancements in computer vision and CNN architectures have continually pushed the boundaries of image recognition, contributing to significant improvements in various applications such as medical imaging (Esteva et al. 2017), autonomous driving (X. Chen et al. 2016; J. G. Song and Lee 2023) and more.

2.5.3 The Modern Era of Computer Vision

The field of computer vision has continued to evolve, moving beyond traditional CNNs to embrace new architectures and training paradigms. Two significant developments in this modern era are the introduction of transformers and the rise of unsupervised learning. These advancements have not only improved the accuracy and efficiency

Table 2.1 | Overview of CNNs. Summary of different CNN models performance on ImageNet classification task. Top1 Accuracy values taken from (Feng et al. 2019) except when indicated with a *.

Model	Year	Key Features	Top1 Acc	Params	Reference
AlexNet	2012	Introduced ReLU activation, dropout, data augmentation	57.2%	60M	(Krizhevsky et al. 2012)
VGGNet	2014	Deep architecture with 16 or 19 layers, small 3×3 filters	71.5%	138M	(Simonyan and Zisserman 2014)
GoogLeNet	2014	Inception module for efficiency, fewer parameters	69.8%	6.8M	(Szegedy et al. 2014)
ResNet	2015	Residual connections to enable very deep networks	78.6%	55M	(He et al. 2015)
DenseNet	2016	Dense connections between layers for improved feature reuse	79.2%	25.6M	(Huang et al. 2016)
EfficientNet	2019	Compound scaling for balancing network depth, width, resolution	79.8%*	9.2M	(M. Tan and Le 2019)
ConvNeXT	2022	Refined convolutional architecture with improved performance and efficiency	83.1%*	50M	(Liu et al. 2022)

of computer vision models, but also expanded their applicability to a wide range of complex problems, including biological research.

Transformers

A transformer is a deep learning model that leverages a mechanism called "self-attention" to effectively capture long-range dependencies in data (Vaswani et al. 2017). Self-attention enables the model to determine which parts of the input data are most important for making predictions. It works by creating three different vectors (query, key, and value) for each input element. The model then calculates how much attention each element should pay to every other element by comparing their query and key vectors, resulting in a set of attention scores. These scores are normalised into weights, which are used to combine the value vectors of the input elements (Lin et al. 2021). In the so called feed forward layers, self-attention outputs are processed to generate more refined representations of the input data, which are then passed through additional layers to produce the final prediction results. The name "transformer" reflects the model's ability to transform input data into meaningful representations through these layers. This allows the model to dynamically focus on relevant parts of the input, and capture relationships and dependencies within the data.

Transformers, while originally developed for natural language processing (NLP) tasks

(Vaswani et al. 2017), have since also been adapted to image recognition tasks. Unlike CNNs, which process images using convolutional layers, Vision Transformers (ViTs), divide images into patches and treat them as sequences, similar to words in a sentence, allowing the model to learn spatial relationships and patterns across the entire image (Dosovitskiy, Beyer, et al. 2020). Vision Transformers offer advantages over traditional CNNs, including improved performance and the ability to capture global context in images more effectively. However, they require very large datasets for initial training, as they lack some of the inductive biases inherent to CNNs, such as translational equivariance (that shifting input results in a corresponding shift in the output) and locality (the ability to focus on local areas of the input data when making decisions). This leads to poor generalisation when insufficient training data is used (Dosovitskiy, Beyer, et al. 2020). But, much like NLP Transformers, ViTs achieve excellent performance when pre-trained on large datasets with a subsequent fine-tuning on a specific task with fewer datapoints.

Self-supervised Learning

A key principle in the training of deep neural networks is the use of a loss function to quantify the difference between the predicted and actual values, which then guides the adjustment of the model's weights accordingly. However, generating labeled datasets is time-consuming and limits the amount of available input data for training, which in turn restricts model performance. As a result, self-supervised learning, where models are trained without a priori labeled data, has gained increasing interest in computer vision. In this approach, the model uses the data itself to generate labels or training signals, thereby circumventing the need for manually labeled datasets.

Self-supervised learning in computer vision typically involves one of two main paradigms: reconstruction tasks and contrastive tasks. In reconstruction tasks, the model learns to predict missing parts of the data based on the remaining parts, effectively generating useful features from the data itself (Caron et al. 2021). Contrastive tasks, on the other hand, involve learning representations by distinguishing between similar and dissimilar examples, which helps the model to capture meaningful patterns and structures in the data (Dosovitskiy, Fischer, et al. 2014; T. Chen et al. 2020). One technique for generating similar examples involves using data augmentation techniques to create different views of the same data instance, such as by applying transformations like rotations or cropping. Dissimilar examples are often generated by sampling from

distinct, unrelated data instances, ensuring a broad range of contrasts for effective learning.

2.5.4 Applications of Deep Learning to Biological Problems

The recent advancements in deep learning have found manifold application in the analysis of complex biological data and have provided new solutions to challenges in various biological domains, including protein structure prediction (Jumper et al. 2021; Krishna et al. 2024), genomic analysis (Poplin et al. 2018; Theodoris et al. 2023; Cui et al. 2024) or mass spectrometry (Zeng et al. 2022; Wallmann et al. 2024).

One of the most prominent applications is in protein structure prediction. *AlphaFold*, developed by DeepMind, revolutionised this field with a breakthrough performance that demonstrated a significant leap in accuracy over previous methods (Jumper et al. 2021).

In the realm of cellular imaging, deep learning has dramatically transformed the process of image segmentation. While previous approaches mainly relied on thresholding and watershed algorithms, segmentation via deep learning models has become the new state-of-the-art. These approaches mainly utilise the U-Net architecture, which consists of an encoder-decoder structure with skip connections to capture both high-level and fine-grained details (Ronneberger et al. 2015). Variant implementations such as *DeepCell* (Valen et al. 2016) or *CellPose* (Stringer et al. 2021), have further improved cellular segmentation and tracking of cells in live-cell imaging.

Furthermore, deep learning has found application in the generation of image representations on the basis of microscopy images to predict cellular function. A recent example is *scDINO*, an adapted ViT model trained on multichannel microscopy data, which showed good performance in classifying immune cells in peripheral blood (Pfaendler et al. 2023). In another application, scientists used a large perturbational microscopy dataset containing cells treated with different chemical compounds, to train a variety of self-supervised models to predict drug targets and gene family classification (V. Kim et al. 2024). While so far no breakthrough performance akin to that of *AlphaFold* has been achieved, these types of applications are promising. Through improved training paradigms in combination with better training datasets, as well as the integration with other data modalities, these types of models could, in the future, be used to create "foundation models" that would be able to decode cellular function.

3 Aims of the Thesis

In this thesis, I aimed to find a way to investigate phenotypes that capture the spatial arrangement of cellular components in genetic screens. The developed technology should provide sufficient throughput to scale to genome-wide applications in human cells, while retaining a high resolution to also permit the capture of subtle phenotypes. This type of technology would allow me to directly link specific genes to their subcellular phenotypes.

Furthermore, I wanted to develop a tool that could identify these subtle phenotypes robustly at the multi-million cell scale I was aiming at. In a first proof of concept, I aimed to use supervised deep learning methods to characterise previously defined phenotypes. Due to the current successes with deep learning, I later aimed to expand this to include the unbiased identification of novel phenotypes.

Finally, the application of deep-learning techniques to microscopy images is partially hindered by the lack of standardised frameworks for representing image-based phenotypes at a single-cell level and making single-cell images available for deep learning applications. Vast resources of microscopy images are publicly available, that as of now, are not yet being fully utilised due to an absence of easy-to-use, open, standardised computational tools.

To bridge this gap, I aimed to develop a computational framework that not only provides end-to-end processing of raw microscopy images into single-cell image datasets, which is compatible with deep-learning based cellular phenotyping, but also establishes a standardised data format for saving these images, which interfaces with current state-of-the art deep learning platforms like PyTorch. Through the development of such a framework, I hope to facilitate the generation of better computer vision models, as well as to facilitate the training of multimodal models that are fully able to decode cellular function.

4 Publications

4.1 IKKβ primes inflammasome formation by recruiting NLRP3 to the trans-Golgi network

The protein NLRP3 is a sensor of the innate immune system that indirectly detects pathogen-derived molecules as well as cellular damage. Upon activation, it assembles a large protein complex called the inflammasome which induces a strongly pro-inflammatory form of cell death known as pyroptosis.

NLRP3 has been implicated in mediating the inflammatory component of diseases like gout, atherosclerosis, Alzheimer's disease as well as several infectious diseases, but its molecular activation mechanism remained unclear. It was known that various triggers have the capacity to activate NLRP3 inflammasome formation, but the specific signal that is recognised by NLRP3 was unknown. In addition, a variety of signalling pathways have been described that modulate responsiveness of NLRP3 to these signals - for example through PTMs. In the field of NLRP3 research this activity modulation is referred to as "priming", describing an increased capacity of NLRP3 to become activated in response to an activating stimulus without already forming an inflammasome. Our understanding of what differentiates primed from non-primed NLRP3 on a molecular level remains limited.

In this publication, we identified the recruitment of NLRP3 to the trans-Golgi network as a priming modality of the NLRP3 inflammasome. Using cellular fractionation coupled with mass-spectrometry, we demonstrated that in response to a priming stimulus, NLRP3 shifted to a cellular fraction strongly enriched for proteins of the trans-Golgi network but not the cis-Golgi network. This observation corroborated findings obtained via microscopy imaging, showing that primed NLRP3 co-localised with phosphatidylinositol-4-phosphate (PI4P), a membrane phospholipid strongly associated with the trans-Golgi network. Together, our findings shed light on the rules

governing NLRP3 inflammasome formation by uncovering a new priming modality of the sensor protein NLRP3 .

The following research article was originally published here:

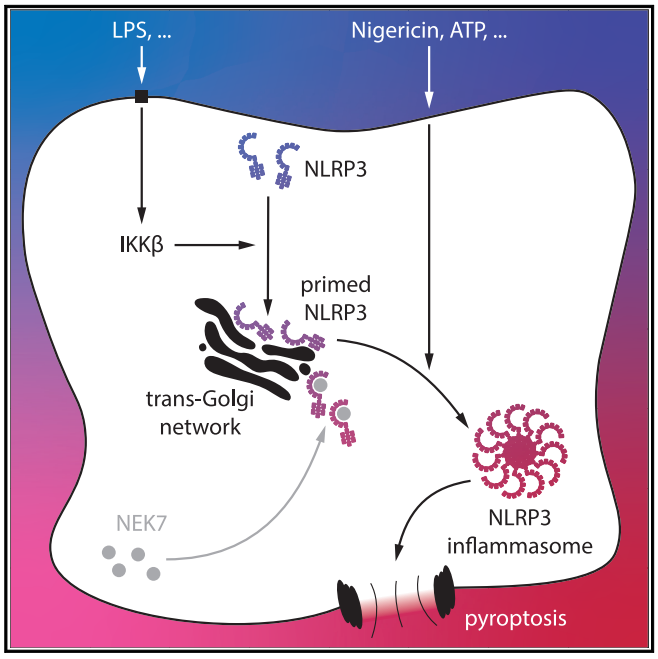
Schmacke, N. A., O'Duill, F., et al. (2022). "IKKβ primes inflammasome formation by recruiting NLRP3 to the trans-Golgi network". In: *Immunity* 55.12, 2271–2284.e7. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2022.10.021

Article

Immunity

IKKβ primes inflammasome formation by recruiting NLRP3 to the *trans*-Golgi network

Graphical abstract



Authors

Niklas A. Schmacke, Fionan O'Duill, Moritz M. Gaidt, ..., Matthias Mann, Heinrich Leonhardt, Veit Hornung

Correspondence

hornung@genzentrum.lmu.de

In brief

The NLRP3 inflammasome causes lytic cell death and inflammation, but its activation mechanism remains enigmatic. Schmacke et al. now show that the kinase IKKβ provides an essential priming signal for inflammasome formation by recruiting NLRP3 to the *trans*-Golgi network. Human cells predominantly use IKKβ instead of the priming factor NEK7.

Highlights

- iPSC-derived human macrophages form an NLRP3 inflammasome independently of NEK7
- NLRP3 priming by IKKβ proceeds independently of transcription and substitutes NEK7
- IKKβ activity constitutes the predominant NLRP3 priming mechanism in human cells
- IKKβ drives NEK7-independent priming by recruiting NLRP3 to the phospholipid PI4P

Schmacke et al., 2022, Immunity *55*, 2271–2284
December 13, 2022
https://doi.org/10.1016/j.immuni.2022.10.021

CellPress

Immunity

CellPress

Article

IKKβ primes inflammasome formation by recruiting NLRP3 to the *trans*-Golgi network

Niklas A. Schmacke,[1] Fionan O'Duill,[1] Moritz M. Gaidt,[1,7] Inga Szymanska,[1] Julia M. Kamper,[1] Jonathan L. Schmid-Burgk,[1,8] Sophia C. Mädler,[2] Timur Mackens-Kiani,[1] Tatsuya Kozaki,[3] Dhruv Chauhan,[1] Dennis Nagl,[1] Che A. Stafford,[1] Hartmann Harz,[4] Adrian L. Fröhlich,[1] Francesca Pinci,[1] Florent Ginhoux,[3,5,6] Roland Beckmann,[1] Matthias Mann,[2] Heinrich Leonhardt,[4] and Veit Hornung[1,9,*]

[1]Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, 81377 Munich, Germany
[2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[3]Singapore Immunology Network (SIgN), Agency for Science, Technology & Research (A*STAR), 8A Biomedical Grove, Immunos Building #3-4, Biopolis, Singapore 138648, Singapore
[4]Faculty of Biology, Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany
[5]Shanghai Institute of Immunology, Shanghai JiaoTong University School of Medicine, 280 South Chongqing Road, Shanghai 200025, China
[6]Translational Immunology Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore 169856, Singapore
[7]Present address: Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC) Campus-Vienna-Biocenter 1, 1030 Vienna, Austria
[8]Present address: Institute of Clinical Chemistry and Clinical Pharmacology, University of Bonn and University Hospital Bonn, Bonn, Germany
[9]Lead contact
*Correspondence: hornung@genzentrum.lmu.de
https://doi.org/10.1016/j.immuni.2022.10.021

SUMMARY

The NLRP3 inflammasome plays a central role in antimicrobial defense as well as in the context of sterile inflammatory conditions. NLRP3 activity is governed by two independent signals: the first signal primes NLRP3, rendering it responsive to the second signal, which then triggers inflammasome formation. Our understanding of how NLRP3 priming contributes to inflammasome activation remains limited. Here, we show that IKKβ, a kinase activated during priming, induces recruitment of NLRP3 to phosphatidylinositol-4-phosphate (PI4P), a phospholipid enriched on the *trans*-Golgi network. NEK7, a mitotic spindle kinase that had previously been thought to be indispensable for NLRP3 activation, was redundant for inflammasome formation when IKKβ recruited NLRP3 to PI4P. Studying iPSC-derived human macrophages revealed that the IKKβ-mediated NEK7-independent pathway constitutes the predominant NLRP3 priming mechanism in human myeloid cells. Our results suggest that PI4P binding represents a primed state into which NLRP3 is brought by IKKβ activity.

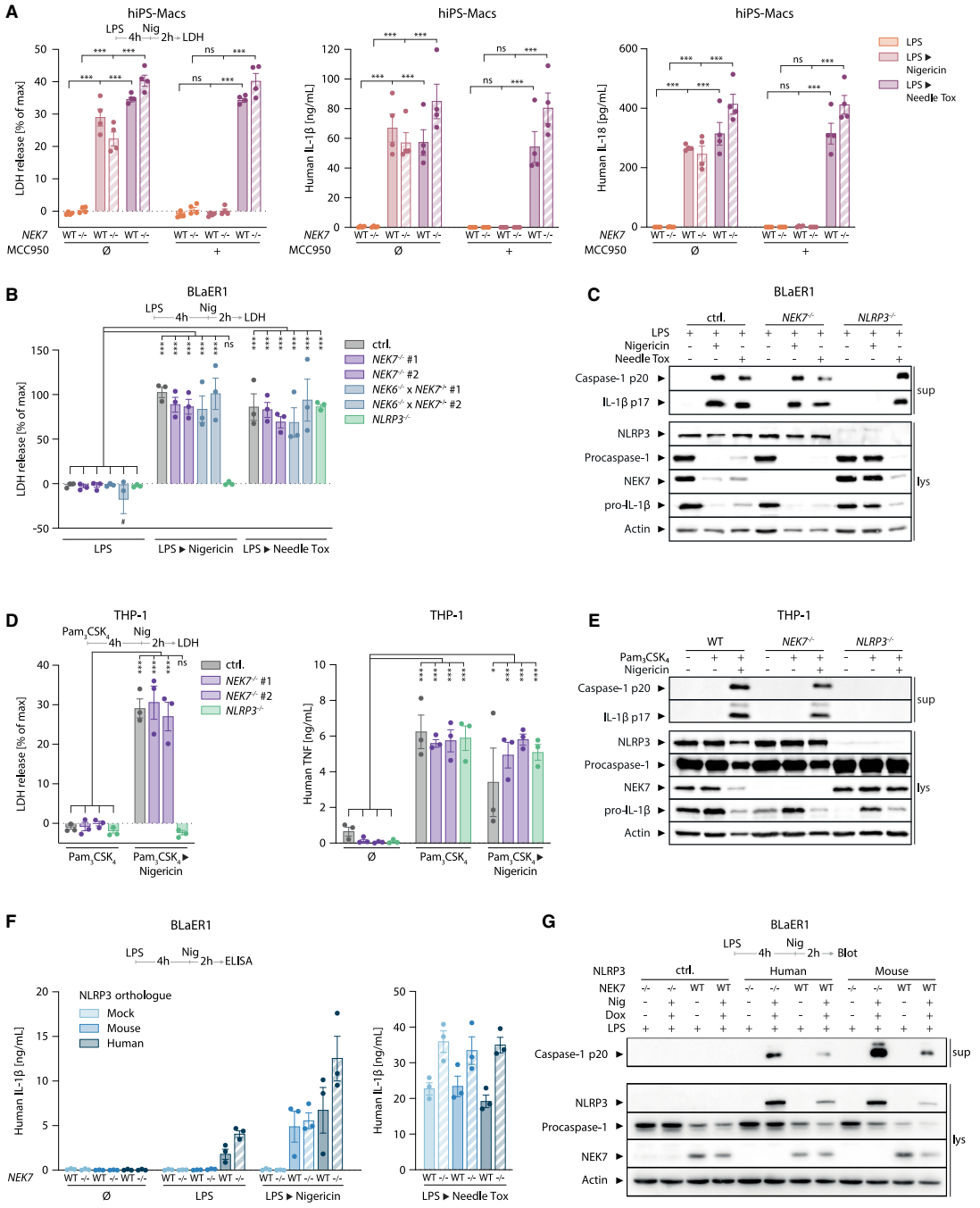
INTRODUCTION

Cells of the innate immune system employ a repertoire of so-called pattern recognition receptors (PRRs) to discriminate self from non-self. Engagement of these PRRs triggers a broad array of effector functions geared toward eliminating a microbial threat. The inflammasome pathway constitutes a special class of this PRR system that is signified by the activation of the cysteine protease caspase-1 in a large supramolecular protein complex.[1] Activation of caspase-1 causes maturation of pro-inflammatory cytokines, most prominently IL-1β,[2] as well as the induction of a special type of cell death, known as pyroptosis.[3] Among inflammasome sensors, NLRP3 plays a pivotal role in antimicrobial defense as well as sterile inflammatory diseases.[4] This is owed to the fact that NLRP3 is a highly sensitive, yet non-specific PRR. In this regard, NLRP3 has been shown to respond to the perturbation of cellular homeostasis by a broad array of diverse stimuli, rather than being activated by a specific microbe-derived molecule.[5] K^+ efflux from the cytosol has been identified as a common denominator of many NLRP3 triggers.[6] In this function, several types of lytic cell death have been shown to result in secondary engagement of the NLRP3 inflammasome pathway.[7] However, K^+ efflux-independent NLRP3 stimuli have also been described,[8,9] and a recent report has identified dispersal of the *trans*-Golgi network (TGN) as a common denominator of both K^+ efflux-dependent and -independent NLRP3 triggers.[10]

Unlike other inflammasome sensors, NLRP3 critically depends on the engagement of a priming step.[11] This priming signal can be provided by different types of receptors, typically PRRs that trigger NF-κB activation. Lipopolysaccharide (LPS) activating TLR4 is commonly used to provide a priming signal preceding the actual NLRP3 activation step. Initially, the necessity of priming had been ascribed to the fact that NLRP3 is expressed at limiting amounts in murine macrophages. In this respect, it has been shown that in a process now also called "transcriptional

CellPress

Immunity
Article



(legend on next page)

priming," NF-κB activating stimuli drive the expression of *Nlrp3*, thereby facilitating its activation.[12,13] In line with these findings, inhibition of transcription blocks this mode of NLRP3 priming, whereas transgenic expression of NLRP3 bypasses the requirement of transcriptional priming.[12,13] Extending this concept, NLRP3 can also be primed non-transcriptionally, e.g., by a short pulse of LPS treatment.[14–16] These modes of priming have been ascribed to a variety of post-translational modifications of NLRP3, including phosphorylation, de-phosphorylation, de-ubiquitination, and de-sumoylation.[17,18] Although being mechanistically unrelated, these events are commonly referred to as post-translational or non-transcriptional priming. The fact that many cells already express NLRP3 at sufficient amounts under steady-state conditions underscores the importance of non-transcriptional priming.[19]

Despite considerable insight into pathways that result in NLRP3 priming, the activation step of the NLRP3 inflammasome and its interconnection with priming have remained enigmatic. In this regard, we and others have identified the mitotic spindle kinase NEK7 (NIMA-related kinase 7) as a critical cofactor in NLRP3 activation in murine cells.[20–22] Notably, this role of NEK7 is distinct from its function in the cell cycle, as its kinase activity is not required for NLRP3 activation.[21,22] NEK7 has been suggested to interact with NLRP3 in a K^+ efflux-dependent manner, and deletion of NEK7 does not affect transcriptional NLRP3 priming.[21,22] This, in combination with a study modeling a NEK7-containing NLRP3 pyroptosome based on a cryo-EM structure of the NLRP3/NEK7 complex,[23] has led to the conclusion that NEK7 is involved in NLRP3 activation downstream of K^+ efflux.[24] Of note, studies identifying NEK7 as an indispensable factor for NLRP3 activation have mainly been conducted in murine models. Here, we report that reductionist genetic dissection of NLRP3 signaling in human cells revealed an additional pathway of NLRP3 priming that enables NLRP3 inflammasome activation independently of NEK7.

RESULTS

Human iPSC-derived macrophages and human myeloid cell lines activate NLRP3 independently of NEK7

We and others have previously described NEK7 to be essential for the activation of the NLRP3 inflammasome in the murine system.[20–22] To study the role of NEK7 in the human system, we adopted a recently described iPSC-derived macrophage model, in which human iPS cells are differentiated into macrophages *in vitro* (hiPS-Macs).[25] hiPS-Macs are fully capable of inflammasome activation: after priming with LPS, activation of the NLRP3 inflammasome with the ionophore Nigericin or the NAIP-NLRC4 inflammasome with Needle Tox resulted in pyroptosis (LDH release) accompanied by the release of IL-1β and IL-18 (Figures S1A and S1B). Both cytokine and LDH release in response to Nigericin, but not Needle Tox, were sensitive to the NLRP3 inhibitor MCC950 (Figures S1A and S1B). To investigate the role of NEK7 in NLRP3 inflammasome activation in hiPS-Macs, we generated *NEK7*$^{-/-}$ iPS cell clones via CRISPR-Cas9 genome editing. NEK7 deficiency neither affected macrophage differentiation nor did it lead to altered NLRP3 expression levels (Figure S1C). Contrasting previous reports from mouse cells,[20–22] NEK7-deficient hiPS-Macs showed no major impairment of their NLRP3 inflammasome response (Figures 1A and S1D). Cytokine and LDH release following Nigericin stimulation remained sensitive to MCC950 in *NEK7*$^{-/-}$ hiPS-Macs, confirming that Nigericin-induced pyroptosis was still mediated by NLRP3 in these cells (Figure 1A). As expected, NAIP-NLRC4 activation and IL-6 release also proceeded unperturbed in *NEK7*$^{-/-}$ hiPS-Macs (Figures 1A and S1D).

We then sought to further characterize NEK7-independent NLRP3 activation in human cells. To this end, we used the BLaER1 transdifferentiation system that we have previously adopted to study innate immune sensing.[26,27] Mirroring hiPS-Macs, NEK7-deficiency showed no impact on NLRP3 inflammasome activation as assessed by release of LDH and IL-1β (Figures 1B and S1E). To address whether the role of NEK7 for NLRP3 activation in human cells was overshadowed by a functional redundancy with its close homolog NEK6, we generated cells deficient for both NEK6 and NEK7. Analogous to NEK7-deficient cells, *NEK6*$^{-/-}$ × *NEK7*$^{-/-}$ BLaER1 cells displayed unimpaired activation of the NLRP3 inflammasome (Figures 1B and S1E). As expected, *NLRP3*$^{-/-}$ BLaER1 cells showed no response to Nigericin stimulation, whereas they remained responsive to NAIP-NLRC4 inflammasome activation (Figures 1B and S1E). In line with these observations, caspase-1 maturation upon Nigericin treatment also proceeded independently of NEK7 (Figure 1C). Pretreatment with the

Figure 1. Human iPSC-derived macrophages and human myeloid cell lines activate the NLRP3 inflammasome independently of NEK7

(A) Four clones per indicated genotype of human iPSCs were differentiated into macrophages (hiPS-Macs), primed with LPS for 4 h and then treated with the inflammasome activators Nigericin (NLRP3) or Needle Tox (NAIP-NLRC4) in the presence of the NLRP3 inhibitor MCC950 as indicated before release of LDH (left), IL-1β (middle), and IL-18 (right) was measured. Dots represent separately differentiated iPS cell clones of the indicated genotypes.

(B and C) BLaER1 monocytes of the indicated genotypes were primed with LPS for 4 h and subsequently stimulated with Nigericin or Needle Tox. LDH release (B) of one or two clones per genotype is depicted. (C) One representative immunoblot of three independent experiments is shown.

(D) Three clones of THP-1 cells of the indicated genotypes were primed with Pam_3CSK_4 for 4 h and subsequently stimulated with Nigericin for 2 h before release of LDH (left) and TNF (right) were measured. Two different sgRNAs against *NEK7* were used (#1 and #2). Dots represent individual clones.

(E) THP-1 cells of the indicated genotypes were primed with Pam_3CSK_4 for 4 h and subsequently stimulated with Nigericin for 2 h before immunoblotting. One representative immunoblot of three independent experiments is shown.

(F) *NLRP3*$^{-/-}$ BLaER1 cells expressing the indicated NLRP3 orthologs under the control of a doxycycline-inducible promoter were treated with doxycycline for the last 24 h of differentiation, primed with LPS for 4 h and subsequently stimulated with Nigericin (left) or Needle Tox (right) for 2 h. The same vector expressing mCherry instead of NLRP3 was used as a mock control.

(G) Western blot of cells treated as in (F), one representative of three independent experiments is shown.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise (one outlier in B is not depicted #). ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Tukey's test (A, B, and D: TNF) or Šidák's test (D: LDH). See also Figures S1 and S2.

CellPress

Immunity
Article

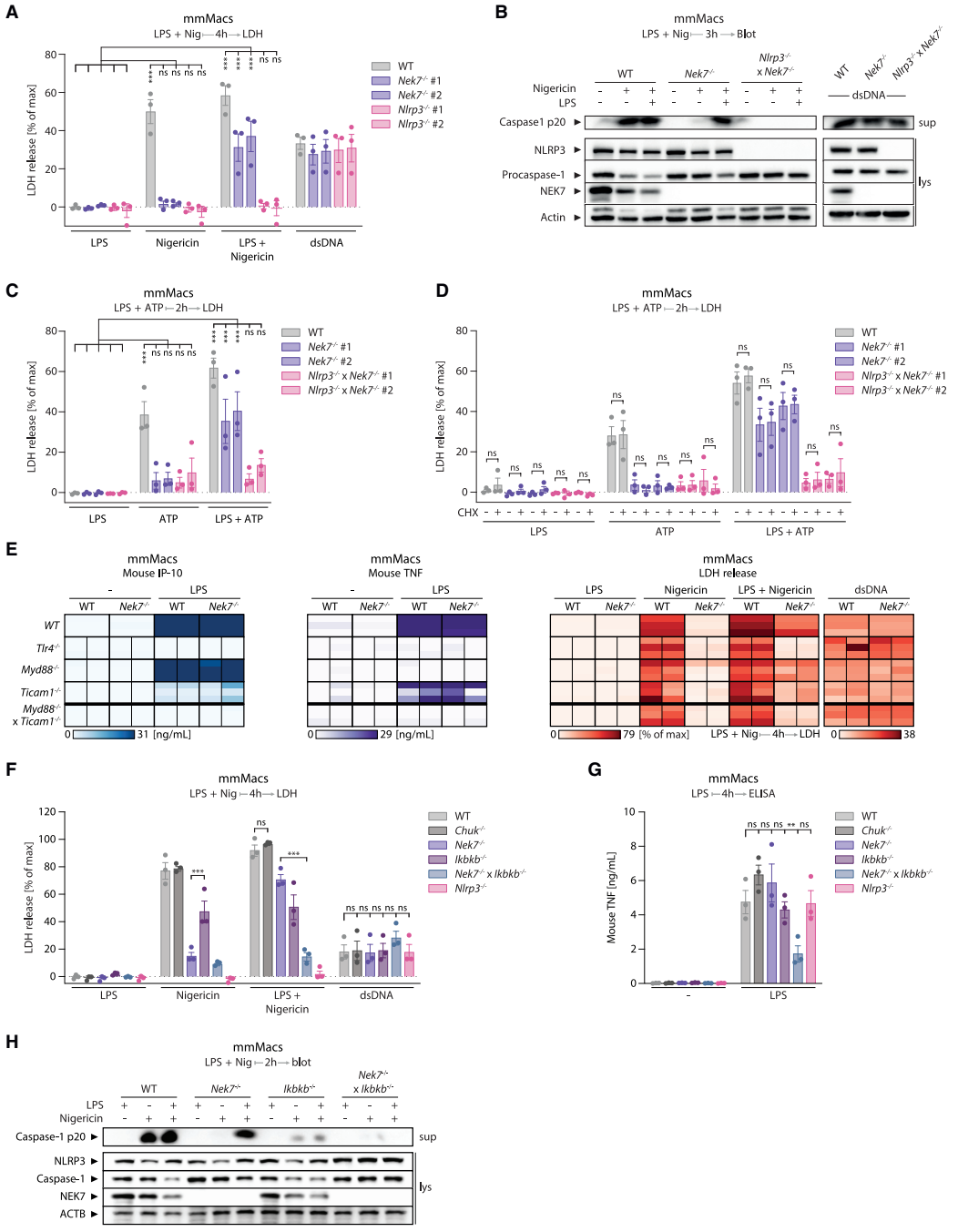


Figure 2. Priming activates IKKβ to bypass NEK7 via a translation-independent mechanism in mouse cells

(A–C) Mouse macrophages constitutively expressing mmNlrp3 (mmMacs) of the indicated genotypes were stimulated with LPS + Nigericin simultaneously for 4 h, with DNA for 28 h or with LPS + ATP for 2 h. (B) One immunoblot representative of two clones from two independent experiments is shown.

(legend continued on next page)

NLRP3-specific inhibitor MCC950[28] or prevention of K^+ efflux by increased extracellular K^+ concentration[6] blunted NLRP3 activation in wild type, *NEK7*$^{-/-}$ and *NEK6*$^{-/-}$ × *NEK7*$^{-/-}$ cells stimulated with Nigericin, whereas it left the NAIP-NLRC4 inflammasome intact (Figures S1F and S2A–S2D), indicating that Nigericin still relied on inducing K^+ efflux to trigger NLRP3 inflammasome activation in absence of NEK7. In line with the results obtained in BLaER1 cells, THP-1 cells deficient in NEK7 showed no attenuation of Nigericin-triggered inflammasome activation, whereas *NLRP3*$^{-/-}$ THP-1 cells were completely defective (Figures 1D, 1E, and S2E).

NEK7-independent NLRP3 activation in human cells contrasts with NEK7-dependent NLRP3 activation in mouse cells published by us and others.[20–22] To investigate if this difference is caused by species-specific features of the human and mouse orthologs of NLRP3, we reconstituted *NLRP3*$^{-/-}$ BLaER1 cells with different NLRP3 orthologs. Phenocopying the human ortholog, NEK7-deficient BLaER1 cells expressing mouse NLRP3 (mmNlrp3) mounted an unperturbed response to Nigericin (NLRP3) and Needle Tox (NAIP-NLRC4) (Figures 1F, 1G, and S2F). Taken together, these results demonstrate that unlike mouse cells, human cells are intrinsically capable of activating NLRP3 in a NEK6- and NEK7-independent manner.

Priming activates IKKβ to enable NEK7-independent NLRP3 inflammasome formation

Having established that human cells activate NLRP3 in absence of NEK7, we wondered whether the NEK7-independent pathway could be triggered in mouse cells where NLRP3 activation has been shown to depend on NEK7.[21] Here, we used an immortalized mouse macrophage cell line constitutively expressing mmNlrp3 (mmMacs) in which we had initially discovered the requirement of NEK7 for NLRP3 activation through a forward genetic screen.[20] These cells do not require transcriptional priming of NLRP3 for inflammasome activation, and stimulation with Nigericin alone already activated NLRP3 in a fully NEK7-dependent manner (Figures 2A and 2B). When testing different priming modalities, we found that simultaneous treatment with LPS and Nigericin led to NLRP3 activation independently of NEK7, as determined by LDH release and caspase-1 maturation 4 h after stimulation (Figures 2A and 2B). Concurrent stimulation with Pam_3CSK_4 or R848 instead of LPS (Figures S3A–S3D) and with ATP instead of Nigericin (Figure 2C) similarly resulted in a NEK7-independent response. Of note, this NEK7 bypass triggered by concurrent priming and stimulation was only uncovered when studying the inflammasome response several hours after treatment (Figure S3E). Indeed, the NLRP3 inflammasome response 1 h following concurrent LPS + Nigericin treatment was still NEK7-dependent (Figure S3F). However, concomitant LPS treatment enhanced this early NEK7-dependent NLRP3 inflammasome response compared with Nigericin treatment alone. This is consistent with previous reports on rapid, non-transcriptional NLRP3 priming enabling accelerated inflammasome formation.[14,15,29] Taken together, these results suggest that NEK7-mediated priming and the LPS-induced NEK7 bypass pathway are not only functionally redundant but may also act synergistically to accelerate NLRP3 activation.

LPS sensing initiates diverse transcriptional programs. However, NEK7-independent priming remained functional in the presence of translation-blocking concentrations of cycloheximide (CHX), indicating that it does not require *de novo* protein synthesis (Figures 2D and S3G). To elucidate the signaling cascade of NEK7-independent post-translational priming, we genetically perturbed TLR4 and its downstream signaling adaptors TRIF (*Ticam1*) and MyD88 in either unmodified or *Nek7*$^{-/-}$ mmMac cells. NLRP3 activation in response to Nigericin treatment remained intact in *Ticam1*$^{-/-}$ or *Myd88*$^{-/-}$ cells (Figure 2E, right panel; Table S1), whereas these cells displayed a selective lack of antiviral (IP-10) or pro-inflammatory (TNF) gene expression, respectively (Figure 2E, left and middle panels). TLR4 deficiency abrogated LPS-dependent cytokine production altogether (Figure 2E, left and middle panels). Accordingly, unlike their TLR4-sufficient counterparts, *Nek7*$^{-/-}$ × *Tlr4*$^{-/-}$ cells were fully defective in NLRP3 activation (Figure 2E). In contrast, *Nek7*$^{-/-}$ cells additionally deficient in either MyD88 or TRIF were still able to mount an NLRP3 inflammasome response after LPS + Nigericin treatment, albeit less effectively (Figure 2E). As expected, *Myd88*$^{-/-}$ × *Ticam1*$^{-/-}$ cells deficient in NEK7 were fully defective in NEK7-independent NLRP3 activation (Figure 2E). Altogether, these results indicate that the NEK7 bypass can be induced downstream of both Myd88 and TRIF signaling.

To identify the common factor mediating the NEK7 bypass, we turned our attention to the TAK and IKK complexes that constitute the apical kinase complexes governing pro-inflammatory signal transduction downstream of both MyD88 and TRIF. When we used the small molecule Takinib to block the activity of TAK1, the key kinase of the TAK complex, we found that the NEK7 bypass was largely inhibited, whereas NLRP3 activation in response to Nigericin remained intact (Figure S4A). We obtained analogous results when we blocked IKKβ, a kinase in the IKK complex, using TPCA-1 (Figure S4B). Of note, for both inhibitors, the NEK7 bypass was not fully abrogated; however, it was attenuated to the same extent as the production of the NF-κB-dependent cytokine TNF (Figures S4A and S4B). The NEK7 bypass was blocked when we deleted *Ikbkb*, the gene coding for IKKβ, but remained unperturbed when we deleted *Chuk*, the gene coding for IKKα (Figures 2F–2H). *Nek7*$^{-/-}$ × *Ikbkb*$^{-/-}$ mmMac cells were almost completely defective in NLRP3 inflammasome activation, whereas AIM2 inflammasome activation in response to dsDNA transfection

(D) mmMacs were pretreated with cycloheximide (CHX) for 30 min and stimulated as in (C).
(E) Two mmMacs clones per genotype were stimulated as indicated. Release of IP-10 (left), TNF (middle), and LDH (right) of two clones (sub-columns) from three independent experiments (sub-rows) are depicted as heatmaps.
(F and G) mmMacs of the indicated genotypes were stimulated as in (A) before the release of LDH (F) and TNF (G) was measured.
(H) mmMacs of the indicated genotypes were stimulated as in (A) for 2 h. One representative of three independent biological replicates is shown.
Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Tukey's test (A, C, and F), Šidák's test (D), or Dunnett's test (G).
See also Figures S3 and S4 and Table S1.

CellPress

Immunity
Article

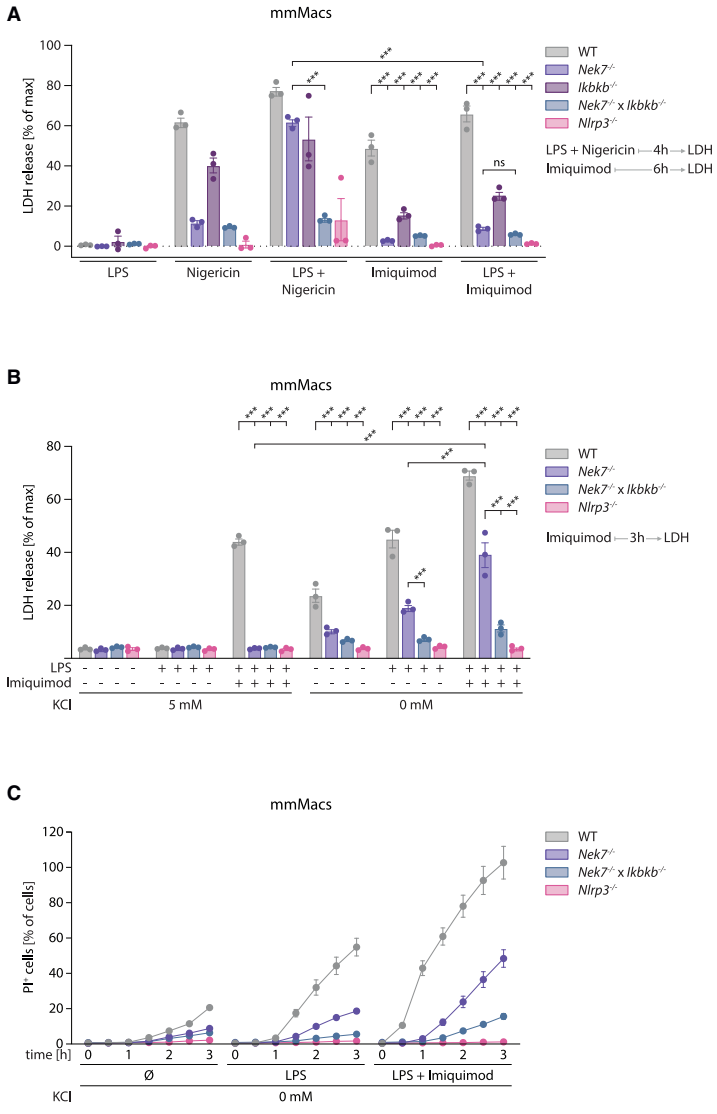


Figure 3. K⁺ efflux works synergistically with IKKβ to bypass NEK7

(A) mmMacs of the indicated genotypes were stimulated with the NLRP3 inflammasome inducers Nigericin or Imiquimod in the presence of LPS as indicated for 4 h (Nigericin) or 6 h (Imiquimod).

(B and C) mmMacs of the indicated genotypes were stimulated as indicated and imaged every 30 min (C) for 3 h in Hank's balanced salt solution with or without potassium + 10% FCS and 5 μg/mL propidium iodide before LDH release was measured (B). Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. ***p < 0.001, **p < 0.01, *p < 0.05, ns p ≥ 0.05 calculated by two-way ANOVA followed by Tukey's test.

remained intact (Figures 2F–2H). Priming with R848 or NLRP3 activation with ATP similarly resulted in IKKβ-dependent NLRP3 inflammasome formation independently of NEK7 (Figures S4C and S4D). In conclusion, since IKKβ is activated downstream of the TAK1 complex, these findings suggest that IKKβ constitutes the critical kinase mediating NEK7-independent NLRP3 inflammasome formation.

RIPK1 and caspase-8 have been implicated in non-transcriptional NLRP3 priming.[30] Although the NEK7 bypass continued to function in *Nek7*$^{-/-}$ × *Ripk1*$^{-/-}$ mmMac cells (Figure S4E), *Nek7*$^{-/-}$ × *Casp8*$^{-/-}$ mmMacs were fully defective in activating the NLRP3 inflammasome despite *Casp8*$^{-/-}$ cells displaying unperturbed NLRP3 activation (Figure S4F). ASC specking was also abrogated in *Nek7*$^{-/-}$ × *Casp8*$^{-/-}$ mmMacs in response to LPS + Nigericin (Figure S4G), showing that caspase-8 deficiency affects NEK7-independent NLRP3 priming upstream of inflammasome formation. Since we found IKKβ to be crucial for the NEK7 bypass, we checked whether caspase-8 deficiency had an effect on IKKβ activity.[31] Indeed, we observed reduced IKKβ phosphorylation after LPS stimulation of *Casp8*$^{-/-}$ mmMacs (Figure S4H), suggesting that reduced IKKβ activity, rather than a specific role of caspase-8, explains the inability of *Nek7*$^{-/-}$ × *Casp8*$^{-/-}$ mmMacs to activate NLRP3 in response to LPS + Nigericin.

In contrast to ATP and Nigericin, which depend on K^+ efflux to engage NLRP3, the TLR7 agonist Imiquimod (R837) has been shown to induce NEK7-dependent NLRP3 inflammasome formation independently of K^+ efflux.[9] In mmMac cells, Imiquimod strongly depended on NEK7 for NLRP3 activation even in combination with LPS (Figure 3A). Given that all K^+ efflux-dependent stimuli tested here can engage the NEK7 bypass with concurrent IKKβ activation, we investigated whether K^+ efflux might boost Imiquimod-driven NLRP3 activation in *Nek7*$^{-/-}$ mmMacs. Indeed, under low extracellular K^+ conditions that facilitate K^+ efflux,[10] Imiquimod stimulation together with LPS led to a NEK7-independent response that was significantly increased over LPS stimulation alone and not detectable with a physiological extracellular K^+ concentration of 5 mM (Figures 3B and 3C). Although the relative contributions of LPS- or Imiquimod-induced IKKβ activity and K^+ efflux- or Imiquimod-induced NLRP3 activation remain unclear, these data indicate that K^+ efflux enhances the NEK7-bypassing effect of IKKβ activation.

Human myeloid cells use IKKβ instead of NEK7 to prime NLRP3 by default

Moving back into the human system, we wondered whether NLRP3 priming through IKKβ was also responsible for the NEK7-independence of NLRP3 activation in human cells. Using the hiPS-Mac model, we found that *IKBKB*$^{-/-}$ cells showed a strong defect in NLRP3 inflammasome activation, whereas NAIP-NLRC4 activation proceeded normally, with IL-18 release being partially compromised (Figures S5A and S5B). However, we also observed a reduction in IL-6 amounts in IKKβ-deficient hiPS-Macs following LPS stimulation (Figure S5C). IKKβ, by governing NF-κB-dependent NLRP3 expression and also mediating the non-transcriptional NEK7 bypass, fulfills a dual role in NLRP3 priming. Hence, any effects on NLRP3 priming in *IKBKB*$^{-/-}$ hiPS-Macs cannot unequivocally be ascribed to either transcriptional or non-transcriptional NLRP3 priming based on these experiments. Although these results establish that IKKβ is critical for NLRP3 priming in human cells, the relative contributions of transcriptional and non-transcriptional priming remain unclear in the hiPS-Mac model.

To clarify whether transcriptional or non-transcriptional NLRP3 priming is the predominant priming modality in the human system, we employed the BLaER1 model system. Given that hiPS-Macs express NLRP3 under steady-state conditions without transcriptional priming, we first sought to clarify if transcriptional priming was required for NLRP3 activation in BLaER1 cells. Although BLaER1 cells deficient in TAK1 (*MAP3K7*), in which NF-κB-mediated transcription after LPS sensing is completely abrogated, did indeed not produce pro-IL-1β upon LPS treatment anymore, they still expressed NLRP3 (Figure S5D). Congruently, blocking protein translation with CHX did not affect NLRP3 activation in these cells (Figure S5E). These data show that in BLaER1 cells, transcriptional priming is not required for NLRP3 inflammasome activation. Still, again mirroring hiPS-Macs, stimulation with Nigericin alone was not sufficient to activate NLRP3, but additional treatment with LPS was required to enable NLRP3 inflammasome formation in BLaER1 cells (Figure S5F). The NAIP-NLRC4 inflammasome formed in response to Needle Tox irrespectively of LPS priming as expected (Figure S5F). These data demonstrate that BLaER1 cells require non-transcriptional priming of NLRP3 for inflammasome activation. In line with these findings, a short pulse of concomitant LPS + Nigericin treatment led to robust NLRP3 activation in BLaER1 cells (Figure S5G). RIPK1, RIPK3, and caspase-8 were dispensable for NLRP3 activation in response to Nigericin and NLRC4 activation, but in *GSDMD*$^{-/-}$ BLaER1 cells, LDH release for both inflammasomes was blunted (Figure S5H).

Given that non-transcriptional priming was still dependent on TAK1 in BLaER1 cells and that TAK1 activates IKKβ, we then assessed NLRP3 activation in BLaER1 cells deficient for IKKβ. Corroborating our findings from hiPS-Macs and the murine system, LDH release and caspase-1 maturation following NLRP3 activation were blunted in *IKBKB*$^{-/-}$ BlaER1 cells (Figures 4A, 4B, and S5I). In contrast, cells deficient in IKKα (*CHUK*), a close homolog of IKKβ, did not display a defect in inflammasome formation (Figure 4A). Cells deficient in both IKKα and IKKβ (*CHUK*$^{-/-}$ × *IKBKB*$^{-/-}$) phenocopied *IKBKB*$^{-/-}$ cells (Figures 4A and 4B). As expected, given the steady-state expression of NLRP3 in BLaER1 cells, *RELA*$^{-/-}$ × *RELB*$^{-/-}$ cells displayed unperturbed NLRP3 activation (Figures 4A and S5I) despite strongly reduced pro-inflammatory cytokine transcription (Figure S5J). Reconstitution of *IKBKB*$^{-/-}$ BlaER1 cells with wild-type IKKβ, but not IKKβ-K44M, a kinase-dead mutant of IKKβ,[32] rescued NLRP3 activation, showing that the kinase activity of IKKβ was required for non-transcriptional NLRP3 priming (Figures 4C and 4D).

To investigate the kinetics of IKKβ-mediated non-transcriptional NLRP3 priming, we added the IKKβ inhibitor TPCA-1 to BLaER1 cells at different time points pre and post NLRP3 priming. Expectedly, adding TPCA-1 concurrently with LPS blocked all priming and abrogated NLRP3 activity (Figure 4E). However, adding TPCA-1 concurrently with or 30 min after Nigericin also blocked or strongly reduced NLRP3 activity, respectively (Figure 4E). Experiments with primary human monocytes corroborated these findings (Figure S5K). In summary, these data show that rapid, non-transcriptional priming by IKKβ is required for NLRP3 activation, further suggesting that human cells are NLRP3 inflammasome competent in the absence of NEK7 because they engage IKKβ by default.

Synergistically with IKKβ, NEK7 can accelerate NLRP3 activation human cells

Having demonstrated that IKKβ activation constitutes the predominant priming pathway in the human system, we wondered whether NEK7-mediated priming could be used by human cells at all. A hallmark of NEK7-mediated NLRP3 priming is the direct interaction of NEK7 and NLRP3.[21] NLRP3 co-immunoprecipitated with NEK7 from THP-1 cells, indicating that the human NEK7 protein (hsNEK7) could in principle function to prime NLRP3 (Figures 5A and S6A). Of note, this interaction was independent of K^+ efflux. We then reconstituted NLRP3 inflammasome signaling in HEK-293T cells, which normally do not express NLRP3 or ASC, the core signaling components of the NLRP3 inflammasome (Figures S6B and S6C). Notably, in this reconstitution system, inflammasome activation is driven by overexpression of NLRP3 and proceeds without stimulation by Nigericin. Hence, we consider inflammasome formation in this HEK-293T inflammasome assay to directly report the priming status of NLRP3. Here, we found that the mouse and human orthologs of NEK7 enhanced the activation of both NLRP3 orthologs, showing that hsNEK7 is capable of priming NLRP3 (Figures 5B and S6D). To investigate if NEK7 has a physiological role in NLRP3 priming, we went back to our hiPS-Mac system. Since we had found NLRP3 activation to require both NEK7 and LPS priming after concomitant LPS + Nigericin stimulation at early time points in mouse cells (Figure S3F), we tested the same condition in hiPS-Macs. Indeed, concomitant stimulation with LPS + Nigericin for 1 h resulted in NEK7-dependent release of LDH, whereas 4 h of LPS + Nigericin stimulation rendered NLRP3 activation NEK7-independent (Figures 5C and 5D).

From these data, we conclude that IKKβ, which is required to activate NLRP3 in all human cell lines tested here, operates in synergy with NEK7 to drive NLRP3 priming. NEK7 can accelerate NLRP3 priming at early time points, when IKKβ is not yet fully active. At later time points, IKKβ becomes redundant with NEK7.

Recruitment of NLRP3 to PtdIns4P induces NEK7-independent inflammasome activation

Finally, we investigated how IKKβ activation enables NEK7-independent NLRP3 activation. As it has recently been reported that

CellPress

Immunity
Article

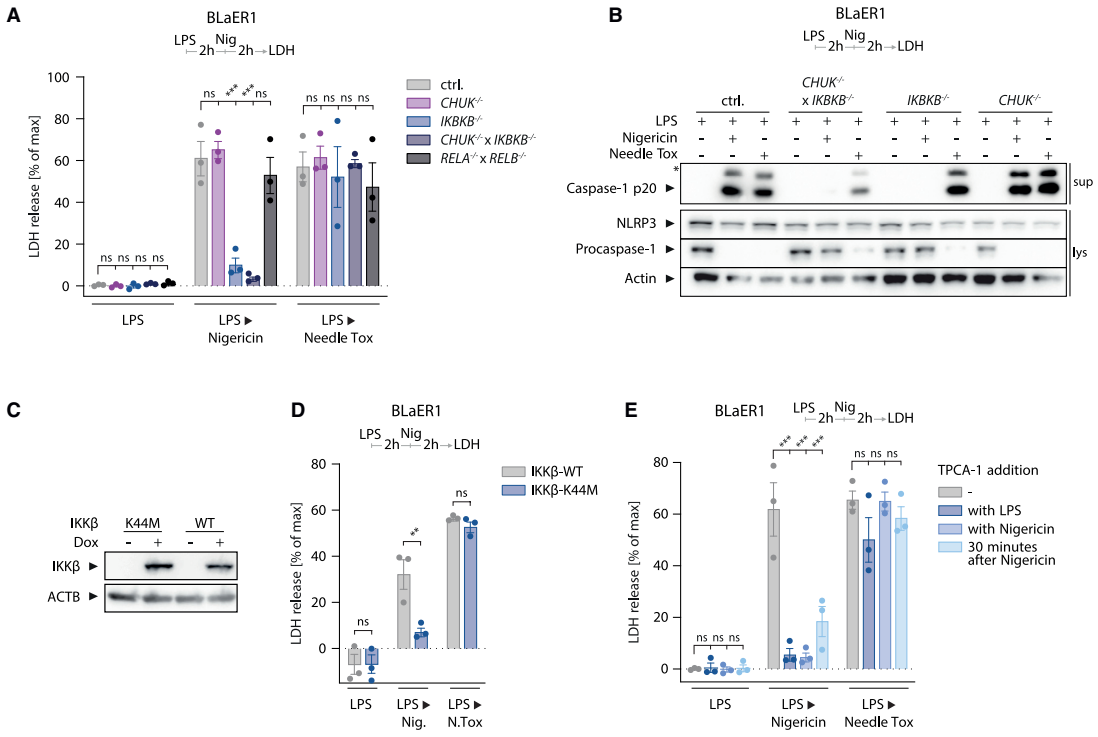


Figure 4. NLRP3 priming through IKKβ is required for inflammasome activation in human myeloid cell lines

(A) LDH release from BLaER1 clones of the indicated genotypes primed with LPS for 2 h and subsequently treated with Nigericin or Needle Tox for 2 h.

(B) One representative of three immunoblots from cells treated as in (A). The asterisk denotes an unspecific band.

(C) Immunoblot of *IKBKB*$^{-/-}$ BLaER1 cells expressing wild-type IKKβ or kinase-dead IKKβ-K44M under the control of a doxycycline-inducible promoter treated with doxycycline during the last 8 h of differentiation.

(D) LDH release from BLaER1 cells as in (C) primed with LPS for 2 h and subsequently treated with Nigericin or Needle Tox as indicated.

(E) LDH release from BLaER1 monocytes primed with LPS for 2 h before stimulation with Nigericin or Needle Tox. TPCA-1 was added at different time points as indicated.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Dunnett's test (A and E) or Šidák's test (D).

See also Figure S5.

interaction of NLRP3 with phosphatidylinositol-4-phosphate (PI4P) on the TGN is an essential requirement for inflammasome formation,[10] we investigated the subcellular localization of NLRP3 during priming. To this end we generated *Pycard*$^{-/-}$ J774 mouse macrophages expressing a fusion protein of the PI4P-binding pleckstrin homology (PH)-domain of oxysterol-binding protein (OSBP) and mCherry (OSBP[PH]-mCherry). In these cells, we found LPS treatment to result in the accumulation of NLRP3 at PI4P-rich sites (Figures 6A and 6B). Of note, this translocation cannot be caused by NLRP3-mediated pyroptosis, since *Pycard*$^{-/-}$ cells are incapable of NLRP3 inflammasome formation. The recruitment of NLRP3 to PI4P was markedly reduced by the IKKβ inhibitor TPCA-1 (Figures 6A and 6B). In line with our findings on LPS-dependent non-transcriptional priming in human and mouse cells, NLRP3 recruitment to PI4P occurred rapidly, generally within 30 min after LPS stimulation (Figure S6E). We did not observe NLRP3 translocation to mitochondria—in fact, PI4P-rich sites appeared mostly distinct from mitochondria (Figure S6F). To identify the cellular compartment that NLRP3 is recruited to, we fractionated lysates of *Pycard*$^{-/-}$ J774 cells. Post-nuclear lysates were centrifuged at 5,000 × *g* to obtain a pellet (P5) and supernatant (S5) fraction. The S5 fraction was further subjected to centrifugation at 100,000 × *g* to yield a pellet (P100) and supernatant (S100) fraction. We found NLRP3 in all fractions irrespectively of LPS priming or concomitant IKKβ inhibition (Figure 6C). However, when we further fractionated P100 across a linear sucrose gradient, we found NLRP3 to become enriched in the top fractions upon LPS stimulation, where we also found the PI4P-binding OSBP(PH)-mCherry fusion protein (Figure 6D). This enrichment of NLRP3 was blocked in the presence of TPCA-1, and, in line with our imaging data, unstimulated cells showed some NLRP3 enrichment on both ends of the gradient. Of note, the mitochondrial membrane protein TOMM40 was also present in

Immunity
Article

CellPress

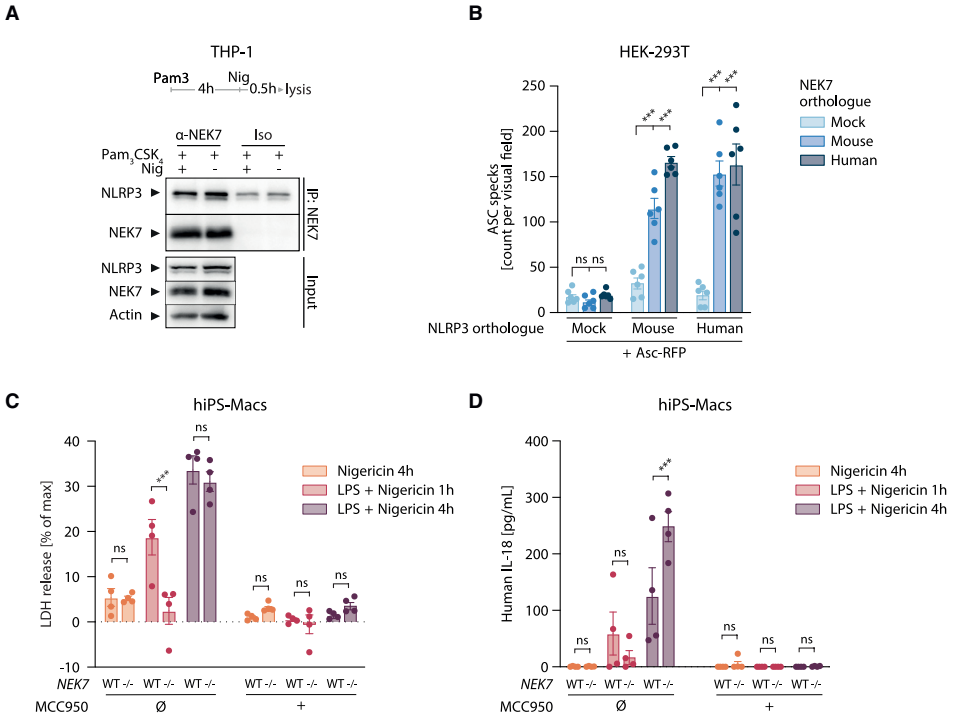


Figure 5. NEK7 accelerates NLRP3 activation at early priming time points in iPSC-derived human macrophages

(A) THP-1 cells were primed with Pam_3CSK_4 for 4 h and then stimulated with Nigericin for 30 min before lysates were immunoprecipitated with anti-NEK7 antibody or isotype control. One representative immunoblot of three independent experiments is shown.

(B) *NEK7*$^{-/-}$ HEK293T cells were transiently transfected with plasmids driving expression of an ASC-RFP fusion protein and mouse or human orthologs of NLRP3 and NEK7 as indicated. ASC-RFP specks were imaged 24 h after transfection. Dots represent technical replicates from one representative of three independent experiments.

(C and D) Four clones per genotype of *NEK7*$^{-/-}$ or wild-type human iPS cells were differentiated into hiPS-Macs and treated with Nigericin or LPS + Nigericin for 4 h or LPS + Nigericin for 1 h in the presence of the NLRP3 inhibitor MCC950 as indicated. Dots represent individual clones.

***p < 0.001, **p < 0.01, *p < 0.05, ns p ≥ 0.05 calculated by two-way ANOVA followed by Dunnett's test (B) or Šidák's test (C and D).

See also Figure S6.

the P100 fraction, but at the opposite end of where the OSBP(PH)-mCherry construct was found. We then analyzed the organelles present in fractions #2 and #11 via mass spectrometry (Table S2). The TGN, but not the *cis*-Golgi network, was highly enriched in fraction #2 along with weakly PI4P$^+$ organelles such as endosomes[33] (Figures 6E and S6G). Taken together, upon priming, NLRP3 translocates to PI4P-rich sites mostly on the TGN.

Based on these data, we hypothesized that the accumulation of NLRP3 on PI4P-rich sites induces NEK7-independent NLRP3 activation. To confirm this hypothesis, we directly tethered NLRP3 to PI4P by fusing it to the PH-domain of OSBP as reported before.[10] Although a previously described K127A, K128A, K129A, and K130A quadruple mutant of mmNlrp3 (Nlrp3(4KA)) was incapable of localizing to the TGN in J774 cells, Nlrp3(4KA-OSBP(PH)) constitutively localized to the TGN as expected (Figure S6H). When we expressed wild-type Nlrp3, Nlrp3(4KA), and Nlrp3(4KA-OSBP(PH)) in *Nlrp3*$^{-/-}$ J774 mouse macrophages, we found wild-type Nlrp3 to facilitate caspase-1 maturation in a NEK7-dependent manner and Nlrp3(4KA-OSBP(PH)) to activate caspase-1 independently of NEK7 (Figure 6F). Nlrp3(4KA) expectedly did not lead to any detectable caspase-1 processing (Figure 6F). Of note, these cells did not require priming with LPS, as they expressed Nlrp3 under the control of a doxycycline-inducible promoter, mirroring above results (Figure 2).

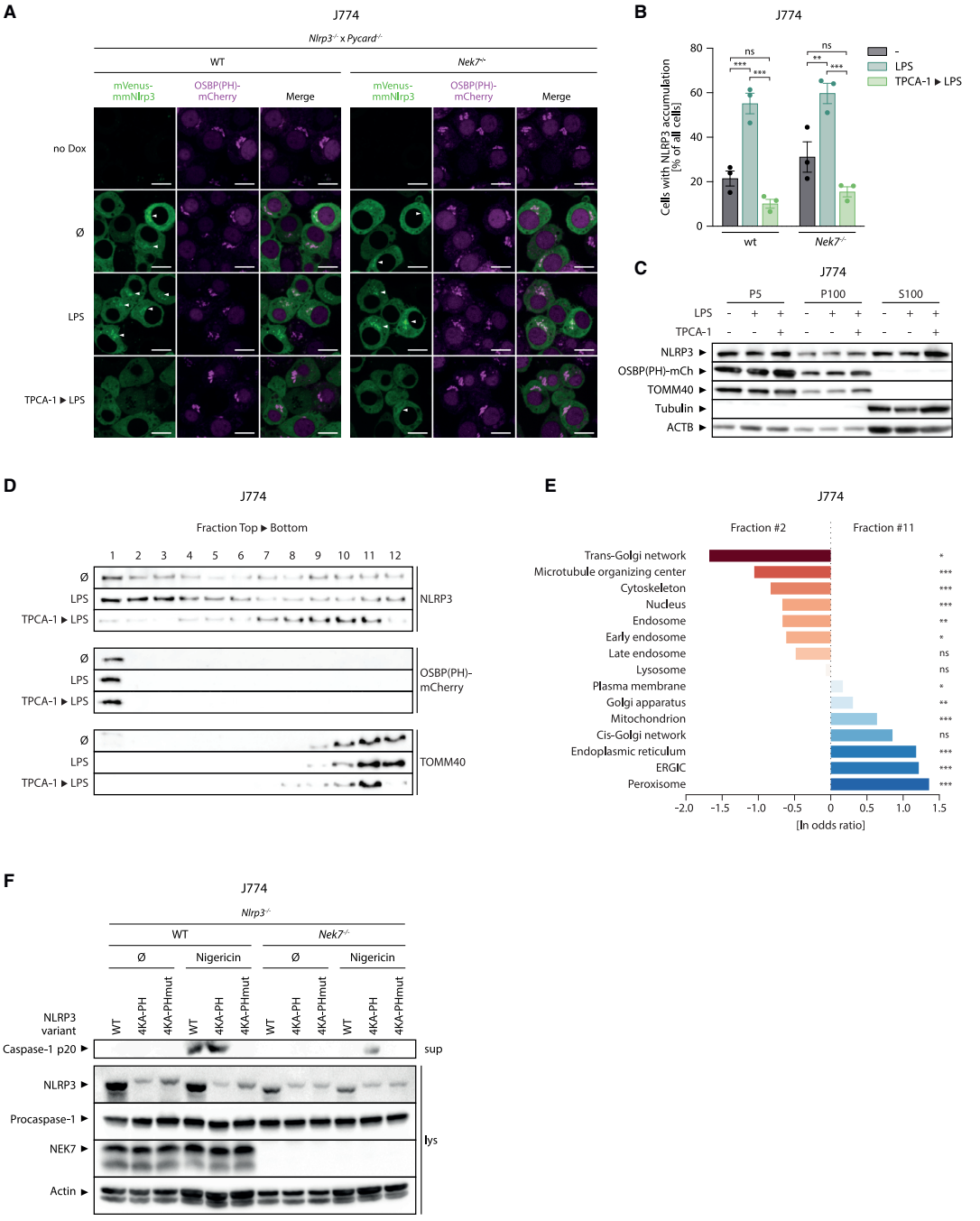
Together, these results demonstrate that IKKβ induces NEK7-independent NLRP3 priming by increasing the recruitment of NLRP3 to PI4P and establish PI4P-recruitment of NLRP3 as a priming modality of the inflammasome (Figure S6I).

DISCUSSION

Since its first description in 2001,[34] NLRP3 has attracted much attention as a key driver of antimicrobial and sterile inflammation.[7] Nonetheless, despite being in the focus for almost two decades, the molecular mechanism of NLRP3 activation has remained obscure. The two-step model of inflammasome priming and activation predates the discovery of NLRP3 and

CellPress

Immunity
Article



(legend on next page)

inflammasomes altogether, originating from the notion that both a pro-inflammatory and a cell-death inducing signal are required to release mature IL-1β from murine bone marrow-derived macrophages.[35] In retrospect, these early studies had assessed NLRP3 inflammasome activation employing a K^+ efflux-inducing trigger. Subsequent studies have revealed that the pro-inflammatory signal indeed serves two independent functions in the context of NLRP3 inflammasome activation. Although this signal is critically required to induce pro-IL-1β expression, it is also necessary to render NLRP3 activatable in the first place. This became apparent when studying the maturation of caspase-1, the expression of which is independent of a pro-inflammatory signal, as a proxy of NLRP3 inflammasome activation. Here, it has been revealed that unprimed macrophages do not mature caspase-1 upon K^+ efflux-inducing stimuli[13,36] but that additional priming by a pro-inflammatory signal was required to facilitate this step. Of note, this unique requirement of NLRP3 priming by a pro-inflammatory signal (referred to as signal 1 or priming in this manuscript) must not be confused with the signal that induces pro-IL-1β expression. Indeed, although both signals can be provided through the same PRR, they can also be separated, and the pro-IL-1β inducing stimulus is not necessary for NLRP3 inflammasome activation.

Although the two-step activation model constitutes an important conceptual framework for NLRP3 activation, it has proven to be an enormous conundrum because it is not trivial to allocate signaling events upstream of NLRP3 to either priming or activation. The fact that several pathways toward NLRP3 priming have been described[37] is likely attributable to stimulus-, cell type-, and species-dependent aspects as well as temporal dynamics playing an important role in this context. We conceptualize that priming serves the function to increase the cellular pool of NLRP3 molecules that are able to respond to an activating stimulus, either by upregulating production of the NLRP3 protein or by lowering the activation threshold of individual NLRP3 molecules. In this regard, we would interpret the existence of multiple redundant NLRP3 priming pathways as the possibility to integrate diverse pro-inflammatory inputs to achieve this activatable state. In fact, we consider this pleiotropy to be a key trait of NLRP3 priming, but not activation pathways.

The mitotic spindle kinase NEK7 has been shown to be an essential cofactor of NLRP3 activation,[20–22] and it has been suggested that NEK7 facilitates inflammasome formation by mediating recognition of the second signal.[21,23] Studying the role of NEK7 in iPS-cell-derived human macrophages, we made the unexpected discovery that NLRP3 activation can be fully operational in the absence of NEK7. By genetically dissecting NLRP3 inflammasome signaling, we uncovered that these cells employ a NEK7-independent signaling cascade instead that drives IKKβ-dependent, post-translational priming of NLRP3. Although this IKKβ-dependent priming signal is the default pathway by which human cells engage the NLRP3 inflammasome, murine macrophages predominantly rely on NEK7 for NLRP3 priming. However, they can bypass NEK7 and switch to IKKβ-dependent priming under pro-inflammatory conditions signified by, for example, TLR activation. The NEK7-independence in human myeloid cells could not be attributed to species-specific constitutions of the NEK7 or NLRP3 molecules themselves: immunoprecipitation and reconstitution experiments showed that human NEK7 interacted with human NLRP3 and that NEK7 was able to facilitate NLRP3 activity. In line with this notion, iPSC-derived human macrophages also employ NEK7 to activate the NLRP3 inflammasome; however, this requires LPS priming and indicates a synergy between NEK7 and IKKβ only observed at an early time point, when the IKKβ post-translational priming mechanism is not yet fully operational. Indeed, in these cells, NEK7 becomes obsolete after prolonged LPS-priming when the IKKβ priming cascade is active. Mechanistically, IKKβ activity recruited NLRP3 to PI4P, a phospholipid enriched on the TGN. Tethering NLRP3 to PI4P led to inflammasome activation independently of NEK7, confirming that increased PI4P interaction serves to prime NLRP3 for inflammasome formation. Based on the redundancy between IKKβ and NEK7 in facilitating NLRP3 inflammasome formation, we conclude that NEK7 serves as a priming factor of the NLRP3 inflammasome.

NEK7 holds a unique position among NLRP3 priming pathways in that it is constitutively expressed and apparently uncoupled from upstream signals in its pro-inflammatory capacity. It has been suggested that NEK7 is employed for NLRP3 activation to avoid inflammasome formation during mitosis, when NEK7 is not available.[22] Furthermore, it has been speculated that the cellular perturbation triggering NLRP3 commonly occurs during mitosis, and thus, the dependency on NEK7 prevents

Figure 6. IKKβ-mediated recruitment of NLRP3 to PI4P enables NEK7-independent inflammasome formation

(A) *Nlrp3*$^{-/-}$ × *Pycard*$^{-/-}$ J774 cells of the indicated *Nek7* genotypes expressing mCherry tethered to phosphatidylinositol-4-phosphate (PI4P) via the PH domain of OSBP (OSBP(PH)-mCherry) and doxycycline-inducible mVenus-mmNlrp3 were treated with doxycycline for 24 h and TPCA-1 for 1 h before stimulation with LPS for 30 min. Scale bars represent 10 μm.

(B) Quantification of at least 10 randomly chosen fields of view per experimental condition from three independent experiments described in (A). Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days.

(C) Lysates of J774 cells pretreated with TPCA-1 for 1 h and then stimulated with LPS for 30 min were depleted of nuclei (5 min 1,000 × *g*), and the supernatant was then centrifuged at 5,000 × *g* for 10 min (pellet P5) followed by 100,000 × *g* for 20 min (pellet P100, supernatant S100) before immunoblotting. One representative of three independent experiments is shown.

(D) P100 fractions from (C) were further fractionated across a linear sucrose gradient (20%–60%) into 12 fractions which were then immunoblotted. One representative of three independent biological replicates is shown.

(E) Enrichment of organelle-specific protein sets identified via mass spectrometry analysis of the protein content of fractions #2 and #11. p-values for set enrichment were calculated based on proteins differing between the two fractions (FC ≥ 1.5, FDR < 0.05) using Fisher's exact test with Benjamini-Hochberg correction.

(F) *Nlrp3*$^{-/-}$ J774 cells of the indicated *Nek7* genotypes expressing doxycycline-inducible variants of Nlrp3 as indicated were treated with doxycycline for 18 h followed by 2 h of Nigericin before immunoblotting.

***p < 0.001, **p < 0.01, *p < 0.05, ns p ≥ 0.05 calculated by two-way ANOVA followed by Tukey's test unless indicated otherwise.

See also Figure S6 and Table S2.

inadvertent inflammasome activation during cell division.[23] However, the here-uncovered redundancy of NEK7 priming with other cell cycle-independent priming pathways (e.g., IKKβ) advocates against a specific de-coupling of NLRP3 inflammasome activation and proliferation. This is also in line with the fact that many NLRP3 inflammasome-competent cells of the innate immune system are postmitotic. As such, despite detailed mechanistic insight into how NEK7 can accelerate NLRP3 inflammasome activation, the physiological role of NEK7 remains to be determined. The redundancy of NEK7 with a priming factor that acts by enhancing the interaction of NLRP3 and PI4P suggests that NEK7 itself might be involved in recruiting NLRP3 to PI4P at the TGN.

The role of K^+ efflux is currently debated in the field: although it was recently shown that K^+ efflux alone is not sufficient to drive inflammasome activation in primed BMDMs and consequently argued that K^+ efflux only promotes recruitment of NLRP3 to the TGN,[10] an older report demonstrated that K^+ efflux does indeed suffice: inflammasome activation did occur in response to K^+ efflux in primed BMDMs.[6] Our study shows that recruitment of NLRP3 to PI4P can be induced by IKKβ activation independently of K^+ efflux. In line with the latter report, K^+ efflux was still required for inflammasome formation following IKKβ-mediated PI4P recruitment of NLRP3, hinting at a role of K^+ efflux beyond recruiting NLRP3 to PI4P. Whether K^+ efflux or dispersal of the TGN serves as the ultimate trigger of NLRP3 inflammasome formation remains to be investigated. From the fact that both IKKβ- and NEK7-mediated NLRP3 priming still require K^+ efflux for inflammasome formation but that IKKβ-mediated priming can bypass NEK7, we conclude that NEK7 itself acts as a priming factor upstream of K^+ efflux. Of note, K^+ efflux-independent NLRP3 activators have also been described.[8,9] For one such agonist, Imiquimod, the NEK7 bypass was only activated in the presence of K^+ efflux, suggesting that K^+ efflux boosts NEK7-independent NLRP3 activation synergistically with IKKβ.

Another study recently implicated IKKβ in the recruitment of NLRP3 to the TGN.[38] In contrast with our findings, in their setting, Nigericin stimulation was still required for TGN recruitment of NLRP3, as reported previously.[10] The authors concluded that IKKβ enhances Nigericin-dependent TGN dispersal, which they suggested to be the cause of increased NLRP3 activity.[38] However, whether increased TGN dispersal is a cause or an effect of increased cell death cannot be concluded from their work. In our study, we observed that IKKβ activation recruited NLRP3 to PI4P on an undispersed TGN independently of Nigericin stimulation or TGN dispersal in pyroptosis-deficient *Pycard*$^{-/-}$ cells. We showed that recruiting NLRP3 to an intact TGN was sufficient for subsequent inflammasome formation independently of an additional priming stimulus. Hence, it is unlikely that the increased TGN dispersal observed by Nanda and colleagues would explain the priming effect of IKKβ that we describe here. Rather, given that we observe K^+ efflux to act synergistically with IKKβ and NEK7, increased recruitment of NLRP3 to the TGN might explain the previously reported effects.[38]

This study establishes NEK7 as a priming rather than an activation signal for NLRP3. Moreover, in its capacity as a priming factor NEK7 does not constitute an absolute requirement for NLRP3 inflammasome activation. Instead, a priming signal emanating from IKKβ can fully compensate for NEK7 by enhancing the interaction of NLRP3 and PI4P. This signal supersedes the NEK7 requirement in human myeloid cell lines and also represents the dominant priming entity in iPSC-derived human macrophages.

Limitations of the study

We have shown that NEK7-independent priming of NLRP3 depends on the kinase activity of IKKβ but does not require *de novo* translation. However, the target that is phosphorylated by IKKβ remains to be determined in future studies. To confirm that recruitment of NLRP3 to PI4P is sufficient for NEK7-independent inflammasome activation, we overexpressed an engineered fusion protein of NLRP3(4KA) and the PI4P-interacting PH domain of the protein OSBP that constitutively interacts with PI4P, as reported previously.[10] Although we controlled for unspecific NLRP3 activation by expressing NLRP3 fused to a non-PI4P-binding point-mutated version of the same PH domain, we cannot exclude that engineering NLRP3 influenced its dependency on NEK7. Finally, owing to the fact that *Nek7*$^{-/-}$ mice are not viable,[39] this study does not include an experiment showing that IKKβ activation bypasses NEK7 *in vivo*.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - BLaER1 cells
  - THP-1 cells
  - mmMacs and J774 mouse macrophages
  - hiPSC, hiPS-Macs cell culture
  - Differentiation of hiPSCs into hiPS-Macs
  - HEK-293T cells
- METHOD DETAILS
  - Trans-Golgi network imaging
  - ASC speck imaging
  - Immunoblotting
  - ELISA and LDH assay
  - Stimulation of immune signaling
  - Inhibition of translation
  - Doxycyclin-inducible gene expression
  - Inhibition of TAK1, IKKβ and NLRP3
  - Inhibition and induction of K^+ efflux
  - Sucrose gradient fractionation
  - Mass spectrometry sample preparation
  - Analysis of MS samples
  - Transient Transfection of HEK-293T cells
  - Plasmid DNA purification
  - Preparation of lentiviral particles
  - Genome editing and overexpression
  - Plasmids
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.immuni.2022.10.021.

ACKNOWLEDGMENTS

We kindly thank Larissa Hansbauer, Jochen Rech, Claudia Ludwig, and Andreas Wegerer (Gene Center, LMU) for great technical support; the BioSysM FACS Core Facility (Gene Center, LMU) for cell sorting; the BioSysM Liquid Handling Unit (Gene Center, LMU) for lab automation; the Center for Advanced Light Microscopy (CALM) for support with confocal microscopy; Adam O'Neill and Magdalena Götz (Department of Physiological Genomics, LMU) for providing us with the hiPSCs and helping us set up experiments with these cells; Russell Vance (UC Berkeley, USA) for providing us with the Needle Tox expression plasmid; and Manuela Moldt and Karl-Peter Hopfner (Gene Center, LMU) for help in producing the Needle Tox protein. This work was funded by the European Research Council grant ERC-2020-ADG ENGINES (101018672), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) CRC 1403 (project number 414786233), and the Fondation Bettencourt Schueller to V.H.

AUTHOR CONTRIBUTIONS

Conceptualization: N.A.S., M.M.G., J.L.S.-B., and V.H.; formal analysis, software: N.A.S. and S.C.M.; investigation: N.A.S., F.O., M.M.G., I.S., J.M.K., J.L.S.-B, S.C.M., T.M.-K., D.C., D.N., C.A.S., H.H., A.L.F., and F.P.; resources: F.G., R.B., M.M., H.L., and V.H.; writing: N.A.S., M.M.G., and V.H. with input from all authors; funding acquisition: V.H.; supervision: V.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 14, 2022
Revised: July 17, 2022
Accepted: October 26, 2022
Published: November 15, 2022

REFERENCES

1. Broz, P., and Dixit, V.M. (2016). Inflammasomes: mechanism of assembly, regulation and signalling. Nat. Rev. Immunol. *16*, 407–420. https://doi.org/10.1038/nri.2016.58.
2. Dinarello, C.A. (2018). Overview of the IL-1 family in innate inflammation and acquired immunity. Immunol. Rev. *281*, 8–27. https://doi.org/10.1111/imr.12621.
3. Vande Walle, L., and Lamkanfi, M. (2016). Pyroptosis. Curr. Biol. *26*, R568–R572. https://doi.org/10.1016/j.cub.2016.02.019.
4. Patel, M.N., Carroll, R.G., Galván-Peña, S., Mills, E.L., Olden, R., Triantafilou, M., Wolf, A.I., Bryant, C.E., Triantafilou, K., and Masters, S.L. (2017). Inflammasome priming in sterile inflammatory disease. Trends Mol. Med. *23*, 165–180. https://doi.org/10.1016/j.molmed.2016.12.007.
5. Swanson, K.V., Deng, M., and Ting, J.P.-Y. (2019). The NLRP3 inflammasome: molecular activation and regulation to therapeutics. Nat. Rev. Immunol. *19*, 477–489. https://doi.org/10.1038/s41577-019-0165-0.
6. Muñoz-Planillo, R., Kuffa, P., Martínez-Colón, G., Smith, B.L., Rajendiran, T.M., and Nuñez, G. (2013). K^+ efflux is the common trigger of NLRP3 inflammasome activation by bacterial toxins and particulate matter. Immunity *38*, 1142–1153. https://doi.org/10.1016/j.immuni.2013.05.016.
7. Gaidt, M.M., and Hornung, V. (2018). The NLRP3 inflammasome renders cell death pro-inflammatory. J. Mol. Biol. *430*, 133–141. https://doi.org/10.1016/j.jmb.2017.11.013.
8. Gaidt, M.M., Ebert, T.S., Chauhan, D., Schmidt, T., Schmid-Burgk, J.L., Rapino, F., Robertson, A.A.B., Cooper, M.A., Graf, T., and Hornung, V. (2016). Human monocytes engage an alternative inflammasome pathway. Immunity *44*, 833–846.
9. Groß, C.J., Mishra, R., Schneider, K.S., Médard, G., Wettmarshausen, J., Dittlein, D.C., Shi, H., Gorka, O., Koenig, P.A., Fromm, S., et al. (2016). K(+) efflux-independent NLRP3 inflammasome activation by small molecules targeting mitochondria. Immunity *45*, 761–773. https://doi.org/10.1016/j.immuni.2016.08.010.
10. Chen, J., and Chen, Z.J. (2018). PtdIns4P on dispersed trans-Golgi network mediates NLRP3 inflammasome activation. Nature *564*, 71–76. https://doi.org/10.1038/s41586-018-0761-3.
11. Hornung, V., and Latz, E. (2010). Critical functions of priming and lysosomal damage for NLRP3 activation. Eur. J. Immunol. *40*, 620–623. https://doi.org/10.1002/eji.200940185.
12. Bauernfeind, F.G., Horvath, G., Stutz, A., Alnemri, E.S., MacDonald, K., Speert, D., Fernandes-Alnemri, T., Wu, J., Monks, B.G., Fitzgerald, K.A., et al. (2009). Cutting edge: NF-kappaB activating pattern recognition and cytokine receptors license NLRP3 inflammasome activation by regulating NLRP3 expression. J. Immunol. *183*, 787–791. https://doi.org/10.4049/jimmunol.0901363.
13. Kahlenberg, J.M., Lundberg, K.C., Kertesy, S.B., Qu, Y., and Dubyak, G.R. (2005). Potentiation of caspase-1 activation by the P2X7 receptor is dependent on TLR signals and requires NF-κB-driven protein synthesis. J. Immunol. *175*, 7611–7622. https://doi.org/10.4049/jimmunol.175.11.7611.
14. Juliana, C., Fernandes-Alnemri, T., Kang, S., Farias, A., Qin, F., and Alnemri, E.S. (2012). Non-transcriptional priming and deubiquitination regulate NLRP3 inflammasome activation. J. Biol. Chem. *287*, 36617–36622. https://doi.org/10.1074/jbc.M112.407130.
15. Lin, K.-M., Hu, W., Troutman, T.D., Jennings, M., Brewer, T., Li, X., Nanda, S., Cohen, P., Thomas, J.A., and Pasare, C. (2014). IRAK-1 bypasses priming and directly links TLRs to rapid NLRP3 inflammasome activation. Proc. Natl. Acad. Sci. USA *111*, 775–780.
16. Bauernfeind, F., Niepmann, S., Knolle, P.A., and Hornung, V. (2016). Aging-associated TNF production primes inflammasome activation and NLRP3-related metabolic disturbances. J. Immunol. *197*, 2900–2908. https://doi.org/10.4049/jimmunol.1501336.
17. Shim, D.-W., and Lee, K.-H. (2018). Posttranslational regulation of the NLR family pyrin domain-containing 3 inflammasome. Front. Immunol. *9*, 1054. https://doi.org/10.3389/fimmu.2018.01054.
18. Barry, R., John, S.W., Liccardi, G., Tenev, T., Jaco, I., Chen, C.H., Choi, J., Kasperkiewicz, P., Fernandes-Alnemri, T., Alnemri, E., et al. (2018). SUMO-mediated regulation of NLRP3 modulates inflammasome activity. Nat. Commun. *9*, 3001. https://doi.org/10.1038/s41467-018-05321-2.
19. McKee, C.M., and Coll, R.C. (2020). NLRP3 inflammasome priming: a riddle wrapped in a mystery inside an enigma. J. Leukoc. Biol. *108*, 937–952. https://doi.org/10.1002/JLB.3MR0720-513R.
20. Schmid-Burgk, J.L., Chauhan, D., Schmidt, T., Ebert, T.S., Reinhardt, J., Endl, E., and Hornung, V. (2015). A genome-wide CRISPR screen identifies NEK7 as an essential component of NLRP3 inflammasome activation. J. Biol. Chem. *291*, 103–109. https://doi.org/10.1074/jbc.C115.700492.
21. He, Y., Zeng, M.Y., Yang, D., Motro, B., and Nuñez, G. (2016). NEK7 is an essential mediator of NLRP3 activation downstream of potassium efflux. Nature *530*, 354–357. https://doi.org/10.1038/nature16959.
22. Shi, H., Wang, Y., Li, X., Zhan, X., Tang, M., Fina, M., Su, L., Pratt, D., Bu, C.H., Hildebrand, S., et al. (2016). NLRP3 activation and mitosis are mutually exclusive events coordinated by NEK7, a new inflammasome component. Nat. Immunol. *17*, 250–258. https://doi.org/10.1038/ni.3333.
23. Sharif, H., Wang, L., Wang, W.L., Magupalli, V.G., Andreeva, L., Qiao, Q., Hauenstein, A.V., Wu, Z., Nuñez, G., Mao, Y., and Wu, H. (2019). Structural mechanism for NEK7-licensed activation of NLRP3 inflammasome. Nature *570*, 338–343. https://doi.org/10.1038/s41586-019-1295-z.

24. He, Y., Hara, H., and Nuñez, G. (2016). Mechanism and regulation of NLRP3 inflammasome activation. Trends Biochem. Sci. *41*, 1012–1021. https://doi.org/10.1016/j.tibs.2016.09.002.

25. Takata, K., Kozaki, T., Lee, C.Z.W., Thion, M.S., Otsuka, M., Lim, S., Utami, K.H., Fidan, K., Park, D.S., Malleret, B., et al. (2017). Induced-pluripotent-stem-cell-derived primitive macrophages provide a platform for modeling tissue-resident macrophage differentiation and function. Immunity *47*. 183–198.e6. https://doi.org/10.1016/j.immuni.2017.06.017.

26. Rapino, F., Robles, E.F., Richter-Larrea, J.A., Kallin, E.M., Martinez-Climent, J.A., and Graf, T. (2013). C/EBPα induces highly efficient macrophage transdifferentiation of B lymphoma and leukemia cell lines and impairs their tumorigenicity. Cell Rep. *3*, 1153–1163. https://doi.org/10.1016/j.celrep.2013.03.003.

27. Gaidt, M.M., Ebert, T.S., Chauhan, D., Ramshorn, K., Pinci, F., Zuber, S., O'Duill, F., Schmid-Burgk, J.L., Hoss, F., Buhmann, R., et al. (2017). The DNA inflammasome in human myeloid cells is initiated by a STING-cell death program upstream of NLRP3. Cell *171*. 1110–1124.e18. https://doi.org/10.1016/j.cell.2017.09.039.

28. Coll, R.C., Robertson, A.A.B., Chae, J.J., Higgins, S.C., Muñoz-Planillo, R., Inserra, M.C., Vetter, I., Dungan, L.S., Monks, B.G., Stutz, A., et al. (2015). A small-molecule inhibitor of the NLRP3 inflammasome for the treatment of inflammatory diseases. Nat. Med. *21*, 248–255.

29. Fernandes-Alnemri, T., Kang, S., Anderson, C., Sagara, J., Fitzgerald, K.A., and Alnemri, E.S. (2013). Cutting edge: TLR signaling licenses IRAK1 for rapid activation of the NLRP3 inflammasome. J. Immunol. *191*, 3995–3999. https://doi.org/10.4049/jimmunol.1301681.

30. Kang, S., Fernandes-Alnemri, T., Rogers, C., Mayes, L., Wang, Y., Dillon, C., Roback, L., Kaiser, W., Oberst, A., Sagara, J., et al. (2015). Caspase-8 scaffolding function and MLKL regulate NLRP3 inflammasome activation downstream of TLR3. Nat. Commun. *6*, 7515. https://doi.org/10.1038/ncomms8515.

31. DeLaney, A.A., Berry, C.T., Christian, D.A., Hart, A., Bjanes, E., Wynosky-Dolfi, M.A., Li, X., Tummers, B., Udalova, I.A., Chen, Y.H., et al. (2019). Caspase-8 promotes c-Rel-dependent inflammatory cytokine expression and resistance against Toxoplasma gondii. Proc. Natl. Acad. Sci. USA *116*, 11926–11935. https://doi.org/10.1073/pnas.1820529116.

32. Mercurio, F., Zhu, H., Murray, B.W., Shevchenko, A., Bennett, B.L., Li, J., Young, D.B., Barbosa, M., Mann, M., Manning, A., and Rao, A. (1997). IKK-1 and IKK-2: cytokine-activated IkappaB kinases essential for NF-kappaB activation. Science *278*, 860–866. https://doi.org/10.1126/science.278.5339.860.

33. Posor, Y., Jang, W., and Haucke, V. (2022). Phosphoinositides as membrane organizers. Nat. Rev. Mol. Cell Biol. Published online May 19, 2022. https://doi.org/10.1038/s41580-022-00490-x.

34. Hoffman, H.M., Mueller, J.L., Broide, D.H., Wanderer, A.A., and Kolodner, R.D. (2001). Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. Nat. Genet. *29*, 301–305. https://doi.org/10.1038/ng756.

35. Hogquist, K.A., Nett, M.A., Unanue, E.R., and Chaplin, D.D. (1991). Interleukin 1 is processed and released during apoptosis. Proc. Natl. Acad. Sci. USA *88*, 8485–8489. https://doi.org/10.1073/pnas.88.19.8485.

36. Mariathasan, S., Newton, K., Monack, D.M., Vucic, D., French, D.M., Lee, W.P., Roose-Girma, M., Erickson, S., and Dixit, V.M. (2004). Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. Nature *430*, 213–218. https://doi.org/10.1038/nature02664.

37. Groslambert, M., and Py, B.F. (2018). Spotlight on the NLRP3 inflammasome pathway. J. Inflamm. Res. *11*, 359–374. https://doi.org/10.2147/JIR.S141220.

38. Nanda, S.K., Prescott, A.R., Figueras-Vadillo, C., and Cohen, P. (2021). IKKbeta is required for the formation of the NLRP3 inflammasome. EMBO Rep. *22*, e50743. https://doi.org/10.15252/embr.202050743.

39. Salem, H., Rachmin, I., Yissachar, N., Cohen, S., Amiel, A., Haffner, R., Lavi, L., and Motro, B. (2010). Nek7 kinase targeting leads to early mortality, cytokinesis disturbance and polyploidy. Oncogene *29*, 4046–4057. https://doi.org/10.1038/onc.2010.162.

40. Rauch, I., Tenthorey, J.L., Nichols, R.D., Al Moussawi, K., Kang, J.J., Kang, C., Kazmierczak, B.I., and Vance, R.E. (2016). NAIP proteins are required for cytosolic detection of specific bacterial ligands in vivo. J. Exp. Med. *213*, 657–665. https://doi.org/10.1084/jem.20151809.

41. Cavlar, T., Deimling, T., Ablasser, A., Hopfner, K.-P., and Hornung, V. (2013). Species-specific detection of the antiviral small-molecule compound CMA by STING. EMBO J. *32*, 1440–1450. https://doi.org/10.1038/emboj.2013.86.

42. Camargo Ortega, G., Falk, S., Johansson, P.A., Peyre, E., Broix, L., Sahu, S.K., Hirst, W., Schlichthaerle, T., De Juan Romero, C., Draganova, K., et al. (2019). The centrosome protein AKNA regulates neurogenesis via microtubule organization. Nature *567*, 113–117. https://doi.org/10.1038/s41586-019-0962-4.

43. Franklin, B.S., Bossaller, L., De Nardo, D., Ratter, J.M., Stutz, A., Engels, G., Brenker, C., Nordhoff, M., Mirandola, S.R., Al-Amoudi, A., et al. (2014). The adaptor ASC has extracellular and 'prionoid' activities that propagate inflammation. Nat. Immunol. *15*, 727–737.

44. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784. https://doi.org/10.1038/nmeth.3047.

45. Schmidt, T., Schmid-Burgk, J.L., and Hornung, V. (2015). Synthesis of an arrayed sgRNA library targeting the human genome. Sci. Rep. *5*, 14987. https://doi.org/10.1038/srep14987.

46. Schmid-Burgk, J.L., Schmidt, T., Gaidt, M.M., Pelka, K., Latz, E., Ebert, T.S., and Hornung, V. (2014). OutKnocker: a web tool for rapid and simple genotyping of designer nuclease edited cell lines. Genome Res. *24*, 1719–1723. https://doi.org/10.1101/gr.176701.114.

47. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100. https://doi.org/10.1186/gb-2006-7-10-r100.

48. Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H., and Valen, E. (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. *47*, W171–W174. https://doi.org/10.1093/nar/gkz365.

49. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372. https://doi.org/10.1038/nbt.1511.

50. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740. https://doi.org/10.1038/nmeth.3901.

51. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.

52. Kutner, R.H., Zhang, X.-Y., and Reiser, J. (2009). Production, concentration and titration of pseudotyped HIV-1-based lentiviral vectors. Nat. Protoc. *4*, 495–505. https://doi.org/10.1038/nprot.2009.22.

53. Dull, T., Zufferey, R., Kelly, M., Mandel, R.J., Nguyen, M., Trono, D., and Naldini, L. (1998). A third-generation lentivirus vector with a conditional packaging system. J. Virol. *72*, 8463–8471. https://doi.org/10.1128/JVI.72.11.8463-8471.1998.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-Caspase-1 (p20) (human), mAb (Bally-1)	AdipoGen, San Diego, CA	Cat# AG-20B-0048-C100
anti-Caspase-1 (p20) (mouse), mAb (Casper-1)	AdipoGen	Cat# AG-20B-0042-C100
anti-NEK7	Abcam, Cambridge, UK	Cat# ab133514
anti-NLRP3/NALP3, mAb (Cryo-2)	AdipoGen	Cat# AG-20B-0014-C100
anti-Human IL-1 beta /IL-1F2	R&D Systems Inc, Minneapolis, MN	Cat# AF-201-NA
Chemicals, peptides, and recombinant proteins		
1-Thioglycerol (MTG)	Sigma-Aldrich, St. Louis, MO	Cat# M6145
Accutase	Stemcell Technologies, Vancouver, Canada	Cat# 07920
Adenosine 5′-triphosphate disodium salt hydrate	Sigma-Aldrich	Cat# A6419
Ascorbic Acid	Sigma-Aldrich	Cat# A4544-100G
B-27 supplement	Thermo Fisher Scientific, Waltham, MA	Cat# 17504-001
Blasticidin S HCl (10 mg/ml)	Thermo Fisher Scientific	Cat# A1113903
BSA	GE Healthcare, Chicago, IL	Cat# SH30574.01
CHIR99021	Miltenyi Biotec, Bergisch Gladbach, Germany	Cat# 130-103-926
Cycloheximide	Carl Roth, Karlsruhe, Germany	Cat# 8682.1
Doxycycline hyclate	Sigma-Aldrich	Cat# D9891-1G
Geltrex	Thermo Fisher Scientific	Cat# A1413302
GeneJuice	Merck, Darmstadt, Germany	Cat# 70967-3
Ham's F12 nutrient mix	Thermo Fisher Scientific	Cat# 21765029
Herring Testis(HT)-DNA sodium salt	Sigma Aldrich	Cat# D6898
Hoechst-33342	Sigma-Aldrich	Cat# B2261-25MG
Human CSF-1 (M-CSF) (iPSC differentiation)	R&D Systems	Cat# 216-MC-005
Human Transferrin	Roche, Basel, Switzerland	Cat# 10-652-202-001
IMDM with GlutaMAX	Thermo Fisher Scientific	Cat# 31980022
Imiquimod (R837)	Invivogen	Cat# tlrl-imq
L-Glutamine	Thermo Fisher Scientific	Cat# 25030024
LFn-YscF	Rauch et al.[40]	N/A
Lipofectamine 2000 Transfection Reagent	Thermo Fisher Scientific	Cat# 11668019
LPS-EB Ultrapure	Invivogen, San Diego, CA	Cat# tlrl-3pelps
LysC	Wako	Cat# 12902541
MCC950	Sigma-Aldrich	Cat# PZ0280
MitoTracker DeepRed	Thermo Fisher Scientific	Cat# M22426
mTeSR1	Stemcell Technologies	Cat# 85850
N-2 Supplement	Thermo Fisher Scientific	Cat# 17502048
Nigericin sodium salt	Sigma-Aldrich	Cat# N7143
Pam3CSK4	Invivogen	Cat# tlrl-pms
Phorbol 12-myristate 13-acetate	ENZO Life Sciences, Farmingdale, NY	Cat# BML-PE160-0005
Protective antigen (pA)	Biotrend, Cologne, Germany	Cat# LL-171E
Puromycin Dihydrochloride	Carl Roth	Cat# 0240.4

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R848	Invivogen	Cat# tlrl-r848-5
Recombinant Human BMP-4	R&D Systems	Cat# 314-BP-010
Recombinant Human CSF-1 (M-CSF) (BlaER1 differentiation)	Recombinantly produced	N/A
Recombinant Human DKK-1	R&D Systems	Cat# 5439-DK-010
Recombinant Human FGF2	R&D Systems	Cat# 233-FB-025
Recombinant Human IL-3	R&D Systems	Cat# 203-IL-010
Recombinant Human IL-3 (BLaER1 differentiation)	Recombinantly produced	N/A
Recombinant Human IL-6	R&D Systems	Cat# 206-IL-010
Recombinant Human SCF	R&D Systems	Cat# 255-SC-010
Recombinant Human VEGF	R&D Systems	Cat# 293-VE-010
ROCK Inhibitor Y-27632	Stemcell Technologies	Cat# 72302
Stempro-34 SFM	Thermo Fisher Scientific	Cat# 10639-011
Takinib	Selleck Chemicals, Houston, TX	Cat# S8663
TPCA-1	R&D Systems	Cat# 2559/10
Trypsin	Sigma-Aldrich	Cat# T6567
β-Estradiol	Sigma-Aldrich	Cat# E8875
Critical commercial assays		
Human IL-1β ELISA Set II	BD Biosciences, San José, CA	Cat# 557953
Human Total IL-18 DuoSet ELISA	R&D Systems	Cat# DY318-05
MiSeq Reagent Kit v2, 300 Cycles	Illumina, San Diego, CA	Cat# MS-102-2002
Mouse CXCL10/IP-10/CRG-2 DuoSet ELISA	R&D Systems	Cat# DY466
Mouse TNF (Mono/Mono) ELISA Set II	BD Biosciences	Cat# 558534
OptEIA Human IL-6 ELISA Set	BD Biosciences	Cat# 555220
OptEIA Mouse IL-1β Elisa Set	BD Biosciences	Cat# 559603
Pierce LDH Cytotoxicity Assay Kit	Thermo Fisher Scientific	Cat# 88954
Deposited data		
Mass spectrometry data of Figure 6E	This study	PRIDE: PXD035302
Immunoblot source data and raw numerical data used to plot the figures	This study	Mendeley data: https://doi.org/10.17632/h7vc8hnb7j.1
Experimental models: Cell lines		
BLaER1	Rapino et al.[26]	N/A
HEK-293T	Cavlar et al.[41]	N/A
iPSC	Camargo Ortega et al.[42]	N/A
Mouse Macrophages, *Nlrp3*, *Asc-CFP*, *Cas9*-expressing	Franklin et al.[43]	N/A
THP-1	ATCC, Manassas, VA	Cat# TIB-202
Target sites of sgRNAs used in this study		
hsMAP3K7	GTAAACACCAACTCATTGCG**TGG**	
hsNEK6	GTCTTTTCGCTGCTCGCTGG**CGG**	
hsNEK7	ATTACAGAAGGCCTTACGAC**CGG**	
hsNLRP3	GCTAATGATCGACTTCAATG**GGG**	
hsIKBKB	ATGAAGGTATCTAAGCGCAG**AGG**	
mmMyd88	GGTTCAAGAACAGCGATAGG**CGG**	
mmNek7	GTCTCTTGGATGGAGTGC**CGG**	
mmNlrp3	CCTCTCTGCTCATAACGACG**AGG**	
mmTicam1	GTACAGGCGAGCCACCGTCC**AGG**	

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
mmTlr4	GATCTACTCGAGTCAGAATG**AGG**	
mmPycard	GTGCAACTGCGAGAAGGCTA**TGG**	
Recombinant DNA		
LentiCas9-Blast	Sanjana et al.[44]	N/A
LentiGuide-Puro	Sanjana et al.[44]	N/A
pBabe-U6-sgRNA-Cas9	Schmidt et al.[45]	N/A
pBlast-hsNEK7	This study	N/A
pBlast-mCherry-OSBP(PH)	This study	N/A
pBlast-mmNek7	This study	N/A
pLIX-hsNLRP3	This study	N/A
pLIX-mmNlrp3	This study	N/A
pLIX-mVenus-mmNlrp3	This study	N/A
pLIX-mVenus-mmNlrp3(4KA)	This study	N/A
pLIX-mVenus-mmNlrp3(4KA-OSBP(PH))	This study	N/A
pLK0.1-sgRNA-CMV-GFP	Schmid-Burgk et al.[46]	N/A
pRP-Asc-RFP	This study	N/A
pRZ-CMV-Cas9	Schmidt et al.[45]	N/A
Software and algorithms		
CellProfiler 3.1.5	Carpenter et al.[47]	https://cellprofiler.org
CHOPCHOP	Labu et al.[48]	https://chopchop.cbu.uib.no
MaxQuant 2.0.3	Cox and Mann[49]	https://maxquant.org
Outknocker	Schmid-Burgk et al.[46]	http://www.outknocker.org
Perseus	Tyanova et al.[50]	https://maxquant.org/perseus/
Prism 9.0	GraphPad, San Diego, CA	https://www.graphpad.com/scientific-software/prism/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Veit Hornung (hornung@genzentrum.lmu.de).

Materials availability

All unique reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

Mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (PRIDE: PXD035302) and are publicly available as of the date of publication.

Immunoblot source data and raw numerical data used to plot the figures were deposited on Mendeley Data: https://doi.org/10.17632/h7vc8hnb7j.1

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

This paper does not report original code.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

BLaER1 cells

BLaER1 cells (female) were cultivated in RPMI supplemented with 10 % FCS, 1 mM pyruvate, 100 U/ml penicillin and 100 μg/ml streptomycin at 37 °C and 5 % CO_2. BLaER1 cells were differentiated in medium containing 10 ng/ml hrIL-3, 10 ng/ml hrCSF-1 (M-CSF) and 100 nM β-estradiol for 5-6 days. In the course of these studies, we serendipitously identified that BLaER1 cells express transcripts of SMRV (squirrel monkey retrovirus) and subsequent experiments confirmed that BLaER1 cells harbor the SMRV proviral genome. Testing early passages of BLaER1 cells by Dr. Thomas Graf (personal communication) confirmed that the parental BLaER1 cell line[26] is positive for SMRV. Of note, extensive characterization of BLaER1

monocytes in comparison to other human myeloid cells has not provided any indication that SMRV positivity would impact on the functionality of these cells as myeloid cells. Samples of other cell lines used in this work were confirmed to be free of SMRV by PCR. All BLaER1 cell experiments were conducted on a *CASP4*$^{-/-}$ background (herein referred to as control).

THP-1 cells

THP-1 cells (male) were obtained from ATCC and cultivated in RPMI supplemented with 10 % FCS, 1 mM pyruvate, 100 U/ml penicillin and 100 μg/ml streptomycin at 37 °C and 5 % CO_2. THP-1 cells were differentiated by adding 100 ng/ml PMA to the medium for 18 hours, rinsed off with ice-cold PBS and replated for experiments.

mmMacs and J774 mouse macrophages

Mouse macrophages were cultivated in DMEM supplemented with 10 % FCS, 1 mM pyruvate 100 U/ml penicillin and 100 μg/ml streptomycin at 37 °C and 5 % CO_2. mmMacs were detached for passaging with 0.05 % Trypsin at 37 °C for 15 minutes after one PBS wash and then rinsed off with DMEM. J774 cells were passaged by scraping in 5 ml fresh DMEM and transferred to new flasks.

hiPSC, hiPS-Macs cell culture

Human induced pluripotent stem cells (hiPSCs) used to make NEK7$^{-/-}$ hiPSCs were kindly provided by Adam O'Neill and Magdalena Götz.[42] hiPSCs for IKBKB$^{-/-}$ were purchased from XCell Science. hiPSCs were cultivated on Geltrex-coated plates in complete mTeSR1 Medium at 37 °C and 5 % CO_2 and detached for passaging using 1.5 ml Accutase for 5 minutes at 37 °C after a PBS wash. After passaging, cells were cultivated in the presence of 5 μM ROCK-Inhibitor overnight.

Differentiation of hiPSCs into hiPS-Macs

Differentiation into iPS-Macs was achieved as described previously.[25] Briefly, 150,000 hiPSC were plated into a one well of a Geltrex-coated 6-well plate and differentiated in StemPro base medium with StemPro Supplement, 200 μg/ml human transferrin, 2 mM glutamine, 0.45 mM MTG and 0.5 mM ascorbic acid (= StemPro medium, ascorbic acid was added just before use) by stimulation with 50 ng/ml VEGF, 5 ng/ml BMP-4 and 2 μM CHIR99021 at 5 % oxygen for two days, followed by two days of stimulation with 50 ng/ml VEGF, 5 ng/ml BMP-4 and 20 ng/ml FGF2. From day four, StemPro medium was supplemented with 15 ng/ml VEGF and 5 ng/ml FGF2. Starting at day six, 10 ng/ml VEGF, 10 ng/ml FGF2, 50 ng/ml SCF, 30 ng/ml DKK-1, 10 ng/ml IL-6 and 20 ng/ml IL-3 were added to StemPro medium until day ten. From day eight, cells were cultivated under normoxic conditions. From day twelve, 10 ng/ml FGF2, 50 ng/ml SCF, 10 ng/ml IL-6 and 20 ng/ml IL-3 were added to StemPro medium. Starting at day sixteen, cells were cultivated in 75 % IMDM with 25 % F12 supplement, N2 supplement, B-27 supplement, 0.05 % BSA and 100 U/ml penicillin and 100 μg/ml streptomycin (= SF-Diff medium) supplemented with 50 ng/ml rhCSF-1 (M-CSF) at least until day 28. Culture medium was exchanged as necessary, but at least every two days. After differentiation, hiPS-Macs were carefully harvested from the supernatant, spun down and replated in RPMI with 10 % FCS, 1 mM Pyruvate, 100 U/ml Penicillin and 100 μg/ml Streptomycin for experiments.

HEK-293T cells

HEK-293T cells were cultivated in DMEM with 10 % FCS, 1 mM pyruvate, 100 U/ml penicillin and 100 μg/ml streptomycin at 37 °C and 5 % CO_2. For passaging, cells were washed with PBS once and then incubated with 0.05 % Trypsin at 37° for 5 minutes. Cells were then rinsed off with DMEM.

METHOD DETAILS

Trans-Golgi network imaging

J774 macrophages expressing mVenus-mmNlrp3 and the PH-domain of hsOSBP (OSBP-PH) fused to mCherry were plated in ibidi 8-well slides (100,000 per well in 200 μl of DMEM) and imaged on a Nikon Eclipse Ti spinning disk confocal microscope with 100x magnification on the following day. Results were manually quantified from at least 10 randomly selected areas per condition per replicate using FIJI.[51] For nuclear staining, Hoechst-33342 was diluted to a final concentration of 10 μg/ml.

ASC speck imaging

ASC specks in transiently transfected HEK-293T cells were imaged 24 hours after transfection on a Leica Hi8 epifluorescence microscope using 10x magnification. Specks were quantified with CellProfiler.[47]

Immunoblotting

Cells were lysed at approximately 5 Mio/ml in 1x Lämmli Buffer and boiled for 5 minutes at 95 °C. For precipitation of total protein from supernatants, stimulations were done in medium containing 3% FCS. Precipitation of total protein from supernatants was achieved by combining 700 μl of supernatant with 700 μl MeOH and 150 μl of $CHCl_3$. Samples were spun down at 20.000 g for 20 minutes, and the upper phase was discarded. Again, 700 μl MeOH were added and samples were centrifuged at 20.000 g for 20 minutes. The pellet was then dried and resuspended in 100 μl 1x Lämmli buffer and boiled at 95 °C for 5 minutes. Samples were run on 12% SDS-PAGE gels at 150 V for 85 minutes and were subsequently transferred onto a nitrocellulose

membrane at 100 V for 75 minutes at 4 °C. Membranes were then blocked in 5 % milk for 1 hour at room temperature. Primary and secondary antibodies were diluted in 1-5 % milk.

ELISA and LDH assay

LDH assays were done on supernatants immediately after experiments. Results are presented relative to a lysis control from the same experiment with the values of unstimulated controls subtracted as background. ELISAs were done according to manufacturer's instructions on supernatants stored at -20 °C.

Stimulation of immune signaling

NLRP3 was primed as indicated with 1 μg/ml Pam_3CSK_4 or 200 ng/ml LPS. NLRP3 was activated with 5 mM ATP or Nigericin at 6.5 μM (BLaER1 cells) or 10 μM (all other cells) as indicated. To activate the AIM2 inflammasome 400 ng HT-DNA were transfected into a 96-well with 0.5 μl Lipofectamine in 50 μl OptiMEM by incubating OptiMEM and Lipofectamine for 5 minutes followed by 20 minutes of incubation of the Lipofectamine-DNA mix in OptiMEM and dropwise addition of the mix to the cells. For immunoblots, transfections were done in a 12-well format. The amount of Lipofectamine and HT-DNA was scaled accordingly by well area. The NAIP-NLRC4 inflammasome was activated with an anthrax toxin lethal factor fused to the *Burkholderia* T3SS needle protein (LFn-YscF, 0.025 μg/ml), which was delivered into cells with protective antigen (pA, 0.25 μg/ml).[40] If not otherwise indicated, cells were stimulated with this construct (herein referred to as Needle Tox) for 2 hours.

Inhibition of translation

For mmMacs, cycloheximide (CHX) was added to the medium 30 minutes before stimulation to a final concentration of 10 μg/ml. For BLaER1 cells, CHX was added to the medium simultaneously with LPS at the indicated concentrations in the range of 1-10 μg/ml.

Doxycyclin-inducible gene expression

In BLaER1 cells and J774 *Pycard*$^{-/-}$ cells transduced with pLIX-Puro derived vectors, gene expression was induced by adding medium to a final concentration of 1 μg/ml doxycycline for the last 24 hours of differentiation. J774 cells transduced with Nlrp3 variants for analysis of caspase-1 processing were stimulated with 1 μg/ml doxycycline for 18 hours before stimulation for inflammasome activation.

Inhibition of TAK1, IKKβ and NLRP3

Takinib was added to a final concentration of 50 μM as indicated. TPCA-1 was used at 5 μM final concentration as indicated. MCC950 was added as indicated to a final concentration of 10 μM.

Inhibition and induction of K^+ efflux

To block K^+ efflux, Potassium chloride (KCl) was added to medium together with the priming stimulus to the indicated final concentrations. The osmolarity of the medium was kept constant over all conditions. To induce K^+ efflux, cells were stimulated in sterile Hank's balanced salt solution with (140 mM NaCl, 5 mM KCl, 1.3 mM $CaCl_2$, 1.0 mM $MgSO_4$, 10 mM HEPES (pH 7.5), 5.5 mM glucose) or without potassium (145 mM NaCl, 1.3 mM $CaCl_2$, 1.0 mM $MgSO_4$, 10 mM HEPES (pH 7.5), 5.5 mM glucose) with 10% FCS as described before.[10]

Sucrose gradient fractionation

For the fractionation experiment, *Nlrp3*$^{-/-}$ x *Pycard*$^{-/-}$ J774 cells stably transduced with pLI-mVenus-mmNLRP3, pBlast-AUG-OSBP(PH)-mCherry were used.

Two days prior to stimulation, 1×10^7 cells were plated per 15 cm dish, using 2 dishes per condition (unstimulated, LPS, TPCA-1 > LPS). 18-20 hours prior to stimulation, doxycycline was added to a final concentration of 1 μg/ml to induce expression of mVenus-mmNLRP3. As indicated, cells were pre-treated with 5 μM TPCA-1 for 30 minutes. Subsequently, cells were stimulated with 200 ng/ml LPS for 30 minutes. Cells were washed once with PBS and scraped using 500 μL of ice-cold isotonic buffer (0.25 M sucrose, 10 mM Tris·HCl (pH 7.5), 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT) supplemented with protease inhibitor . Then, cells were homogenized by performing 30 strokes with a 29G needle (VWR, BDAM324891). Lysates were centrifuged at 1000×g for 5 minutes at 4°C to remove nuclei and any remaining cells. The resulting supernatant was centrifuged at 5000×g for 10 minutes at 4°C to obtain the heavy membrane fraction (pellet, P5). The resulting supernatant was centrifuged at 100,000×g for 20 minutes at 4°C in a TLA 120.2 rotor (Beckman Coulter) to obtain the light membrane fraction (pellet, P100) and the cytosolic fraction (supernatant, S100). The fractions P5 and P100 were washed once with isotonic buffer, pelleted repeating the centrifugation step at 5000×g and 100,000×g, respectively, and resuspended in 500 μL isotonic buffer.

The fraction P100 was then loaded onto a 20%-60% continuous sucrose density gradient (10 mM Tris·HCl (pH 7.5), 100 mM KCl, 1.5 mM MgCl2, 1 mM DTT, and protease inhibitor cocktail). The gradients were centrifuged in an SW40Ti rotor (Beckman Coulter) at 170,085×g for two hours at 4°C and 13 fractions of 0.93 ml each were collected using a BioComp Gradient Station. 30 μL of each fraction were used for SDS-PAGE followed by immunoblotting.

Furthermore, to analyze the distribution of various organelle markers, the fractions P5, P100 and S100 were subjected to SDS-PAGE followed by immunoblotting. Protein concentrations were determined by BCA assay and adjusted between samples (unstimulated, LPS, TPCA-1 > LPS) for each of the fractions separately.

Mass spectrometry sample preparation

Sucrose gradient fractions #2 and #11 were lysed in 1% SDC with 100mM Tris-HCl. Protein amounts from each sample were adjusted to 30 μg with a BCA protein assay kit. Samples were reduced with 10mM tris(2-carboxy(ethyl)phosphine) (TCEP), alkylated with 40mM 2-chloroacetamide (CAA), and digested with trypsin and lysC (1:50, enzyme/protein, w/w) overnight. Digested peptides were desalted using SDB-RPS-stage tips. Desalted peptides were resolubilized in 5μl 2% ACN and 0.3% TFA and about 200 ng of peptides were injected into the mass spectrometer.

Samples were loaded onto 50-cm columns packed in-house with C18 1.9μM ReproSil particles (Dr. Maisch GmbH), with an EASY-nLC 1200 system (Thermo Fisher Scientific) coupled to the MS (Orbitrap Exploris 480, Thermo Fisher Scientific). A homemade column oven maintained the column temperature at 60°C. Peptides were introduced onto the column with buffer A (0.1% formic acid) and were eluted with a 120-min gradient starting at 5% buffer B (80% ACN, 0.1% formic acid) followed by a stepwise increase to 30% in 95 min, 65% in 5 min, 95% in 2×5 min and 5% in 2×5 min at a flow rate of 300 nL/min.

Samples were measured in data-dependent acquisition with a TopN MS method in which one full scan (300–1650 m/z, R=60,000 at 200m/z) at an Automatic Gain Control (AGC) target of 3×10e6 ions was first performed, followed by 15 data-dependent MS/MS scans with higher-energy collisional dissociation (AGC target 1x10e5 ions, maximum injection time at 25ms, isolation window 1.4 m/z, normalized collision energy 30%, and R=15,000 at 200 m/z). Dynamic exclusion of 30 s was enabled.

Analysis of MS samples

The MS raw files were processed in MaxQuant version 2.0.3.0[49] and fragment lists were queried against the mouse UniProt FASTA database (25,320 entries, Oct 2020) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin and lysC as proteases and a maximum of two missed cleavages.

Bioinformatics analysis of the MS data was performed using the Perseus software suite (version 1.6.7.0).[50] After filtering to remove potential contaminants, reverse hits, and proteins only identified by modification sites, the remaining summed intensities were $\log_2$-transformed. Quantified proteins were filtered for at least 2 valid values in one fraction across three biological replicates. Missing values were imputed by sampling from a normal distribution (width 0.3, downshift 1.8) and significantly up- or downregulated proteins were determined by two-sided Student's t-test (FDR < 0.05, S0 $\geq$ 1.5). To determine the systematic enrichment or de-enrichment of a select list of GOCC annotated organelles in each fraction a Fisher's exact test was performed on the significantly differentially regulated proteins between the two fractions.

Transient Transfection of HEK-293T cells

HEK-293T cells were transiently transfected with 400 ng plasmid DNA in 50 μl OptiMEM with 1 μl GeneJuice by incubating GeneJuice with OptiMEM for 5 minutes followed by 15 minutes of incubation of the DNA-GeneJuice mix in OptiMEM. DNA concentrations were kept constant across all conditions using pBluescript as stuffer DNA.

Plasmid DNA purification

Plasmid DNA was purified from *E.Coli* DH5α using a Thermo HiPure Maxiprep Kit according to manufacturer's instructions.

Preparation of lentiviral particles

Lentiviral particles were prepared according to.[52] Briefly, HEK-293T cells were transfected with 20 μg transfer plasmid, 15 μg pCMVΔ8.91 packaging plasmid and 6 μg pMD2.G VSV-G pseudotyping plasmid dish by diluting the plasmids in 1 ml 1x HBS, adding 50 μl 2.5 M Calcium chloride and gently pipetting the mix onto a 10-cm dish with approximately 6 Mio. HEK-293T cells in fresh medium. Alternatively, pMDLg/pRRE and pRSV-REV were used as packaging plasmids.[53] After 8 hours the medium was exchanged. Supernatants were harvested 48 hours later, spun down and filtered before being used to transduce target cells. Successfully transduced cells were selected with 2.5 - 5 μg/ml puromycin or 10 μg/ml blasticidin S for 48 hours, or FACSorted for fluorescence markers.

Genome editing and overexpression

sgRNA oligos were designed using CHOPCHOP[48] and cloned into expression plasmids as described previously.[44,46] BLaER1 cells were electroporated in OptiMEM with 5 μg of plasmids driving expression of Cas9 and an sgRNA on a BioRad GenePulser XCell as described previously.[45] THP-1 cells and murine macrophages were transduced with lentiviral particles driving expression of Cas9 (Lenti-Cas9-Blast[44]) or an sgRNA (LentiGuide-Puro[44]). HEK-293T cells were transiently transfected with plasmids driving expression of Cas9 or an sgRNA.

hiPS cells conditioned to grow as single clones were electroporated with Cas9-crRNA-trRNA complexes (RNPs) targeting *NEK7* on a 4D-Nucleofector (Lonza Bioscience). Grown single clones were duplicated, lysed and out-of-frame editing in *NEK7* was analyzed via deep sequencing as described previously.[46] Several *NEK7*$^{-/-}$ and *NEK7*$^{+/+}$ clones were expanded and used for experiments. To generate *IKBKB*$^{-/-}$ hiPSCs, XCL1 hiPS cells were electroporated with 0.5 μg of plasmids driving expression of Cas9 and an sgRNA

targeting *IKBKB* with a 4D-Nucleofector (Lonza Bioscience). Grown single clones were picked and the sequence of the targeted *IKBKB* region was confirmed by Sanger sequencing.

Plasmids

Cloning of genes of interest into pLIX, pRP and pFUGW backbones was performed by conventional restriction enzyme cloning. pMDLg/pRRE was a gift from Didier Trono (Addgene plasmid #12251; http://n2t.net/addgene:12251; RRID:Addgene_12251), pRSV-Rev was a gift from Didier Trono (Addgene plasmid #12253; http://n2t.net/addgene:12253; RRID:Addgene_12253), pLIX_403 (herein referred to as pLIX) was a gift from David Root (Addgene plasmid #41395; http://n2t.net/addgene:41395; RRID:Addgene_41395). LentiGuide-Puro (Addgene plasmid # 52963; http://n2t.net/addgene:52963; RRID:Addgene_52963) and lentiCas9-Blast (Addgene plasmid #52962; http://n2t.net/addgene:52962; RRID:Addgene_52962) were a gift from Feng Zhang. pTY-zeo-NLRP3(127-128-129-130 4KA)-GFP) and pTY-zeo-Flag-NLRP3(ΔKKKK OSBPPH)) were a gift from Zhijian J. Chen.[10]

QUANTIFICATION AND STATISTICAL ANALYSIS

Numbers of independent replicates (n) are reported in the respective figure legends. p-values were calculated based on two-way ANOVAs followed by Šidák's multiple comparisons test for groups containing two elements, or Tukey's test for larger groups. Dunnett's test was used wherever comparing all experimental conditions to one control instead of all other conditions was appropriate as indicated in the respective figure legends. All statistical analyses were done using GraphPad Prism 9. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ns $p \geq 0.05$.

Immunity, Volume *55*

Supplemental information

IKKβ primes inflammasome formation by recruiting NLRP3 to the *trans*-Golgi network

Niklas A. Schmacke, Fionan O'Duill, Moritz M. Gaidt, Inga Szymanska, Julia M. Kamper, Jonathan L. Schmid-Burgk, Sophia C. Mädler, Timur Mackens-Kiani, Tatsuya Kozaki, Dhruv Chauhan, Dennis Nagl, Che A. Stafford, Hartmann Harz, Adrian L. Fröhlich, Francesca Pinci, Florent Ginhoux, Roland Beckmann, Matthias Mann, Heinrich Leonhardt, and Veit Hornung

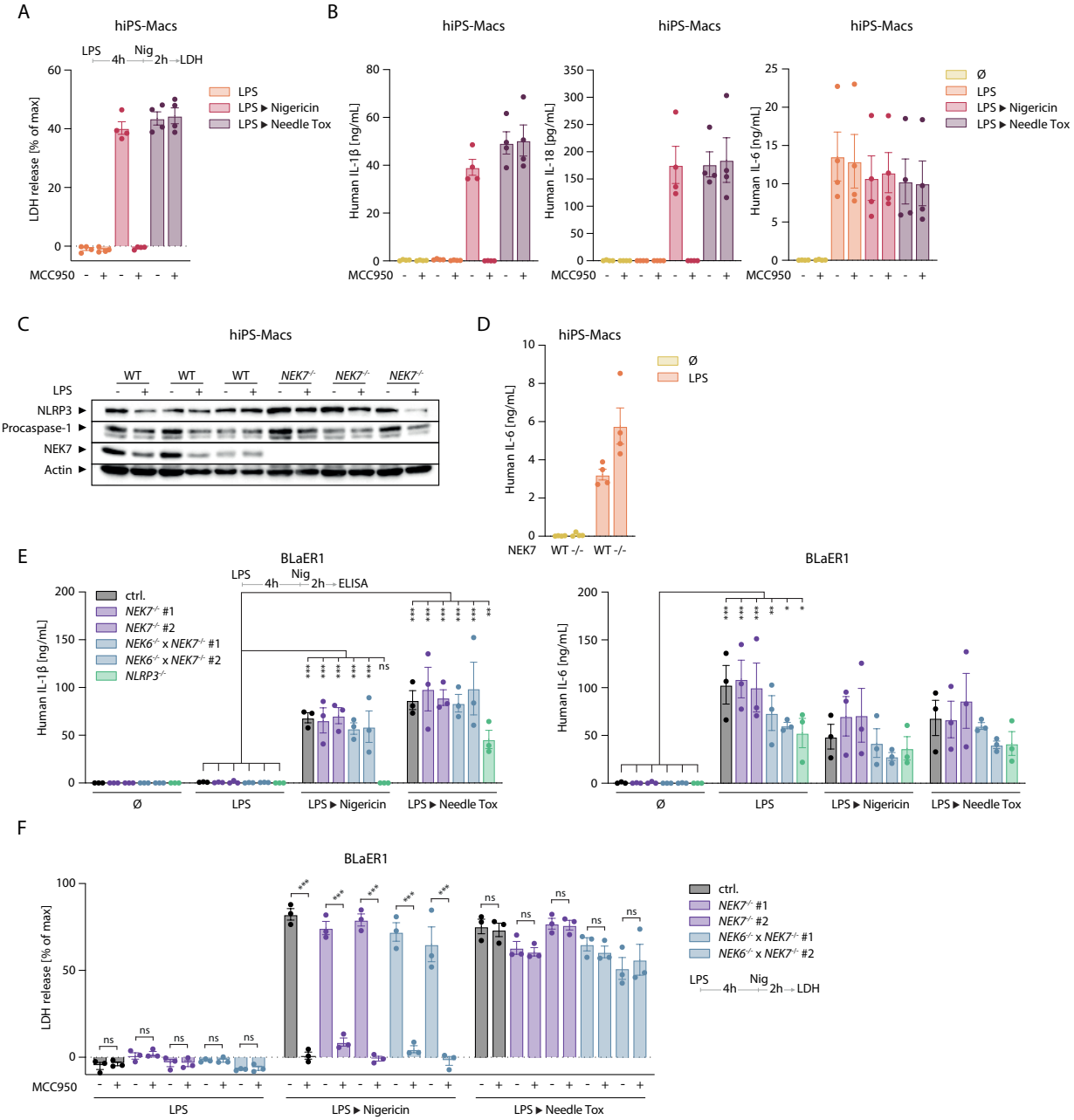

A
hiPS-Macs
LPS
4h
Nig
2h
LDH
LDH release [% of max]
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
MCC950
B
hiPS-Macs
Human IL-1β [ng/mL]
Human IL-18 [pg/mL]
Human IL-6 [ng/mL]
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
C
hiPS-Macs
WT
NEK7-/-
LPS
NLRP3
Procaspase-1
NEK7
Actin
D
hiPS-Macs
Human IL-6 [ng/mL]
NEK7
WT -/-
E
BLaER1
LPS
Nig
ELISA
Human IL-1β [ng/mL]
ctrl.
NEK7-/- #1
NEK7-/- #2
NEK6-/- x NEK7-/- #1
NEK6-/- x NEK7-/- #2
NLRP3-/-
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
F
BLaER1
LDH release [% of max]
MCC950

Figure S1

Figure S1 related to Figure 1. *Human iPSC-derived macrophages and human myeloid cell lines activate NLRP3 independently of NEK7*

(A, B) Four different clones of human iPS cells were differentiated into macrophages (hiPS-Macs), primed with LPS for 4 hours and subsequently stimulated with the inflammasome activators Nigericin (NLRP3) and Needle Tox (NAIP-NLRC4) for 2 hours in the presence of the NLRP3 inhibitor MCC950 as indicated. LDH release (A) and cytokine release (B) are depicted as mean ± SEM with dots representing individual clones.

(C) hiPS-Macs of the indicated genotypes were stimulated with LPS for 4 hours before immunoblotting for NLRP3 inflammasome components.

(D) hiPS-Macs of the indicated genotypes were treated with LPS for 6 hours before IL-6 release was measured.

(E) BLaER1 monocytes of the indicated genotypes were primed with LPS for 4 hours and subsequently stimulated with the inflammasome activators Nigericin (NLRP3) and Needle Tox (NAIP-NLRC4) for 2 hours before release of the indicated cytokines was measured by ELISA.

(F) BLaER1 monocytes of the indicated genotypes were stimulated as in (E) in the presence of the NLRP3 inhibitor MCC950.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Tukey's test (E IL-1β), Dunnett's test (E IL-6) or Šidák's test (F).

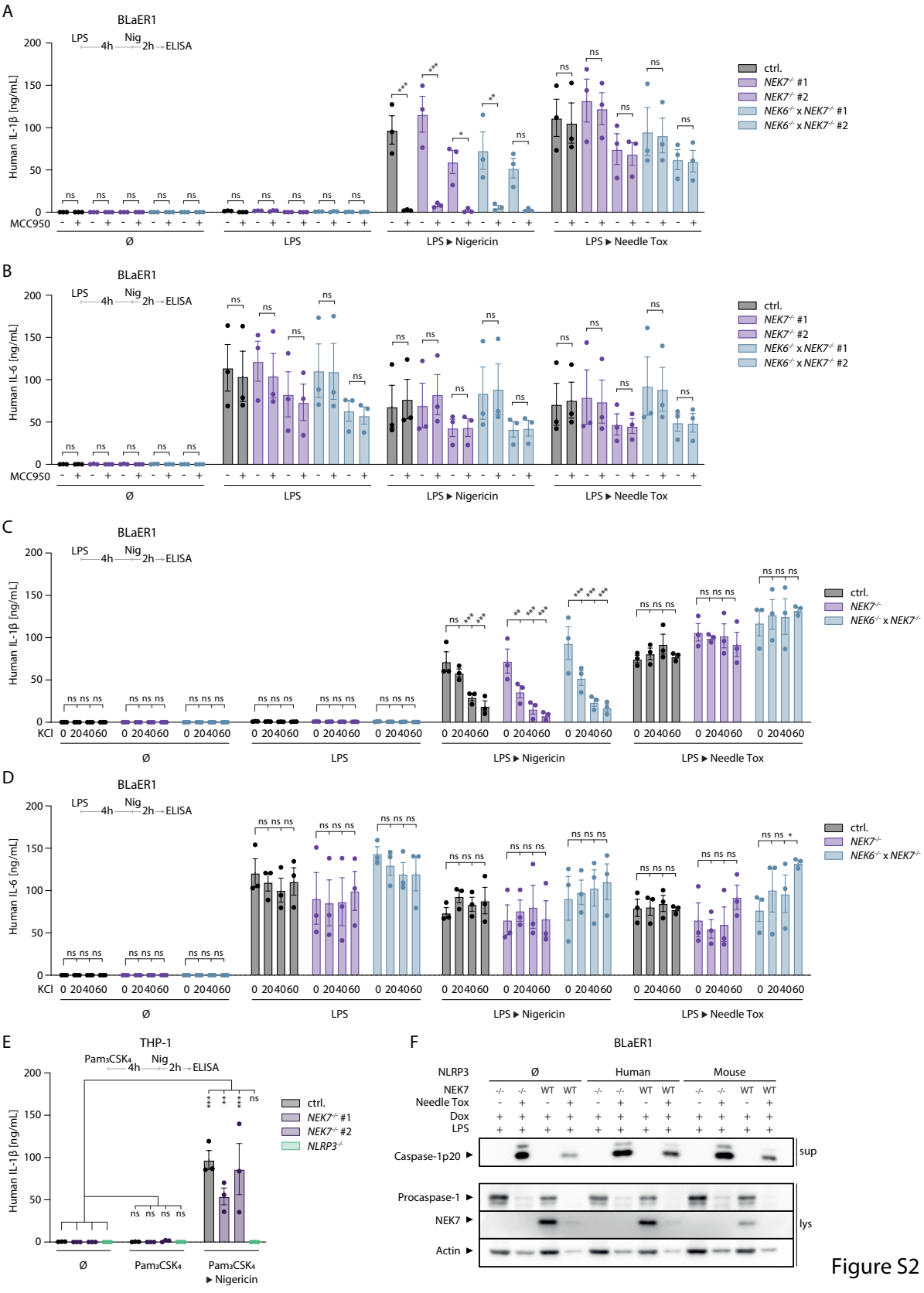

A
BLaER1
LPS 4h Nig 2h ELISA
Human IL-1β [ng/mL]
MCC950
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
ctrl.
NEK7-/- #1
NEK7-/- #2
NEK6-/- x NEK7-/- #1
NEK6-/- x NEK7-/- #2
B
BLaER1
LPS 4h Nig 2h ELISA
Human IL-6 [ng/mL]
MCC950
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
ctrl.
NEK7-/- #1
NEK7-/- #2
NEK6-/- x NEK7-/- #1
NEK6-/- x NEK7-/- #2
C
BLaER1
LPS 4h Nig 2h ELISA
Human IL-1β [ng/mL]
KCl 0 20 40 60
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
ctrl.
NEK7-/-
NEK6-/- x NEK7-/-
D
BLaER1
LPS 4h Nig 2h ELISA
Human IL-6 [ng/mL]
KCl 0 20 40 60
Ø
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox
ctrl.
NEK7-/-
NEK6-/- x NEK7-/-
E
THP-1
Pam3CSK4 4h Nig 2h ELISA
Human IL-1β [ng/mL]
Ø
Pam3CSK4
Pam3CSK4 ▶ Nigericin
ctrl.
NEK7-/- #1
NEK7-/- #2
NLRP3-/-
F
BLaER1
NLRP3
NEK7
Needle Tox
Dox
LPS
Ø
Human
Mouse
Caspase-1p20 ▶
Procaspase-1 ▶
NEK7 ▶
Actin ▶
sup
lys

Figure S2

Figure S2 related to Figure 1. *Human myeloid cell lines activating NLRP3 independently of NEK7 are sensitive to MCC950 and K^+ efflux*

(A, B) BLaER1 monocytes of the indicated genotypes were primed with LPS for 4 hours and subsequently stimulated with Nigericin or Needle Tox in the presence of 10 μM MCC950 as indicated. Release of IL-1β and IL-6 is depicted as mean ± SEM of three independent experiments.

(C, D) BLaER1 monocytes of the indicated genotypes were primed with LPS for 4 hours and subsequently stimulated with Nigericin or Needle Tox in the presence of up to 60 mM potassium chloride (KCl) as indicated.

(E) THP-1 cells of the indicated genotypes were primed with Pam_3CSK_4 for 4 hours and subsequently stimulated with Nigericin for 2 hours before release of IL-1β was measured. Dots represent individual clones. Two different sgRNAs against *NEK7* were used (#1 and #2).

(F) *NLRP3*$^{-/-}$ BLaER1 cells expressing the indicated NLRP3 orthologues under the control of a doxycycline-inducible promoter were treated with doxycycline for the last 24 hours of differentiation, primed with LPS for 4 hours and subsequently stimulated with Needle Tox for 2 hours. The same vector expressing mCherry instead of NLRP3 was used as a mock control. One representative immunoblot of three independent replicates is shown.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Šidák's test (A, B), Dunnett's test (C, D) or Tukey's test (E).

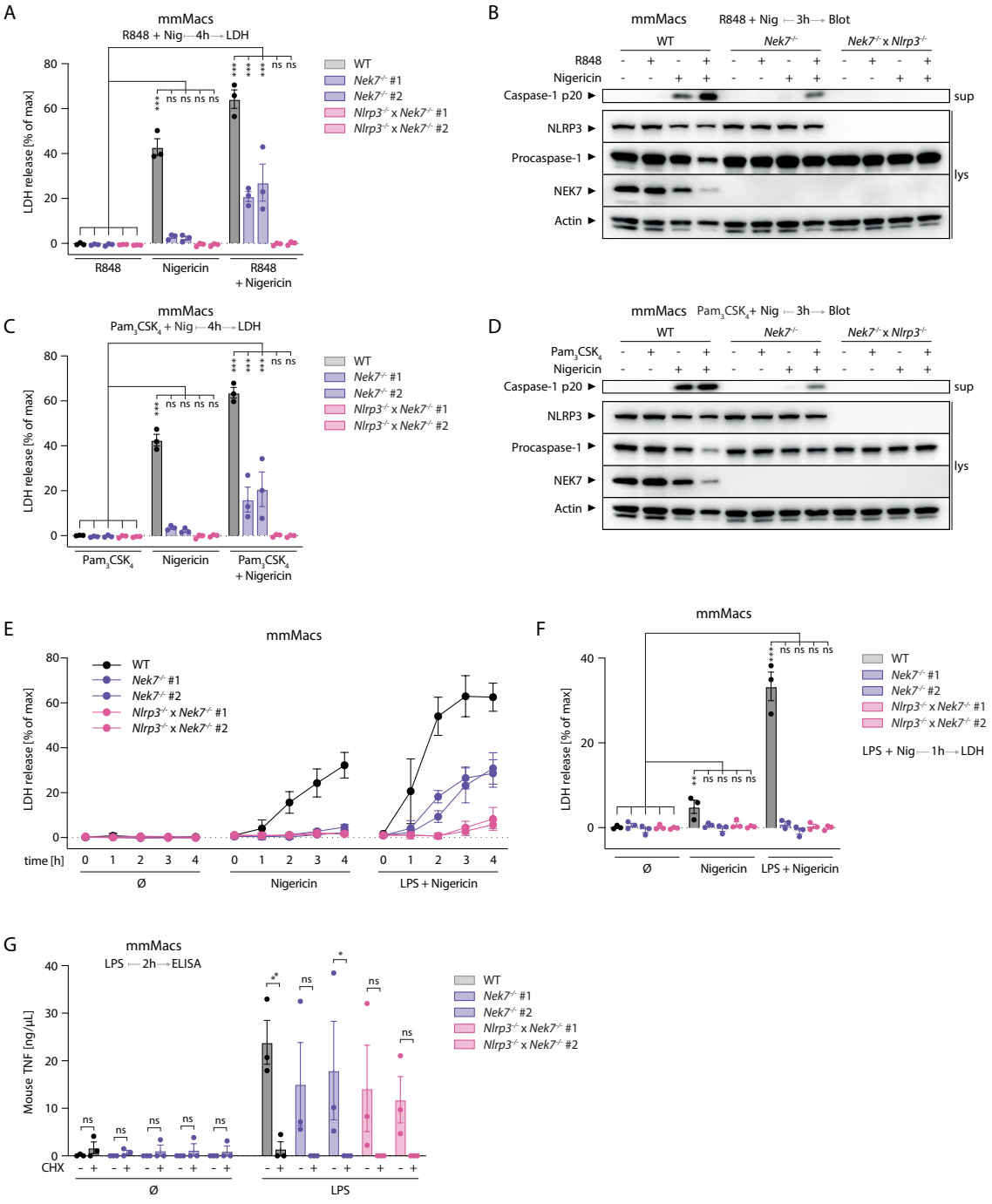

A
mmMacs
R848 + Nig 4h LDH
LDH release [% of max]
R848
Nigericin
R848 + Nigericin
WT
Nek7-/- #1
Nek7-/- #2
Nlrp3-/- x Nek7-/- #1
Nlrp3-/- x Nek7-/- #2
B
mmMacs
R848 + Nig 3h Blot
WT
Nek7-/-
Nek7-/- x Nlrp3-/-
R848
Nigericin
Caspase-1 p20
sup
NLRP3
Procaspase-1
NEK7
Actin
lys
C
mmMacs
Pam3CSK4 + Nig 4h LDH
Pam3CSK4
Nigericin
Pam3CSK4 + Nigericin
D
mmMacs
Pam3CSK4 + Nig 3h Blot
E
mmMacs
time [h]
Ø
Nigericin
LPS + Nigericin
F
mmMacs
Ø
Nigericin
LPS + Nigericin
LPS + Nig 1h LDH
G
mmMacs
LPS 2h ELISA
Mouse TNF [ng/µL]
CHX
Ø
LPS

Figure S3

Figure S3 related to Figure 2. *Priming bypasses NEK7 via a translation-independent mechanism*

(A) Mouse macrophages constitutively expressing mmNlrp3 (mmMacs) of the indicated genotypes were treated with R848 + Nigericin simultaneously for 4 hours.

(B) mmMacs of the indicated genotypes were treated with R848 + Nigericin simultaneously for 3 hours.

(C) mmMacs of the indicated genotypes were treated with Pam_3CSK_4 + Nigericin simultaneously for 4 hours.

(D) mmMacs of the indicated genotypes were treated with Pam_3CSK_4 + Nigericin simultaneously for 3 hours.

(E) mmMacs of the indicated genotypes were stimulated as indicated for up to 4 hours.

(F) mmMacs of the indicated genotypes were treated with LPS + Nigericin simultaneously for 1 hour.

(G) mmMacs of the indicated genotypes were pretreated with cycloheximide (CHX) and stimulated with LPS for 2 hours.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. Western Blots represent one of two clones from one of three independent experiments. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Tukey's test (A, C, F) or Šidák's test (G).

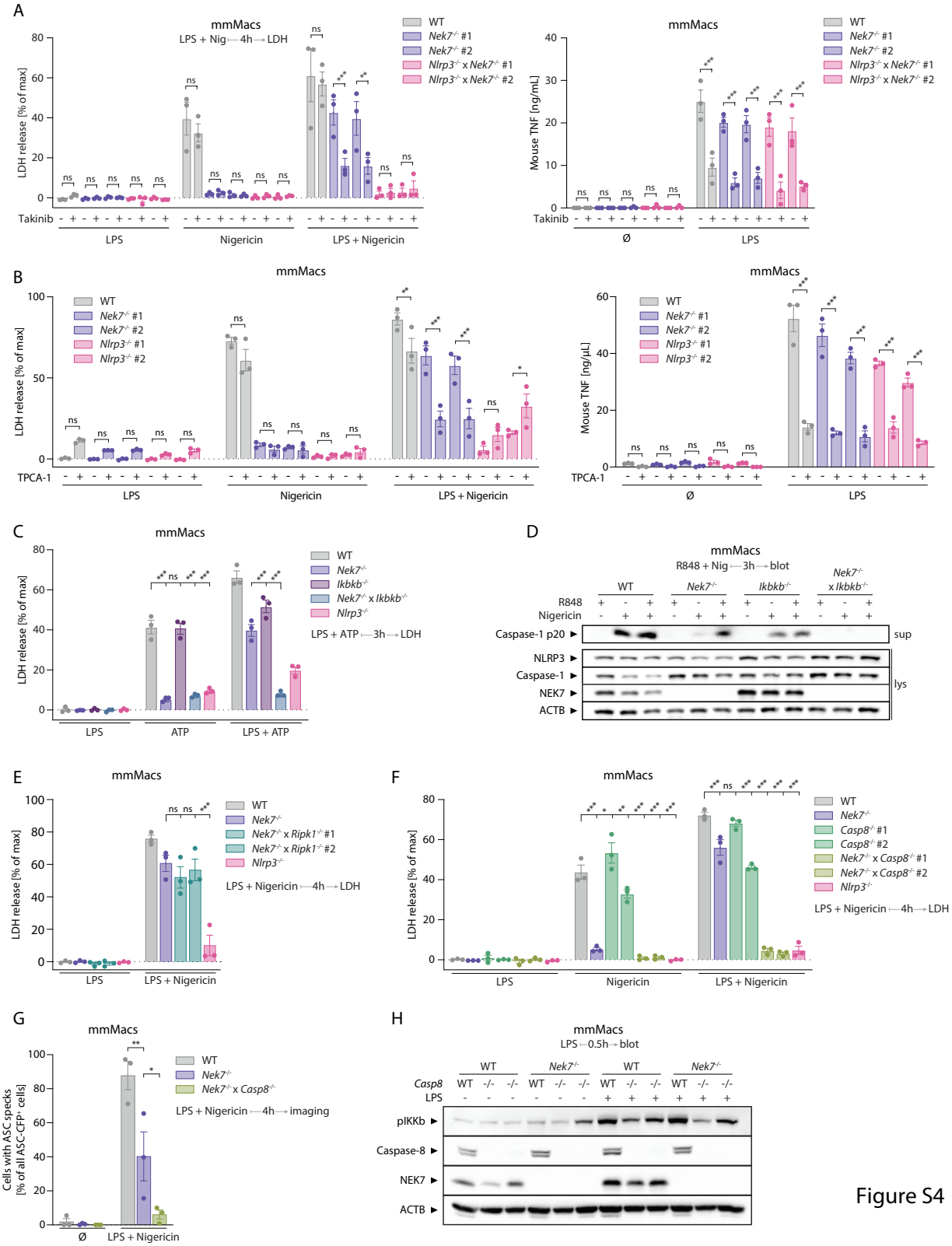

A
mmMacs
LPS + Nig 4h LDH
LDH release [% of max]
WT
Nek7-/- #1
Nek7-/- #2
Nlrp3-/- x Nek7-/- #1
Nlrp3-/- x Nek7-/- #2
Takinib
LPS
Nigericin
LPS + Nigericin
mmMacs
Mouse TNF [ng/mL]
Ø
LPS
B
mmMacs
LDH release [% of max]
WT
Nek7-/- #1
Nek7-/- #2
Nlrp3-/- #1
Nlrp3-/- #2
TPCA-1
LPS
Nigericin
LPS + Nigericin
mmMacs
Mouse TNF [ng/µL]
Ø
LPS
C
mmMacs
LDH release [% of max]
WT
Nek7-/-
Ikbkb-/-
Nek7-/- x Ikbkb-/-
Nlrp3-/-
LPS + ATP 3h LDH
LPS
ATP
LPS + ATP
D
mmMacs
R848 + Nig 3h blot
WT
Nek7-/-
Ikbkb-/-
Nek7-/- x Ikbkb-/-
R848
Nigericin
Caspase-1 p20
sup
NLRP3
Caspase-1
NEK7
ACTB
lys
E
mmMacs
LDH release [% of max]
WT
Nek7-/-
Nek7-/- x Ripk1-/- #1
Nek7-/- x Ripk1-/- #2
Nlrp3-/-
LPS + Nigericin 4h LDH
LPS
LPS + Nigericin
F
mmMacs
LDH release [% of max]
WT
Nek7-/-
Casp8-/- #1
Casp8-/- #2
Nek7-/- x Casp8-/- #1
Nek7-/- x Casp8-/- #2
Nlrp3-/-
LPS + Nigericin 4h LDH
LPS
Nigericin
LPS + Nigericin
G
mmMacs
Cells with ASC specks [% of all ASC-CFP+ cells]
WT
Nek7-/-
Nek7-/- x Casp8-/-
LPS + Nigericin 4h imaging
Ø
LPS + Nigericin
H
mmMacs
LPS 0.5h blot
WT
Nek7-/-
WT
Nek7-/-
Casp8
LPS
pIKKb
Caspase-8
NEK7
ACTB

Figure S4

Figure S4 related to Figure 2. *IKKβ activity enables NEK7-independent NLRP3 activation in mouse cells*

(A) mmMacs of the indicated genotypes were pretreated with Takinib for 30 minutes before stimulation as indicated.

(B) mmMacs of the indicated genotypes were pretreated with TPCA-1 for 30 minutes before stimulation as indicated.

(C) mmMacs of the indicated genotypes were stimulated as indicated for 3 hours.

(D) mmMacs of the indicated genotypes were stimulated as indicated for 3 hours. One representative of three independent biological replicates is shown.

(E, F) mmMacs of the indicated genotypes were stimulated as indicated for 4 hours.

(G) mmMacs of the indicated genotypes were stimulated as indicated for 4 hours. ASC-CFP specking was imaged every hour. The number of ASC-CFP specks in three separate visual fields was averaged for each biological replicate.

(H) mmMacs of the indicated genotypes were stimulated with LPS for 30 minutes. One wildtype and two *Casp8*$^{-/-}$ clones generated in a wildtype or *Nek7*$^{-/-}$ background are shown. One representative of three independent biological replicates is shown.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Šidák's test (A, B) or Tukey's test (C, E, F, G).

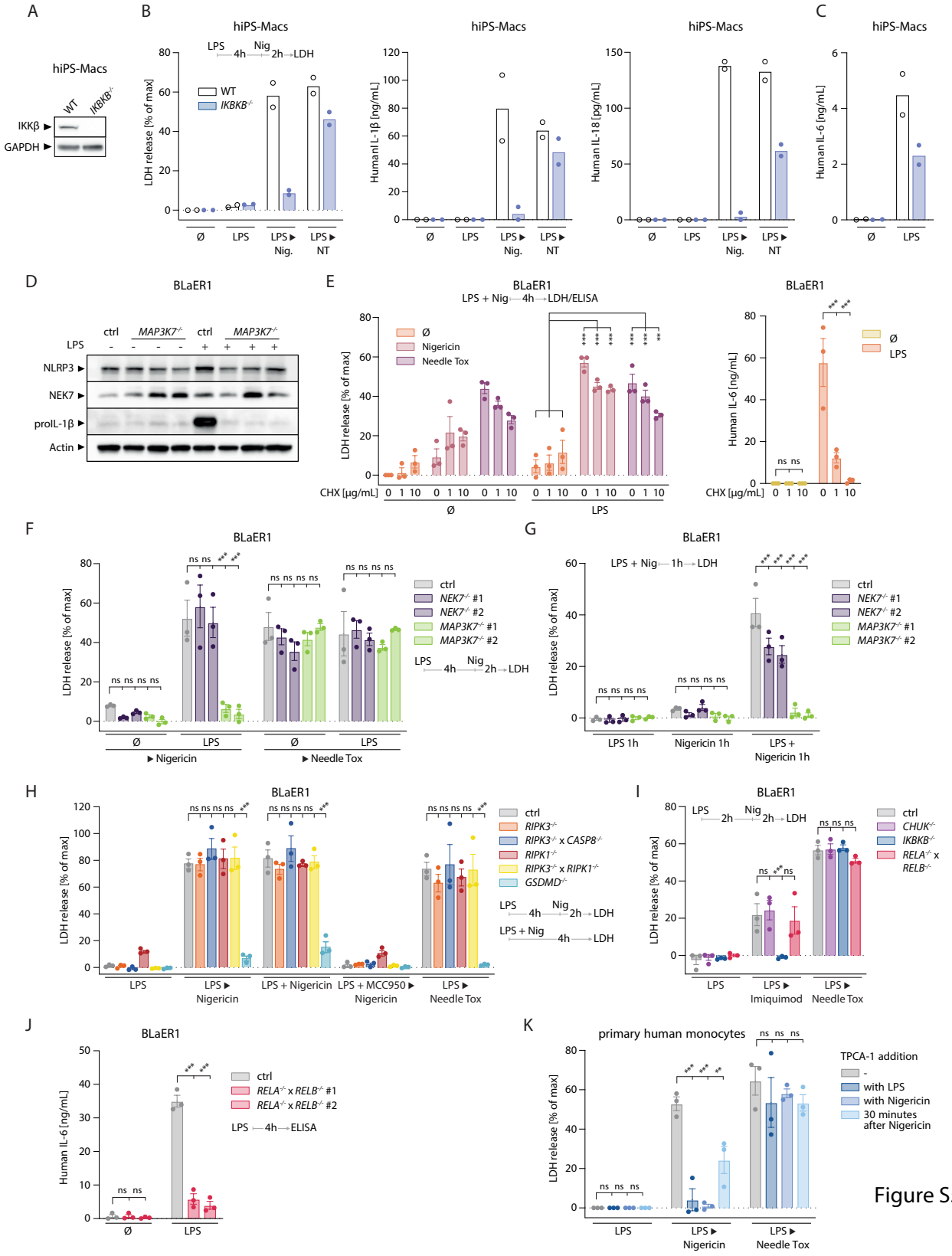

A
hiPS-Macs
WT
IKBKB-/-
IKKβ
GAPDH
B
hiPS-Macs
LPS 4h Nig 2h LDH
WT
IKBKB-/-
LDH release [% of max]
Ø
LPS
LPS ▶ Nig.
LPS ▶ NT
hiPS-Macs
Human IL-1β [ng/mL]
hiPS-Macs
Human IL-18 [pg/mL]
C
hiPS-Macs
Human IL-6 [ng/mL]
D
BLaER1
ctrl
MAP3K7-/-
LPS
NLRP3
NEK7
proIL-1β
Actin
E
BLaER1
LPS + Nig 4h LDH/ELISA
Ø
Nigericin
Needle Tox
LDH release [% of max]
CHX [µg/mL]
BLaER1
Human IL-6 [ng/mL]
F
BLaER1
LDH release [% of max]
ctrl
NEK7-/- #1
NEK7-/- #2
MAP3K7-/- #1
MAP3K7-/- #2
LPS 4h Nig 2h LDH
▶ Nigericin
▶ Needle Tox
G
BLaER1
LPS + Nig 1h LDH
LPS 1h
Nigericin 1h
LPS + Nigericin 1h
H
BLaER1
LDH release [% of max]
ctrl
RIPK3-/-
RIPK3-/- x CASP8-/-
RIPK1-/-
RIPK3-/- x RIPK1-/-
GSDMD-/-
LPS 4h Nig 2h LDH
LPS + Nig 4h LDH
LPS
LPS ▶ Nigericin
LPS + Nigericin
LPS + MCC950 ▶ Nigericin
LPS ▶ Needle Tox
I
BLaER1
LPS 2h Nig 2h LDH
LDH release [% of max]
ctrl
CHUK-/-
IKBKB-/-
RELA-/- x RELB-/-
LPS
LPS ▶ Imiquimod
LPS ▶ Needle Tox
J
BLaER1
Human IL-6 [ng/mL]
ctrl
RELA-/- x RELB-/- #1
RELA-/- x RELB-/- #2
LPS 4h ELISA
K
primary human monocytes
LDH release [% of max]
TPCA-1 addition
-
with LPS
with Nigericin
30 minutes after Nigericin
LPS
LPS ▶ Nigericin
LPS ▶ Needle Tox

Figure S5

Figure S5 related to Figure 4. *NLRP3 activation in human iPSC-derived macrophages and human myeloid cell lines requires non-transcriptional priming via IKKβ*

(A) Lysates of *IKBKB*$^{-/-}$ hiPS-Macs were immunoblotted for IKKβ.

(B, C) *IKBKB*$^{-/-}$ hiPS-Macs were primed with LPS for 4 hours and subsequently treated with the indicated inflammasome agonists before the release of LDH and the indicated cytokines was measured. Cells in (C) were treated with LPS for 4 hours. Dots represent separate, parallel differentiations of the clone shown in (A).

(D) Three clones of TAK1-deficient BLaER1 monocytes (*MAP3K7*$^{-/-}$) were treated with LPS for 6 hours before lysates were immunoblotted. One representative of three independent experiments is shown.

(E) BLaER1 monocytes were treated with LPS and Nigericin or Needle Tox in the presence of the indicated concentrations of cycloheximide (CHX) for 4 hours.

(F) BLaER1 monocytes of the indicated genotypes primed with LPS for 4 hours or left unprimed were stimulated with inflammasome inducers Nigericin (NLRP3) or Needle Tox (NAIP-NLRC4).

(G) BLaER1 monocytes of the indicated genotypes were concurrently stimulated with LPS and Nigericin for 1 hour before LDH release was measured.

(H) BLaER1 monocytes of the indicated genotypes were stimulated as indicated before LDH release was measured. Arrows denote 4 hours of LPS priming followed by the indicated stimulation.

(I) BLaER1 monocytes of the indicated genotypes were primed with LPS for 2 hours and then stimulated with Imiquimod or Needle Tox for 2 hours before LDH release was measured.

(J) BLaER1 monocytes of the indicated genotypes were stimulated with LPS for 4 hours before IL-6 secretion was measured.

(K) LDH release from primary human monocytes primed with LPS for 2 hours before stimulation for 2 hours as indicated. TPCA-1 was added at different timepoints as indicated. Bars depict mean ± SEM of three independent experiments.

Data are represented as mean ± SEM with dots representing biological replicates conducted on separate days unless indicated otherwise. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$ calculated by two-way ANOVA followed by Tukey's test (E LDH, H) or Dunnett's test (E IL-6, F, G, I, J, K).

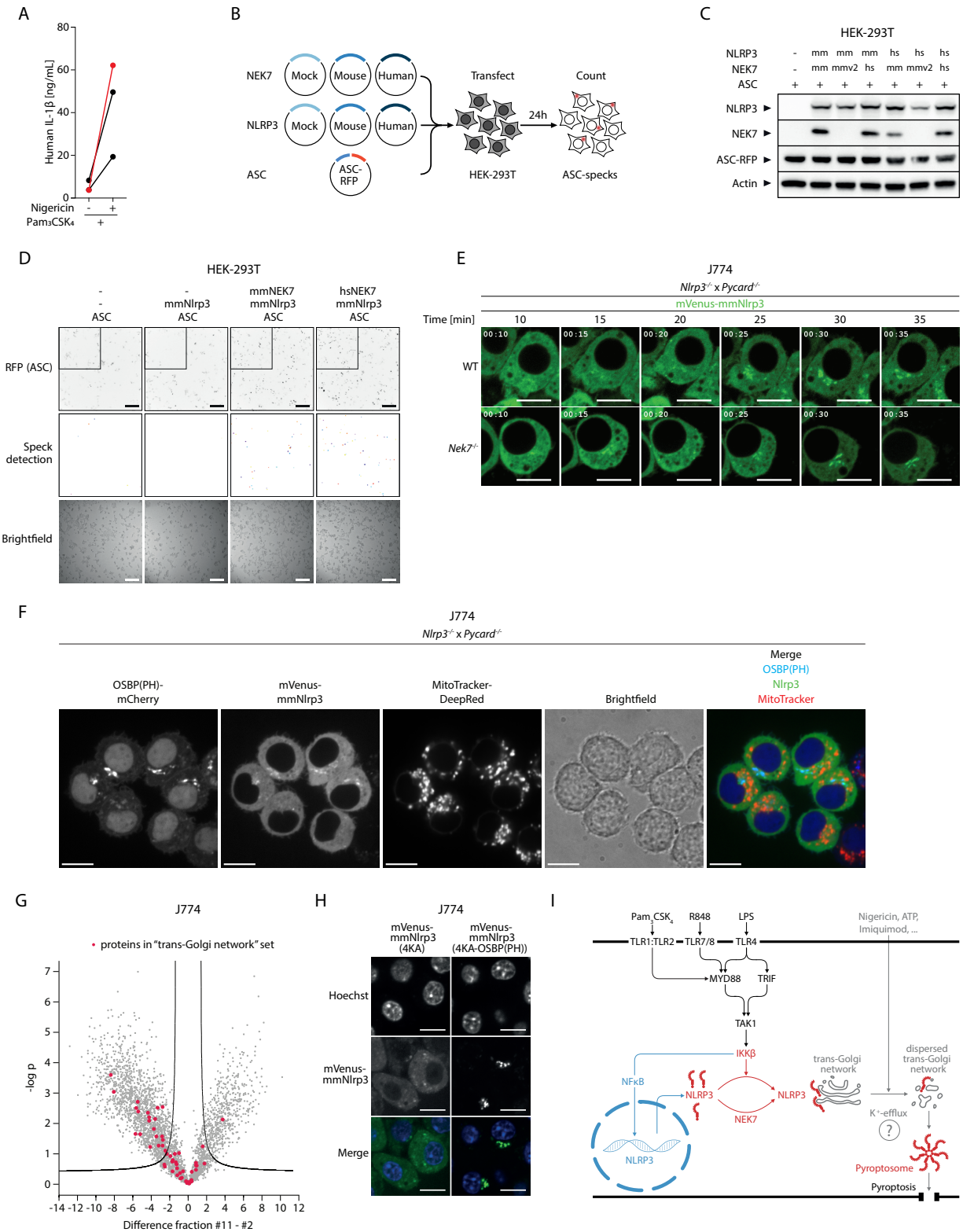

A
Human IL-1β [ng/mL]
Nigericin
Pam3CSK4
B
NEK7
NLRP3
ASC
Mock
Mouse
Human
ASC-RFP
Transfect
24h
Count
HEK-293T
ASC-specks
C
HEK-293T
NLRP3
NEK7
ASC
ASC-RFP
Actin
D
HEK-293T
ASC
mmNlrp3
mmNEK7
hsNEK7
RFP (ASC)
Speck detection
Brightfield
E
J774
Nlrp3-/- x Pycard-/-
mVenus-mmNlrp3
Time [min]
WT
Nek7-/-
F
J774
Nlrp3-/- x Pycard-/-
OSBP(PH)-mCherry
mVenus-mmNlrp3
MitoTracker-DeepRed
Brightfield
Merge
OSBP(PH)
Nlrp3
MitoTracker
G
J774
proteins in "trans-Golgi network" set
-log p
Difference fraction #11 - #2
H
J774
mVenus-mmNlrp3 (4KA)
mVenus-mmNlrp3 (4KA-OSBP(PH))
Hoechst
mVenus-mmNlrp3
Merge
I
Pam3CSK4
R848
LPS
TLR1:TLR2
TLR7/8
TLR4
Nigericin, ATP, Imiquimod, ...
MYD88
TRIF
TAK1
IKKβ
NFκB
NLRP3
NEK7
trans-Golgi network
dispersed trans-Golgi network
K+-efflux
Pyroptosome
Pyroptosis

Figure S6

Figure S6 related to Figures 5 and 6. *J774 mouse macrophages recruit NLRP3 to the TGN upon priming*

(A) THP-1 cells were primed with Pam_3CSK_4 for 4 hours before stimulation with Nigericin for 30 minutes. IL-1β release from three independent replicates is shown. The replicate highlighted in red corresponds to the immunoprecipitation shown in Figure 5A.

(B) Overview of NLRP3 inflammasome reconstitution in *NEK7*$^{-/-}$ HEK-293T cells.

(C) Expression of inflammasome components in *NEK7*$^{-/-}$ HEK-293T cells transiently transfected with NLRP3 inflammasome components as depicted in (B). "mmv2" refers to an annotated shorter transcript of mouse *Nek7*. The immunoblot corresponding to the replicate depicted in Figure 5B is shown.

(D) Micrographs of *NEK7*$^{-/-}$ HEK-293T transiently expressing the indicated NLRP3 inflammasome components. Images of ASC-RFP are shown with inverted colors. The second row shows the result of automated ASC-speck detection in the upper left region indicated in the top row. Scalebars represent 300 µm.

(E) *Nlrp3*$^{-/-}$ x *Pycard*$^{-/-}$ J774 cells of the indicated *Nek7* genotypes expressing mCherry tethered to phosphatidylinositol-4-phosphate (PI4P) via the PH domain of OSBP (OSBP(PH)-mCherry) and doxycycline-inducible mVenus-mmNlrp3 were treated with doxycycline for 24 hours and TPCA-1 for 1 hour before stimulation with LPS for 30 minutes. Images were taken every 5 minutes. Scalebars represent 10 µm.

(F) As (E) but in the presence of MitoTracker DeepRed and images taken 35 minutes after stimulation.

(G) Mass spectrometry analysis of the protein content of fractions #2 and #11 (Figure 6D). Proteins with FC ≥ 1.5 and FDR < 0.05 were considered to be significantly enriched in one fraction over the other and used for organelle enrichment analysis (Figure 6E). Proteins known to be on the TGN are highlighted in red.

(H) *Nlrp3*$^{-/-}$, *Pycard*$^{-/-}$ J774 cells expressing doxycycline-inducible mVenus-mmNlrp3(4KA) or mVenus-mmNlrp3(4KA-OSBP(PH)) were treated with doxycycline for 24 hours and Hoechst-33342 for 30 minutes before imaging. Scalebars represent 10 µm.

(I) Overview of NLRP3 priming pathways including IKKβ-mediated PI4P recruitment of NLRP3. The role of K^+ efflux, and whether it acts up- or downstream of TGN dispersal, is controversial.

Table S1 related to Figure 2E. *Statistical significance of LDH release from murine macrophages deficient in the indicated TLR4 signaling components*

LDH release from murine Macrophages (mmMacs) of the indicated genotypes following the indicated treatments as depicted in Figure 2E was compared to the respective WT or *Nek7*$^{-/-}$ controls. Results of ANOVA followed by Tukey's multiple comparison test are reported as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns $p \geq 0.05$, ns: not significant.

Stimulation	LPS				Nigericin				LPS + Nigericin			
Background	WT		*Nek7*$^{-/-}$		WT		*Nek7*$^{-/-}$		WT		*Nek7*$^{-/-}$	
Clone #	1	2	1	2	1	2	1	2	1	2	1	2
Tlr4$^{-/-}$	ns	ns	ns	ns	ns	ns	ns	ns	*	*	***	***
Myd88$^{-/-}$	ns	ns	ns	ns	ns	ns	ns	ns	ns	*	ns	ns
Ticam1$^{-/-}$	ns	ns	ns	ns	ns	***	ns	ns	ns	*	***	ns
Myd88$^{-/-}$, *Ticam1*$^{-/-}$	ns	ns	ns	ns	ns	ns	ns	ns	*	ns	***	***

4.2 Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome

Proteomics is a powerful technology that allows for the unbiased and large-scale characterisation of protein abundances. We recently developed deep visual proteomics (DVP), a technology which can investigate the molecular composition of tissues in a spatial context (Mund et al. 2021). In DVP, microscopy is used to identify cells of interest in the larger context of the tissue. Subsequently pools of phenotypically similar cells are analysed via mass-spectrometry (MS)-based proteomics. In this publication, by increasing MS sensitivity we were able to facilitate the analysis of single cells from tissues while preserving their spatial information which we termed single-cell DVP (scDVP). We used this approach to investigate the molecular components driving zonation of hepatocytes in the liver. Obtaining high quality images in which to identify cells of interest and map obtained protein abundances back to the spatial context requires an advanced imaging and image processing workflow. We developed an imaging routine that allowed for the high-throughput acquisition of confocal microscopy images on tissue sections. During image processing, we stitch individual image tiles with sub-micrometer accuracy even across entire microscopy slides. This enabled the development of a computational model that could assign proteome classes to all cells across the imaged tissue section on the basis of the acquired images alone. With these innovations, scDVP now allows for the fine-grained proteomic analysis of tissues while retaining the spatial context of the analysed cells through imaging.

The following research article was originally published here:

Rosenberger, F. A., Thielert, M., et al. (2023). “Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome”. In: *Nature Methods* 20.10, pp. 1530–1536. ISSN: 1548-7091. DOI: 10.1038/s41592-023-02007-6

nature methods

Article

https://doi.org/10.1038/s41592-023-02007-6

Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome

Received: 9 December 2022

Accepted: 15 August 2023

Published online: 2 October 2023

Check for updates

Florian A. Rosenberger[1], Marvin Thielert[1], Maximilian T. Strauss[2], Lisa Schweizer[1], Constantin Ammar[1], Sophia C. Mädler[1], Andreas Metousis[1], Patricia Skowronek[1], Maria Wahle[1], Katherine Madden[1], Janine Gote-Schniering[3], Anna Semenova[3], Herbert B. Schiller[3], Edwin Rodriguez[1], Thierry M. Nordmann[1], Andreas Mund[2] & Matthias Mann[1,2]

Single-cell proteomics by mass spectrometry is emerging as a powerful and unbiased method for the characterization of biological heterogeneity. So far, it has been limited to cultured cells, whereas an expansion of the method to complex tissues would greatly enhance biological insights. Here we describe single-cell Deep Visual Proteomics (scDVP), a technology that integrates high-content imaging, laser microdissection and multiplexed mass spectrometry. scDVP resolves the context-dependent, spatial proteome of murine hepatocytes at a current depth of 1,700 proteins from a cell slice. Half of the proteome was differentially regulated in a spatial manner, with protein levels changing dramatically in proximity to the central vein. We applied machine learning to proteome classes and images, which subsequently inferred the spatial proteome from imaging data alone. scDVP is applicable to healthy and diseased tissues and complements other spatial proteomics and spatial omics technologies.

Mass spectrometry (MS)-based single-cell proteomics (scProteomics) has made tremendous progress within just a few years, and can now quantify more than 1,000 proteins in cultured cells[1–3]. While this trajectory is promising, proteome depth, throughput and lack of spatial context limit biological use. We have recently introduced deep visual proteomics (DVP), a spatial technology that combines imaging, cell segmentation, laser microdissection and MS into a single workflow to investigate complex tissues with various cell types and metabolic niches[4]. DVP overcomes depth and throughput limitations with pooling the required number of cells with similar morphological features and staining patterns to identify statistically and analytically robust cellular phenotypes ('biological fractionation'). By its nature, it depends on prior knowledge of adequate markers of the cells of interest that resolve their heterogeneity. These markers might not be available for all subtypes of cells or those tissues that have rapidly changing proteome types such as heterogeneous tumors. To address this, we here developed single-cell DVP (scDVP), a complementary approach that extends scProteomics technologies into the intact tissue context.

In this Article, we use scDVP to explore spatial characteristics of hepatocyte subsets in mammalian liver—a highly organized and functionally repetitive tissue, in which the proteome of hepatocytes is determined by paracrine signaling, as well as oxygen and nutrient gradients[5]. These metabolic gradients require distinct functional cell states along the portal vein (PV) to central vein (CV) axis. This phenomenon of liver zonation has been described by single-cell RNA sequencing (scRNAseq) for hepatocytes[6,7], fluorescence-activated cell sorting (FACS) and MS-based proteomics[8], and multiplexed imaging[9]. Despite this long and varied background, the extent of spatial heterogeneity and proteome variation in hepatocyte remains an open question.

[1]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Comprehensive Pneumology Center (CPC) / Institute of Lung Health and Immunity (LHI), Helmholtz Munich; Member of the German Center for Lung Research (DZL), Munich, Germany.
e-mail: mmann@biochem.mpg.de

Article https://doi.org/10.1038/s41592-023-02007-6

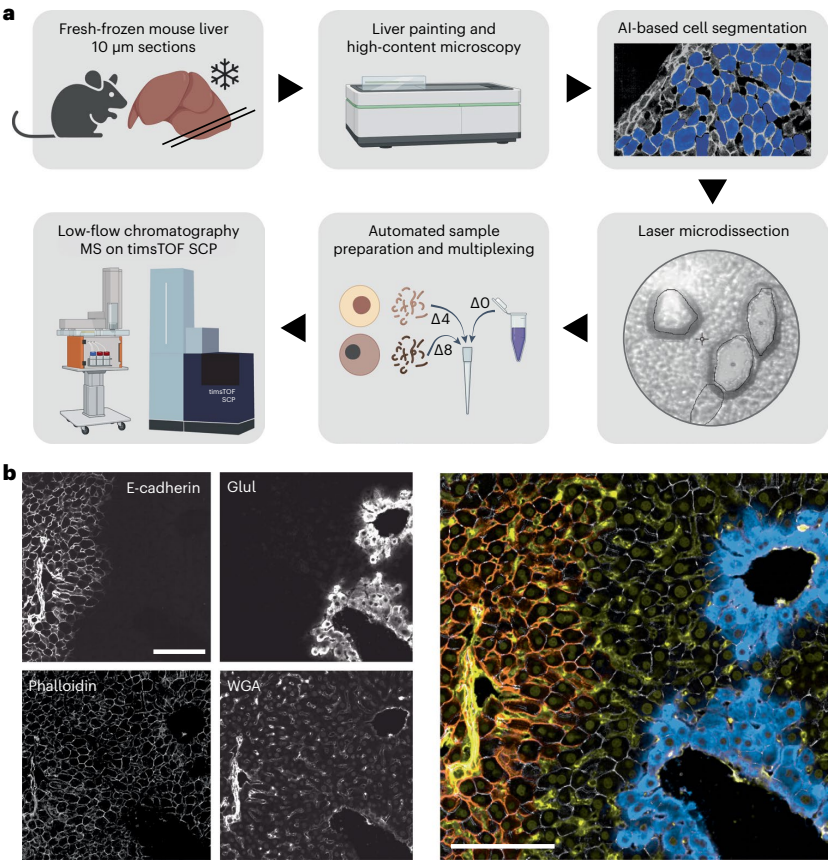


Fig. 1 | Isolation and characterization of individual hepatocyte shapes in situ. a, The scDVP workflow comprised embedding of fresh mouse liver tissue, staining and high-content microscopy, AI-guided hepatocyte segmentation, cutting and sorting of cells on a laser microdissection microscope, and peptide preparation with or without dimethyl labeling. The Δ0 channel contains the reference proteome and Δ4 and Δ8 contain two individual samples, which are all analyzed by ultra-high-sensitivity mass spectrometry. Created with BioRender.com. **b**, Liver painting with four stains. Left: E-cadherin marks PV regions, glutamate-ammonia ligase (Glul) surrounds the CV, the cell segmentation marker phalloidin, and the sinusoidal and nuclear counterstain WGA. Right: false color overlay of all channels: orange, E-cadherin; yellow, WGA, gray, phalloidin; turquoise, Glul. Scale bars, 100 µm.

Results

Robust isolation and characterization of hepatocyte shapes

To map the proteome of mouse hepatocytes at single-cell resolution, we established a modular and automated workflow aimed at loss-less sample preparation of the initial input cell for injection into the mass spectrometer (Fig. 1a). Mice livers were embedded and immediately frozen after cardiac arrest. We fixed 10 µm sections and stained them with a one-step protocol marking PVs and CVs, the sinusoidal architecture, nuclei and cell membranes (Fig. 1b and Methods). Individual cells were segmented by deep learning as before[4], and the resulting masks transferred to a laser microdissection microscope that automatically excised and collected individual shapes in 384-well plates. Given hepatocyte sizes of 20–30 µm, one shape cut from a 10 µm section corresponds to a third or half of a hepatocyte, or approximately 250 pg of protein input, equivalent to the protein content of one HeLa cell. We automated protein extraction and digestion by reagent addition into the same plate, omitting extra transfer steps, followed by peptide separation on the Evosep system[10] and injection into a trapped ion mobility time-of-flight single-cell proteomics (timsTOF SCP) mass spectrometer (Fig. 1a).

To establish an efficient workflow, we applied our scProteomics protocol[2] and titrated the number of cells required to obtain a robust signal (Extended Data Fig. 1a). To confirm biological ground truth, we performed initial experiments on five adjacent shapes per well (corresponding to about two complete hepatocyte cell masses), cut from randomly chosen locations. With these five shapes, we reached a median depth of 1,235 proteins across 230 samples (Extended Data Fig. 1b). The results confirmed expected liver biology, for instance, by differential expression of the PV marker argininosuccinate lyase (Asl) and central Cytochrom P450 2E1 (Cyp2e1, Extended Data Fig. 1c). Using zonation anchor proteins to arrange all the samples in pseudo-space (Extended Data Fig. 1d), we characterized spatially enriched gene sets along the zonation axis. While the protein sets for electron transport chain and oxidative phosphorylation (OXPHOS) were among processes upregulated in proximity to the PV, biotransformation and oxidations by cytochrome P450 were increased proximal to the CV, providing positive controls for low-input proteomics (Extended Data Fig. 2 and Supplementary Tables S1 and S2).

Multiplex-DIA drastically increases proteome depth

Encouraged by these spatial results, we next asked whether single shapes alone could produce deep and interpretable proteomic results. To improve sensitivity further, we adopted and optimized elements of

Article https://doi.org/10.1038/s41592-023-02007-6

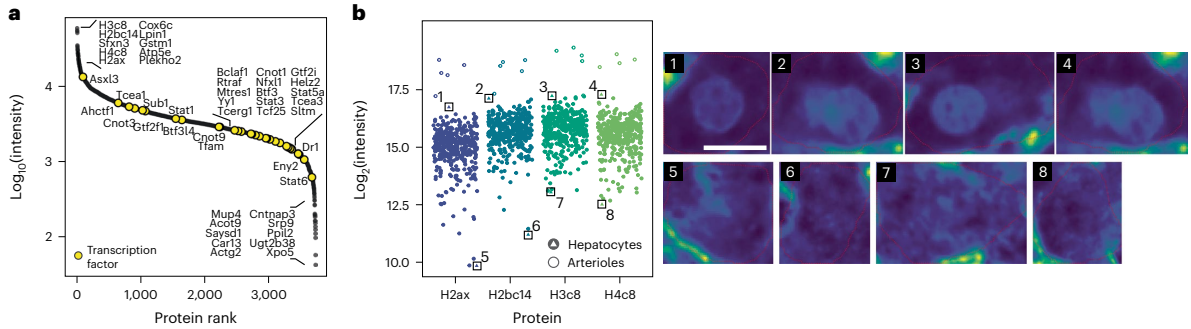


Fig. 2 | Depth of single-shape proteomes and estimation of the nuclear compartment. a, Unique proteins quantified in our scDVP workflow with mDIA ranked by signal intensity (two single-shape and a reference proteome channels, 31 min Evosep gradient, 15 cm column at 100 nl min^{-1}, dia-PASEF with optimized window design, library-dependent search in DIA-NN). Names of highest- and lowest-ranking proteins, as well as transcription factors, are indicated. **b**, Left: intensity of the top four histone proteins across all samples, including hepatocytes and quality control arteriole structures. Colors are specific for the indicated histone subunit. Right: WGA stain of cells corresponding to marked data points in the scatter plot. The color scale is signal intensity. Scale bar, 10 µm. Data from three mice were pooled.

our scProteomics workflow[11]. These include addition of the surfactant *n*-dodecyl-β-D-maltoside (DDM) to maximize peptide recovery[12], lowering the chromatographic flow rate to 100 nl min^{-1} for increased ionization efficiency[2] ('Whisper gradients' on the Evosep system) and achieving higher chromatographic resolution with zero dead volume columns (IonOpticks)[13]. Most importantly, we added a labeled reference channel for multiplexed data-independent acquisition (mDIA) that decouples identification and quantification[11] (Fig. 1a).

For scDVP, we constructed a dimethyl-labeled bulk liver reference. Our robotic sample preparation setup achieved about 99% labeling efficiency in all three channels (Extended Data Fig. 3a). We co-injected 10 ng of the reference proteome together with the labeled proteomes of two single shapes at a mean size of 600 $µm^2$ (Extended Data Fig. 3b). This resulted in a doubling of identified proteins with a median number of 1,712 proteins across three biological replicates and 455 single shapes, at twice the previous throughput (Extended Data Fig. 3c). A maximum of 2,769 proteins were identified in one shape, and 3,738 unique proteins were found across all samples (Fig. 2a and Extended Data Fig. 3c). Four histone components ranked in the top ten, but we also found many transcription factors (Fig. 2a). In contrast, plasma proteins produced in hepatocyte were of medium abundance and hemoglobin subunits were not detected. This suggests little to no contamination from surrounding blood, a common issue in bulk proteomics (Extended Data Fig. 3d). The number of detected proteins correlated logarithmically with the microdissected area (Extended Data Fig. 3e), indicating that scDVP requires the highest possible MS sensitivity. Data completeness across all samples increased with median intensity per protein. Coefficients of variation were less than 50% and strongly depended on cell size and position along the zonation axis, reflecting biological heterogeneity in the data (Extended Data Fig. 3f,g). We hypothesized that the nuclear proportion in the cell slice would correlate with the intensity of these histones. Indeed, shapes with lowest histone intensities did not have any evident nuclear signal, while top intensities were in shapes with large or two nuclei. In addition to this, the intensity of the top four abundant histone proteins was highest in arterioles that we cut as technical control structures, and which are composed of more than one cell and nucleus (Fig. 2b and Extended Data Fig. 3h).

Single-shape proteomes accurately reflect hepatocyte zonation

To test the biological validity of our proteomics data, we first reduced dimensionality in a principal component analysis (PCA), which revealed that PC1 represented the measured distance of a hepatocyte to PV and CV (Fig. 3a,b). Overlays of known liver zonation markers including Cyp2e1 and argininosuccinate lyase (Asl) showed opposite visual enrichment along PC1 (Extended Data Fig. 4a,b). In contrast, PC2 did not correlate with measured distance or hepatocyte zonation markers but rather with cytoskeletal components (Extended Data Fig. 4c,d). PC2 was also the dimension in which portal arterioles, which we excised as technical controls, separated from hepatocytes (Extended Data Fig. 4e).

We asked whether single-cell resolution provides any benefit over combining adjacent shapes. To this end, we iteratively combined shape information and averaged protein levels of cells with the same relative location along the zonation axis ('pseudo-neighbors'). Starting from the combination of as little as two shapes, PC1 continuously gained importance as measured by interquartile range and variance explained, whereas PC2 and all subsequent components dropped in explanatory value (Extended Data Fig. 5). This demonstrates that single-cell data retains subtle biological differences compared to the excision of larger areas.

On the basis of the distance ratio of PV and CV, we grouped the data into 20 spatial bins—approximately the maximum number of cells along the zonation axis. Analysis of variance (ANOVA) testing across all bins revealed that 49% of all proteins detected in at least half of the samples were significantly different between zones (false discovery rate (FDR) <0.05; Extended Data Fig. 6a and Supplementary Table S2). Zonation was also apparent after spatial sorting at the total proteome level (Fig. 3c and Supplementary Table S3) and for known hepatocyte zonation markers (Fig. 3d). Only 5.8% of these proteins were expressed equally in all zones (multiple testing-adjusted Shapiro–Wilk test, $P > 0.05$), including electron transfer flavoprotein β (Etfb), the electron acceptor in mitochondrial fatty acid β-oxidation (Extended Data Fig. 6b).

The first principal component along the zonation axis indicated that portal and periportal regions were more similar to one another than central and pericentral zones (Fig. 3b). Indeed, the spatial expression of the top ten significant zonation markers for each portal and central regions followed a hockey-stick curve from portal to central (Fig.3e), similar to Wnt-controlled transcripts in a scRNAseq dataset[6] and in line with a CV origin of Wnt signaling[14]. In contrast, this pattern was absent for the hits with the highest *P* values (least zonated hits; Extended Data Fig. 6d).

A cross-omics comparison with scRNAseq data[6] confirmed the directionality of the most prominent zonation markers (Pearson's $R = 0.97$, Extended Data Fig. 7a,b), while correlation was low across all proteins and transcripts (Pearson's $R = 0.12$). Notably, a number of proteins were regulated only in the RNA or protein dimension, or

Article https://doi.org/10.1038/s41592-023-02007-6

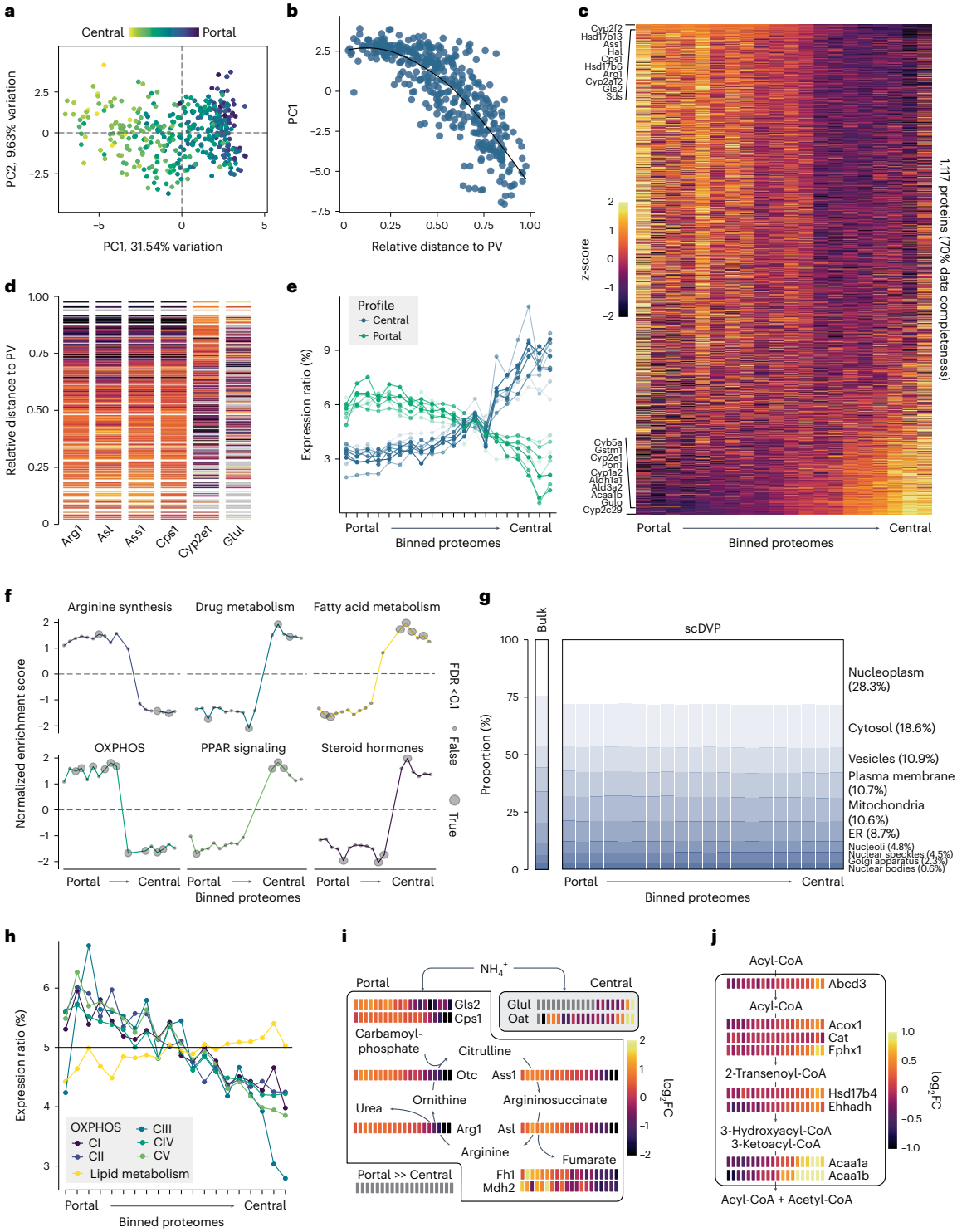


Fig. 3 | Single-shape proteomes are accurate descriptors of zonated hepatocytes. a, PCA of all hepatocytes. The color overlay corresponds to the ratio of measured distance PV over CV in the microscopy image. **b**, Measured distance ratio versus PC1. Relative distance of 0 is at the PV and of 1 is at the CV. Black: smoothing curve. **c**, Heat map of protein expression as *z*-score per protein across all samples. Proteins are ordered according to ANOVA fold change (FC) across 20 spatial equidistant bins, summarizing samples with a similar distance ratio to PV and CV. The ten top and bottom proteins are given. Only proteins that were detected in 70% of all samples are included. **d**, Protein expression as z-score of selected marker proteins, ordered by relative distance to PV and CV. One line is one shape measurement. Gray: protein not detected. **e**, Expression of the top 20 significant proteins in 20 spatial bins, relative to total expression from portal to central. Zonation peak at PV: positive ANOVA fold change ($n = 10$), and vice versa ($n = 10$). **f**, Selected gene sets in individual spatial bins versus all others bins, depicting normalized enrichment score after gene set enrichment analysis. Dot size: significance after multiple testing adjustment. **g**, The proportion of protein signal stratified by subcellular compartment in a bulk mouse liver proteome and the scDVP dataset. Percentages refer to mean across spatial bins in the scDVP data. **h**, Relative expression in 20 bins from PV to CV of proteins constituting mitochondrial OXPHOS components (C) I–V, and mitochondrial lipid metabolism. **i**, Levels of urea cycle and connected enzymes from portal (left) to central (right) bins as $\log_2$FC relative to median expression in the two center bins. Portal box: active in portal regions. Central box: active in central region. **j**, Levels of peroxisomal enzymes related to very-long chain fatty acid degradation, spatially resolved as in **g**. Data from three mice were pooled.

even inversely correlated (Extended Data Fig. 7c), such as dimethylglycine dehydrogenase in the choline catabolic pathway. Similarly, when we compared our data with a FACS-based hepatocyte proteome[8], we found slightly lower correlation of markers and better overlap overall (Pearson's *R* of 0.16 versus 0.12, Extended Data Fig. 7d–f). Members of glutathione metabolism had similar spatial distribution in both datasets (Extended Data Fig. 7g). This underlines that the scDVP dataset provides orthogonal insight into liver physiology instead of merely complementing existing datasets.

Enrichment of functional protein sets across the spatial bins confirmed that arginine biosynthesis and OXPHOS were highly enriched toward the PV (Fig. 3f). When we added subcellular annotations to our dataset, we found negligible differences to a bulk mouse liver proteome for many compartments including the plasma membrane (summed intensity of 10.6% in the library versus 10.7% in this scDVP data), highlighting that laser microdissection is suitable to excise the entire shape (Fig. 3g and Supplementary Table S6). On a biological level, we found only small changes of summed organellar intensities across spatial bins (Extended Data Fig. 8 and Supplementary Table S6), namely decreasing mitochondrial and endoplasmic reticulum mass and increasing Golgi apparatus and nucleoplasm from PV to CV. When cross-mapping the scDVP data with the mitochondrial protein library Mitocarta 3.0 (ref. 15), the five complexes of OXPHOS decreased collectively by more than 25%, yet mitochondrial proteins related to fatty acid metabolism mildly increased conspatially, suggesting differential regulation within the same cellular compartment (Fig. 3h). Remarkably, these protein sets reach their spatial expression at the midpoint between PV and CV in contrast to the hockey-stick distribution of the top ten differentially expressed proteins (Fig. 3e,h), suggesting that the mitochondrial compartment is not dependent on the Wnt-signaling gradient.

The scDVP data correctly confirmed that proteins participating in ammonia fixation of the urea cycle were highly expressed in portal regions, while those involved in ammonia capture on glutamate were strongly pericentral (Fig. 3i). To our surprise, several other signaling-related pathways were also zonated including peroxisome proliferator-activated receptor (PPAR) signaling (Fig. 3f and Supplementary Table S5). This was corroborated by prominent central expression of enzymes required for peroxisomal degradation of very-long-chain fatty acids, and ω-oxidation of dicarboxylic C12 fatty acids, enriched in, for instance, coconut oil (Fig. 3j). We conclude that the spatial proteome data from single hepatocyte shapes is biologically accurate and informative, and furthermore, contains rich biological information to be mined.

Spatial context regulates single-cell proteomes

Combining the single-shape proteomes with their inherent spatial information and staining intensities, scDVP revealed clear dependence of fluorescent intensities with the eight proteome classes established above (Fig. 4a,b).

Encouraged by the evident complementarity between extensive proteomics and spatial data, we reasoned that the microscopic image could contain sufficient information to predict the proteome. To this end, we trained a machine-learned (ML) model on 17 features to predict the proteome classes from imaging data. We grouped the training set into five proteome classes by *k*-means clustering (Extended Data Fig. 9a), and used the information in all imaging channels as predictors (Extended Data Fig. 9b). This model reached an average precision of 0.94 (Extended Data Fig. 9c,d), correctly assigning the proteome class of almost all cells. Errors occurred exclusively between spatially neighboring classes (Fig. 4c).

We tested the model performance on a new section (not used in training), from which we measured 60 single-shape proteomes. Visual inspection indicated that the predicted classes were correctly located in proximity to CV or PV, even in the presence of cutting artifacts (Fig. 4d). We used the class probabilities as weights to predict the spatial proteome, which accurately approximated overall protein intensities ($R = 0.78$ between prediction and measurement, Fig. 4e). When predicting the proteome of a larger section for all quantified proteins, the ML model correctly assigned the spatial directionality of zonation markers, as well as their expected extension into the intermediate zone (Fig. 4f). Thus, the model confirms the accuracy of measured single-shape proteomes, and is furthermore a potent predictor of spatial proteomes across any imaged areas.

Discussion

Here, we present a single-cell spatial map of the murine liver acquired by MS-based proteomics. Our approach successfully combines microscopic imaging data with ultra-high-sensitivity proteomics, building on four major technological advances: (1) artificial intelligence (AI)-assisted segmentation and laser microdissection, (2) multiplex-DIA (mDIA), (3) low-flow gradients and (4) the ultra-high sensitivity of a timsTOF SCP mass spectrometer.

To date, MS-based scProteomics has been exclusively reported for cell suspensions. State-of-the-art workflows currently reach a proteomic depth of up to 2,000 proteins in cultured cells, with about 250 pg of cellular protein mass. This is similar to the protein material in our sliced hepatocytes, taking the section thickness of 10 µm and hepatocyte size of 20–30 µm into account. With our scDVP workflow, we achieved more than 1,700 proteins per single shape (and up to 2,700) despite working from sections that were fixed, stained, imaged and laser dissected. Laser microdissection successfully separated hepatocytes from surrounding material including blood remnants, which holds promise for smaller cell types in more complex tissue environments. The size of our shapes correlated strongly with the number of identified proteins, suggesting that scDVP is currently limited by MS sensitivity and will thus profit from continuous technical developments. We established the scDVP protocol to combine one reference channel with two single shapes (effective two-plex) and used a 40 samples per day chromatography method. This can be

Article https://doi.org/10.1038/s41592-023-02007-6

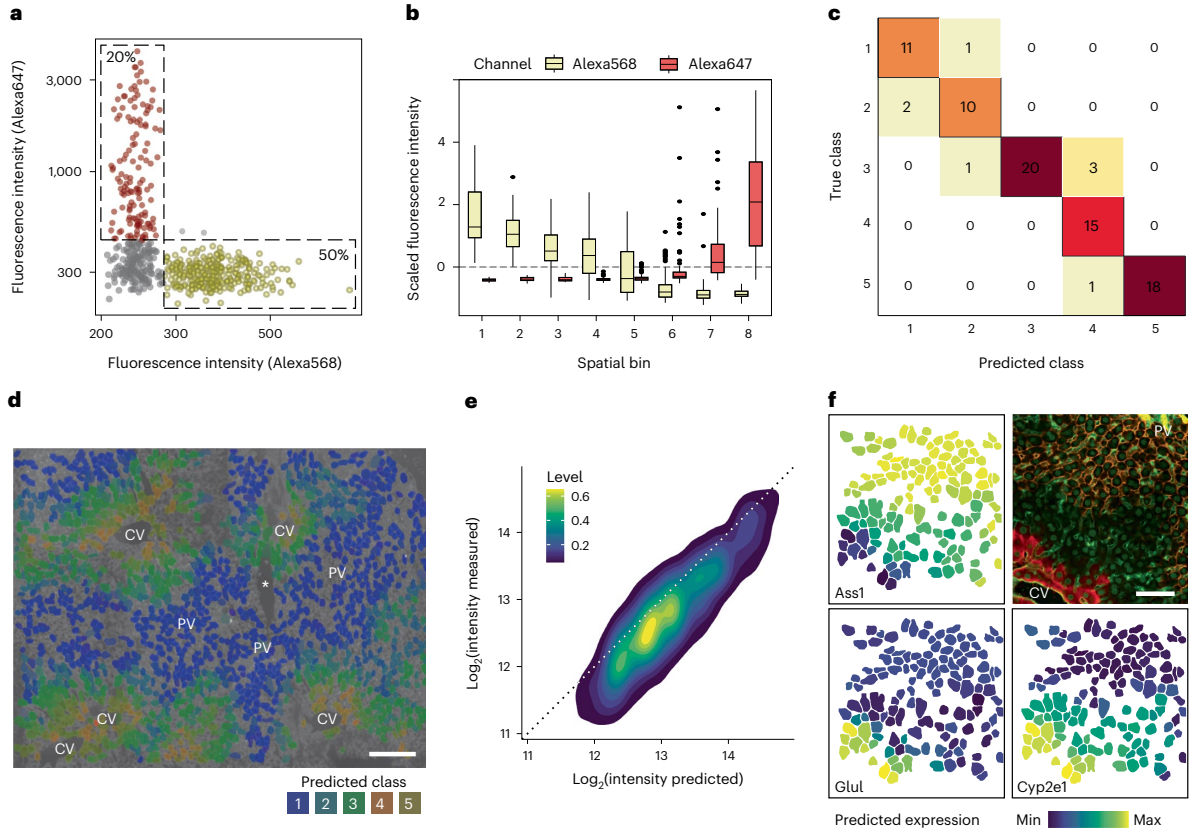


Fig. 4 | Combining imaging and proteome data for a ML model. a, Fluorescence intensities of Alexa568 (PV marker E-cadherin) and Alexa647 (CV marker Glul), with percentages in indicated bins. Each dot represents one shape. **b**, Intensities of the spatial markers as in **a** across eight spatial bins. The boxes are first and third quartiles, the thick line is the median, whiskers are ±1.5 interquartile range and outliers are indicated as individual points. **c**, Confusion matrix of a ML model with five classes, informed by microscopy and proteomics data. Colors scale with counts in each box. **d**, Predicted classes of segmented hepatocytes. The hue is the maximum class probability. **e**, Density plot of predicted versus measured intensities of a section excluded from machine learning ($R = 0.78$). **f**, Spatial depiction of one biological replicate with microscopy ground truth on top right, and three predictions. Orange, E-cadherin; red: Glul; green: WGA. *Sectioning artifact. Scale bar, 150 µm.

further scaled to five-plex (effective four-plex) and 80 samples per day, scaling to 320 shapes per day[11]. Given the more stable core proteome compared with single-cell transcriptomes[2] and the resulting lower required sample number, scDVP experiments encompassing a few hundred single shapes could be done in just a few days. Furthermore, because of the very low quantities and absence of proprietary reagents, marginal costs are extremely low.

Our proteomics data from single shapes correctly and accurately recapitulates hepatocyte physiology by direction, extent and spatial organization of zonation. More than half of quantified proteins were significantly different between portal and central zones, in line with scRNAseq data[6,16] and FACS-based proteomics data[8]. The fact that we detected all of the previously used markers of liver zonation[6] suggests that our proteomic depth is sufficient to integrate into other omics datasets. This became further apparent on the level of functional pathways, including signaling and disease pathways. Interestingly, peroxisomal degradation of very-long-chain fatty acids, as well as dicarboxylic C12 fatty acids, was enriched in proximity to the CV. Biochemical evidence by radiolabeling experiments support the notion that nonmitochondrial fatty acid oxidation localizes to pericentral regions[17]. We report an almost linear decrease of mitochondrial mass and OXPHOS subunits along the zonation axis. This is in line with intravital microscopy data showing decreasing mitochondrial membrane potential[18]. A rhythmic expression pattern has been previously shown for a large number of liver transcripts and proteins[16,19]. While we have not covered the temporal aspect here, the scDVP approach could contribute to such studies by adding a spatial dimension.

In the previously described DVP workflow, we used pools of cells combined on the basis of common features, such as the expression intensity of already known markers, or morphology[4]. This approach allows a deep, rapid and robust proteome characterization that accurately represents the underlying biology. By analyzing single cellular shapes without prior assumptions, scDVP now removes the dependency on established markers or features. This makes it a promising approach in heterogeneous tissues with partially or not defined subtypes of cells, such as in many tumor tissues. Moreover, scDVP can be a method of choice to map proteomic disturbances along gradients of, for instance, signaling factors, nutrients or gases, and in physiological settings that may create impediments for other omics methods, for instance, in extracellular fibrotic scars.

Our results demonstrate that the central challenge of scDVP is the sensitivity of the overall workflow. Although we have here reduced the area

required for laser microdissection by 100-fold compared with our initial DVP report[4], we note that one excised hepatocyte shape contains approximately ten times more protein than the smallest cells of interest, such as typical resting lymphocytes[20]. While the required sensitivity is being developed, the original DVP approach using pooled cells of the same type is a powerful tool for this kind of problems. We also note recent success in drastically scaling down DVP for formalin-fixed and paraffin-embedded samples, which are readily available in many clinical settings[21].

There have been advances in the quantification of posttranslational modifications from ultra-low-input material, such as from 1 µg down to the material corresponding to single cell, for instance in the enrichment protocol µPhos[22]. In combination with scDVP, this holds promise for single cells, although the biological information in single-cell phosphoproteomics data would currently be limited to a few high-abundance proteins with high modification stoichiometries. Subtle signaling events, such as the liver-dominant Wnt signaling, will require additional technological developments for in-depth biological description of signaling in single cells by MS.

We have shown that single-cell data can be used to train an accurate ML model that predicts the proteome class from visual information only. Evidence suggests that morphological features such as nuclear vacuolation and texture associate with zonation, and can even serve as a progression and stratification marker of nonalcoholic fatty liver disease[23]. Combining such easily available features and extensive proteomic sampling can clearly lead to higher precision of the predictive models. Transfer learning might then extend the approach to many new areas, as already shown for single-cell transcriptomics data[24]. The modular nature of scDVP, especially the open format from laser microdissection to 384-well plates for sample preparation, makes it widely applicable and also compatible with other spatial omics technologies such as spatial transcriptomics, epigenomics[25] or multiplexed imaging. In conclusion, scDVP is a powerful tool for basic discovery science, working in concert with DVP and other omics methods to enrich spatial workflows.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-02007-6.

References

1. Kelly, R. T. Single-cell proteomics: progress and prospects. *Mol. Cell Proteomics* **19**, 1739–1748 (2020).
2. Brunner, A.-D. et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* **18**, e10798 (2022).
3. Derks, J. et al. Increasing the throughput of sensitive proteomics by plexDIA. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01389-w (2022).
4. Mund, A. et al. Deep visual proteomics defines single-cell identity and heterogeneity. *Nat. Biotechnol.* **40**, 1231–1240 (2022).
5. Cunningham, R. P. & Porat-Shliom, N. Liver zonation—revisiting old questions with new technologies. *Front. Physiol.* **12**, 732929 (2021).
6. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
7. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
8. Ben-Moshe, S. et al. Spatial sorting enables comprehensive characterization of liver zonation. *Nat. Metab.* **1**, 899–911 (2019).
9. Guilliams, M. et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* **185**, 379–396.e38 (2022).
10. Bache, N. et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell Proteomics* **17**, 2284–2296 (2018).
11. Thielert, M. et al. Robust dimethyl-based multiplex-DIA workflow doubles single-cell proteome depth via a reference channel. *Mol. Syst. Biol.* https://doi.org/10.15252/msb.202211503 (2023).
12. Tsai, C.-F. et al. Surfactant-assisted one-pot sample preparation for label-free single-cell proteomics. *Commun. Biol.* **4**, 1–12 (2021).
13. Sandow, J. J., Infusini, G., Dagley, L. F., Larsen, R. & Webb, A. I. Simplified high-throughput methods for deep proteome analysis on the timsTOF Pro. Preprint at *bioRxiv* https://doi.org/10.1101/657908 (2021).
14. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
15. Rath, S. et al. MitoCarta3.0: an updated mitochondrial proteome now with suborganelle localization and pathway annotations. *Nucleic Acids Res.* **49**, D1541–D1547 (2021).
16. Droin, C. et al. Space–time logic of liver gene expression at sublobular scale. *Nat. Metab.* **3**, 43–58 (2021).
17. Guzmán, M., Bijleveld, C. & Geelen, M. J. H. Flexibility of zonation of fatty acid oxidation in rat liver. *Biochem. J.* **311**, 853–860 (1995).
18. Kang, S. W. S. et al. A spatial map of hepatic mitochondria uncovers functional heterogeneity shaped by nutrient-sensing signaling. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.13.536717 (2023).
19. Robles, M. S., Humphrey, S. J. & Mann, M. Phosphorylation Is a central mechanism for circadian control of metabolism and physiology. *Cell Metab.* **25**, 118–127 (2017).
20. Rosenberger, F. A., Thielert, M. & Mann, M. Making single-cell proteomics biologically relevant. *Nat. Methods* **20**, 320–323 (2023).
21. Makhmut, A. et al. A framework for ultra-low input spatial tissue proteomics. Preprint *bioRxiv* at https://doi.org/10.1101/2023.05.13.540426 (2023).
22. Oliinyk, D., Will, A., Schneidmadel, F. R., Humphrey, S. J. & Meier, F. µPhos: a scalable sensitive platform for functional phosphoproteomics. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.04.535617 (2023).
23. Segovia-Miranda, F. et al. Three-dimensional spatially resolved geometrical and functional models of human liver tissue reveal new aspects of NAFLD progression. *Nat. Med.* **25**, 1885–1893 (2019).
24. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
25. Deng, Y. et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).

Article https://doi.org/10.1038/s41592-023-02007-6

Methods

Mouse experiments and organ collection

Pathogen-free male and female 10-week-old C57BL/6J-rj mice were purchased from Janvier and maintained at the appropriate biosafety level under constant temperature and humidity conditions with a 12 h light cycle. Animals were allowed food and water ad libitum. All experiments were performed on 12- or 13-week-old wild-type mice. These were killed by cervical dislocation, and the liver was rapidly excised through a ventral opening of the peritoneum. The organ was rinsed in cold phosphate-buffered saline (PBS), and the left lateral lobe was divided into three pieces. For this study, the distal-caudal quarter was embedded in optimal cutting temperature medium (Sakura Finetek) in 15 mm disposable cryomolds (Sakura Finetek) and frozen in isopentane that was precooled to dew point in liquid nitrogen. Fully solidified blocks were transferred to dry ice, and then to a −80 °C freezer until further processing. Animal handling and organ withdrawal were performed in accordance with the governmental and international animal welfare guidelines and ethical oversight by the local government for the administrative region of Upper Bavaria (Germany), registered under ROB-55.2-2532.Vet_02-16-208.

Immunofluorescence staining

Two micrometer polyethylene naphthalate membrane slides were pretreated by ultraviolet ionization for 1 h at 254 nm. Without delay, slides were consecutively washed for 5 min each in 350 ml acetone and 7 ml VECTABOND reagent to 350 ml with acetone, and then washed in ultrapure water for 30 s before drying in a gentle nitrogen air flow. For sectioning, tissue blocks were transferred to a cryostat (Leica CM3050) at −18 °C chamber and −15 °C object temperature, and left to equilibrate for 30 min. Blocks were then trimmed, and final sections were cut at 10 µm thickness with a disposable high-profile blade (Leica 818). Frozen sections were transferred to pretreated, cold polyethylene naphthalate membrane slides, and melted for less than 5 s on a room temperature surface. The sections were then fixed in prewarmed 4% paraformaldehyde in PBS at 37 °C, then in 95% ethanol at room temperature and finally again in 4% paraformaldehyde in PBS at 37 °C. Slides were rinsed in PBS and left in 5% BSA–PBS blocking solution for 1 h until staining. Sections were stained for 1 h at 37 °C in a humid and dark chamber with 200 µl of a one-step liver painting in 1% BSA: 1:300 phalloidin coupled to Atto-425 (Sigma 66939), 1:200 wheat germ agglutinin (WGA) coupled to Alexa Fluor 488 (Invitrogen, W11261), 1:100 anti-E-cadherin coupled to Alexa Fluor 555 (BD 560064), anti-glutamine synthase (Abcam, ab176562) and 1:500 anti-rabbit nanobody coupled to Alexa Fluor 647 (Chromotek srbAF647-1-100). Slides were washed three times for 2 min in PBS in the dark, and mounted with 21 FL ProLong Diamond mounting medium (Invitrogen, P36961) and a 22 × 22 mm #1.5 coverslip. Slides were stored until imaging in 50 ml tubes with desiccating material at 4 °C.

High-content imaging

Sections were imaged on an OperaPhenix high-content microscope, controlled with Harmony v4.9 software, at 40× magnification, with binning of two and a per tile overlap of 10%. At excitation wavelengths of 425 nm, 555 nm and 647 nm, an 80% laser intensity were used at an illumination time of 100 ms, while in the 488 nm (CFP) channel, 20% and 20 ms were used. E-cadherin and glutamine synthetase were imaged simultaneously, while phalloidin and WGA were imaged consecutively.

Image postprocessing

Acquired images were flat-field corrected using the Harmony software. Stitching of image tiles was performed using the ashlar Python API (application programing interface)[26] with a max shift value of 30. Stitched images were exported as .tif files and imported into the Biological Image Analysis Software (BIAS, Single-Cell Technologies Ltd.)[4] with the packaged import tool. In BIAS, large tif images were first retiled to 1,024 × 1,024 px at an overlap of 5%. Hepatocytes were identified with a deep neural network for histological cytoplasm segmentation on the basis of phalloidin staining at 1.2 input spatial scaling, 40% detection confidence and 30% contour confidence. Only contours between 135 µm^2 and 1,350 µm^2 were taken into consideration, and no further exclusion criteria were applied. After removal of duplicates and false identifications by supervised machine learning, contours were exported without additional shape offset together with three calibration points that were chosen at characteristic tissue positions. Contour outlines were simplified by removing 99% of data points. For five-shape proteomes, directly adjacent shapes forming a pentagon-like structure were manually picked. Single shapes were randomly picked and every 15–25th shape was assigned to adjacent wells in a 384-well plate. Arterioles were manually assigned based on WGA signal, ellipticity and proximity to the E-cadherin positive PV.

Laser microdissection

Contour outlines were imported after reference point alignment, and shapes were cut by laser microdissection with the LMD7 (Leica) in a semi-automated manner at the following settings: power 59, aperture 1, speed 60, middle pulse count 1, final pulse −1, head current 48–52%, pulse frequency 3,282 and offset 100. For the five-shape experiment, the microscope was controlled with LMD v8.2, with which five directly adjacent shapes were sorted into a low-binding 384-well plate (Eppendorf 0030129547) with one empty well between samples. Single shapes were cut and sorted with the software LMD beta 10 after calibration of the gravitational stage shift into 384-well plates into all wells, leaving the outermost rows and columns empty. A 'wind shield' plate was used on top of the sample stage to avoid erroneous sorting. Plates were sealed, centrifuged at 1,000*g* for 5 min and then frozen at −20 °C until further processing.

Reference peptide preparation for five-shape and single-shape proteomes

The proximal part of two biologically independent lobes of the same mice as in the scDVP experiments was used to construct a library. The tissue embedded in optimal cutting temperature medium was removed from −80 °C and directly disintegrated in a plastic bag with a manual tissue homogenizer (rubber hammer). Pieces of approximately 1 mm^3 were transferred into a low-binding 96-well plate with magnets (BeatBox Tissue Kit, Preomics), covered with 50 µl of 60 mM triethylammonium bicarbonate buffer with 10% acetonitrile (ACN; lysis buffer), and lysed in a BeatBox (Preomics) at standard settings for 10 min. Samples were then boiled at 96 °C for 20 min, transferred to 1.5 ml low-binding tubes, filled up to 500 µl with lysis buffer and sonicated for five times 30 s on/off cycles. After centrifugation at 2,000*g* for 1 min, the protein concentration in the supernatant was estimated on a NanoDrop, and LysC and trypsin were added at a protein-to-enzyme ratio of 1:100. After digest for 20 h, samples were acidified to 1% trifluoroacetic acid (TFA), centrifuged at 3,000*g* for 10 min at room temperature, and dried in a SpeedVac for 30 min. Digest was filled to 1 ml with buffer A (0.1% formic acid (FA)), and desalted on C-18 columns (Waters WAT036820). They were activated and equilibrated with 2 ml of methanol, 2 ml of buffer B (100% ACN, 0.1% FA) and 2 ml of buffer A, before sample loading. Peptides were washed with buffer A two times, eluted in 80% ACN with 0.2% FA and dried down.

Library fractionation for five-shape proteomes

Peptides were reconstituted in 18 µl buffer A* (0.1% FA, 2% ACN) fractionated on a 30-cm-long 1.9 µm ReproSil C-18 column (PepSep) using a 100 min high-pH gradient. The concentration of buffer B was increased from 3% to 30% in 45 min, to 40% in 12 min, to 60% in 5 min, to 95% in 10 min, kept constant for 10 min, reduced to 5% in 10 min and kept constant for 8 min. The eluted peptides were automatically collected into 48 fractions with a concatenation time of 90 s per fraction. The

fractions were dried in a SpeedVac, reconstituted in 0.1% FA and directly loaded onto Evotips.

Labeling of single-shape reference proteome

Peptides were reconstituted to 0.125 µg µl^{-1} in 60 mM triethylammonium bicarbonate buffer with 10% ACN, pH 8.5. The peptides were then dimethyl labeled with 0.15% light formaldehyde (CH_2O) and 0.023 M light sodium cyanoborohydrate ($NaBH_3CN$) for 1 h at room temperature, quenched with 0.13% ammonia and acidified to 1% TFA. After drying in a SpeedVac, pellets were reconstituted in 100 µl buffer A, and desalted via 5 µg C-18 columns on an AssayMap (Agilent) following the standard protocol. The resulting reference peptides were dried, and reconstituted to 1 ng µl^{-1} in buffer A.

Peptide preparation of single shapes and dimethyl labeling for multiplexing

Peptides were prepared semi-automated on a Bravo pipetting robot (Agilent), similarly to as described previously[11]. During each incubation step, plates were tightly sealed with two stacked aluminum lids to avoid evaporation (Thermo Fisher Scientific, AB0626). For this, plates were removed from the freezer and centrifuged. The wells were then washed on the robot with 28 µl of 100% ACN and dried in a SpeedVac (Eppendorf) at 45 °C for 20 min. Shapes were then resuspended in 6 µl of 60 mM triethylammonium bicarbonate buffer (pH 8.5, Sigma) with 0.013% DDM (Sigma), and cooked for 30 min at 95 °C in a PCR cycler at a lid temperature of 110 °C. After addition of 1 µl of 80% ACN (final concentration 10%), samples were incubated for another 30 min at 75 °C, cooled briefly, and 1 µl with 4 ng LysC and 6 ng trypsin was added. The samples were digested for 18 h, and then 1 µl of either intermediate (CD_2O) or heavy formaldehyde ($^{13}CD_2O$) was added to a final concentration of 0.15%. Without delay, either light ($NaBH_3CN$) or heavy ($NaBD_3CN$) sodium cyanoborohydrate were added to 0.023 M to retrieve Δ4 and Δ8 dimethyl-labeled single-shape samples. The sealed plate was then incubated at room temperature for 1 h, and the reaction was quenched to 0.13% ammonia and acidified to 1% TFA.

Peptide loading onto C-18 tips

C-18 tips (Evotip Pure, EvoSep) were activated for 5 min in 1-propoanl, washed twice with 50 µl of buffer B (99.9% ACN, 0.1% FA), activated for 5 min in 1-propanol, and washed twice with 50 µl buffer A (0.1% formic acid). Single-shape samples were then loaded automatically with the Agilent Bravo robot into 30 µl buffer in the tip that was spun through the C-18 disk for a few seconds only. For loading, 10 µl of 1 ng µl^{-1} reference peptides (Δ0) were pipetted first, followed by Δ4, and Δ8 samples with the same tip. Wells were rinsed with 15 µl buffer A that was also loaded onto the tip. After peptide binding, the disk was further washed once with 50 µl buffer A and then overlayed with 150 µl buffer A. All centrifugation steps were performed at 700*g* for 1 min, except sample loading for 2 min.

For five-shape proteomes, plates were treated as above, with the exception of lysis in 4.5 µl 60 mM triethylammonium bicarbonate buffer, pH 8.5 without DDM, and consecutive addition of 1 µl LysC and 1.5 µl trypsin to achieve the same digestion volume as above. Five-shape samples were not dimethyl labeled and multiplexed, but acidified directly after digest, and loaded manually onto Evotips following the protocol described above.

LC–MS/MS analysis of five shapes

Samples were measured with the Evosep One LC system (EvoSep) coupled to a timsTOF SCP mass spectrometer (Bruker Daltonics). The 30 samples per day method was used with the Evosep Performance column 15 cm, 150 µm ID (EV1137 EvoSep) at 40 °C inside a nanoelectrospray ion source (Bruker Daltonics) with a 10 µm emitter (ZDV Sprayer 10, Bruker Daltonics). The mobile phases were 0.1% FA in liquid chromatography (LC)–MS-grade water (buffer A) and 99.9% ACN/0.1% FA (buffer B). We used a dia-PASEF method with 16 dia-PASEF scans separated into four ion mobility windows per scan covering an m/z range from 400 to 1,200 by 25 Th windows and an ion mobility range from 0.6 to 1.6 V s cm^{-2} ('standard scheme'[27]). The mass spectrometer was operated in high sensitivity mode, with an accumulation and ramp time at 100 ms, capillary voltage set to 1,750 V and the collision energy as a linear ramp from 20 eV at $1/K_0$ = 0.6 V s cm^{-2} to 59 eV at $1/K_0$ = 1.6 V s cm^{-2}.

LC–MS/MS analysis of single shapes

Samples were measured with the Evosep One LC system (EvoSep) coupled to a timsTOF SCP mass spectrometer (Bruker Daltonics). The Whisper40 samples per day method was used with the Aurora Elite CSI third generation 15 cm and 75 µm ID (AUR3-15075C18-CS IonOpticks, Australia) at 50 °C inside a nanoelectrospray ion source (Bruker Daltonics). The mobile phases were 0.1% formic acid in LC–MS-grade water (buffer A) and 99.9% ACN/0.1% FA (buffer B). The timsTOF SCP was operated with an optimal dia-PASEF method generated with our Python tool py_diAID[28]. The method contained eight dia-PASEF scans with variable width and two ion mobility windows per dia-PASEF scan, covering an m/z from 300 to 1,200 and an ion mobility range from 0.7 to 1.3 V s cm^{-2}, as previously used on the same gradient and similar input material amount[11]. The mass spectrometer was operated in high sensitivity mode, with an accumulation and ramp time at 100 ms, capillary voltage set to 1,400 V and the collision energy as a linear ramp from 20 eV at $1/K_0$ = 0.6 V s cm^{-2} to 59 eV at $1/K_0$ = 1.6 V s cm^{-2}.

The labeling efficiency was accessed on the same LC–MS/MS in data-dependent acquisition (dda)-PASEF mode with ten PASEF scans per topN acquisition cycle. Singly charged precursors were excluded by their position in the m/z-ion mobility plane using a polygon shape, and precursor signals over an intensity threshold of 1,000 arbitrary units were picked for fragmentation. Precursors were isolated with a 2 Th window below m/z 700 and 3 Th above, as well as actively excluded for 0.4 min when reaching a target intensity of 20,000 arbitrary units. All spectra were acquired within a m/z range of 100–1,700. All other settings were kept as described before.

Spectral library generation

The spectral library was generated on five dda-PASEF single shots from 50 ng mouse reference peptide, using the same chromatography method as above. Spectra were search with FragPipe v18.0 (ref. 29) using MSFragger v3.5, Philosopher v4.4.0 and EasyPQP v0.1.32 against a mouse FASTA reference file with 55,319 entries used throughout this study, excluding 50% decoys. Standard settings of the DIA_SpecLib_Quant workflow were used with the following exceptions: N-terminal and lysine mass shift of 28.0313 Da were set as fixed modifications, and methionine oxidation as variable modification. Carbamidomethylation was unselected as samples were not reduced and alkylated. One missed cleavage was accepted. The precursor charge ranged from 2 to 4. The peptide mass range was set to 300–1,800, and peptide length from 7 to 30. For DIA-NN compatibility, the column 'FragmentLossType' was removed in the output library file.

Spectral search

All 263 files were search together in DIA-NN (version 1.8.1) (ref. 30) against the above-generated library, using a mass and MS1 mass accuracy of 15.0, scan windows of 9, and activated isotopologues, Match-between-Runs (MBR), heuristic protein inference and no shared spectra, in single-pass mode. Proteins were inferred from genes. Library generation was set as 'IDs, RT & IM profiling', and 'Robust LC (high precision)' as the quantification strategy. Dimethyl labeling at N-termini and lysins was set as fixed modification at 28.0313 Da, and Δ4 or Δ8 were spaced 4.0251 Da or 8.0444 Da from the reference Δ0 ({–fixed-mod Dimethyl, 28.0313, nK} and {–channels Dimethyl, 0, nK, 0:0; Dimethyl, 4, nK, 4.0251:4.0251; Dimethyl, 8, nK, 8.0444:8.0444}). Additional settings were {–original-mods}, {–peak-translation}, {–ms1-isotope-quant}, {–report-lib-info}.

Data analysis

(1) RefQuant: to determine the quantities of the precursors in the DIA-NN report.tsv file, we utilized the Python-based RefQuant algorithm[11]. In brief, RefQuant determines the ratio between target and reference channels for each individual fragment ion and MS1 peak that is available. This gives a collection of ratios from which RefQuant estimates a likely overall ratio between target and reference. The ratio between target and reference was rescaled by the median reference intensity over all runs for the given precursor, thereby giving a meaningful intensity value for the target channel. The RefQuant quantification matrix was filtered for 'Lib.PG.Q.Value' <0.01, 'Q.value' <0.01 and 'Channel.Q.Value' <0.15 and was then collapsed to protein groups using the MaxLFQ algorithm[31] as implemented in the R package iq (version 1.9.6) (ref. 32) with median normalization turned off.

(2) Sample filtering and normalization: protein group data were then further analyzed in R v4.2.1 operating in RStudio v2022.07.2. Samples were excluded if the number of detected proteins was below 1.5 or above 3 s.d. from the sample identification median, or within (806, 3,362) identified proteins, resulting into a dropout of 8.9% (41 of 459 samples). Four samples were removed due to their outlier position on the PCA, see Supplementary Table S3. Eight samples were removed due to their cell sizes larger than the BIAS cutoff of 1,350 µm^2. This resulted in 406 included samples, of which 400 were hepatocytes and 6 endothelial structures for validation. After sample filtering, data was median normalized to a set of proteins that were quantified across all samples (175 proteins quantified in 100% of included samples; Supplementary Table S3), thus correcting for the dependence of protein numbers on shape size. For hepatocyte specific analysis, the arteriole proteomes were removed before normalization.

(3) Figure generation: we chose 20 classes for all comparative spatial analyses as this matches the approximate number of cells from PV to CV, and five classes for machine learning as a compromise between meaningful separation and having enough samples per class. Proteome bins were based on an equidistant split of PC1, distance classes accordingly on a split of PV over CV distance ratios, and applied as indicated. PCA were performed with the PCAtools v2.8.0 package. Limma v3.52.4 was used for statistical testing across proteome bins on a 50%-complete protein data matrix. 'FDR' was applied for multiple testing correction, and an FDR cutoff of 5% was considered significant. Heat mapping was performed with pheatmap 1.0.12, the completeness of the data matrix is indicated in the figure legends. Proteomic gene set enrichment analyses were done with WebGestalt 2019 (ref. 33) in an R environment using Kyoto Encyclopedia of Genes and Genomes metabolic pathways or Wikipathway as functional library and an FDR threshold for reporting of 1. Significance was defined as FDR <10%, and normalized enrichment scores are reported here. Subcellular localization and mitochondrial functional protein sets were retrieved from mouse Mitocarta 3.0 (ref. 15). Urea cycle and peroxisomal fatty acid degradation proteins were manually curated. Normality was assessed with Shapiro–Wilk's test, and *P* values were corrected for multiple testing and expressed as FDR. Spatial data from xml files was plotted with the package sf v1.0-9. For comparisons to scRNAseq data, the dataset of Halpern et al.[6] was used, for which we binned the proteome data into nine equidistant spatial bins as described above. We used the dataset by Ben-Moshe et al.[8] to compare scDVP data with FACS-based proteomics data, binning our samples into eight equidistant spatial bins.

Image processing

Image data analysis was done in Python (3.8.11). Image shapes were extracted from the stitched tiles using Pillow (9.0.0). For each shape, the bounding box was calculated by taking the floor and ceiling of each shape coordinate and taking the maximum and minimum in x and y. The bounding rectangle was used to crop out the respective region of interest of the image. For image with offset extraction, the center of each bounding rectangle was calculated and rounded to the next integer. An offset of 1,000 was added to each direction to additionally capture the surrounding environment, and the bounding box was highlighted. For composite images, each image was exported per channel with matplotlib (3.5.1), reloaded, merged with NumPy (1.4.2) and saved again. ImageJ was used to manually measure the distance of a shape to its proximal PV and CV.

Machine learning

For each shape and in all four channels (cyan fluorescent protein, Alexa488, Alexa568 and Alexa647), the mean, median, minimum and maximum intensity of each bounding box were calculated, as well as the shape area. This feature list was saved with pandas (1.22.3). Proteomics data were clustered with a k-means algorithm into five clusters. Next, we used a supervised learning approach to classify the proteomic clusters based on the feature list. The training was performed using the scikit-learn package (1.0.2). Data (n = 408) were randomly split in train and test datasets (split of 0.2). For classification, we used a RandomForestClassifier (n_estimators=200) and achieved a testing accuracy of 0.90. To export probabilities, we used the predict_proba functionality of RandomForest. Diagnostic plots were generated using the Yellowbrick package (1.5). The random state was set to 23 for train/test-split and RandomForestClassifier.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[34] partner repository with the ID PXD038699. Imaging data has been deposited to BioImages[35] with the accession number S-BIAD596.

Code availability

The R and Python code used to produce the figures can be downloaded from the Mann lab Github repository via https://github.com/MannLabs/single-cell-DVP.

References

26. Muhlich, J. L. et al. Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac544 (2022).
27. Meier, F. et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
28. Skowronek, P. et al. Rapid and in-depth coverage of the (phospho-)proteome with deep libraries and optimal window design for dia-PASEF. *Mol. Cell. Proteomics* **21**, 100279 (2022).
29. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
30. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

31. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ*. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
32. Pham, T. V., Henneman, A. A. & Jimenez, C. R. iq: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics. *Bioinformatics* **36**, 2611–2613 (2020).
33. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
34. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
35. Hartley, M. et al. The BioImage Archive—building a home for life-sciences microscopy data. *J. Mol. Biol.* **434**, 167505 (2022).

Acknowledgements

We thank our colleagues at the Department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry as well as our colleagues at the Center for Proteome Research in Copenhagen for their input and support. We are particularly grateful for input and help from I. Bludau, A. Brunner and L. Zeitler. We thank Peter H. and Single-Cell Technologies Ltd. for their technical support. F.A.R. is an EMBO postdoctoral fellow (ALTF 399-2021). S.C.M. is a PhD fellow of the Boehringer Ingelheim Fonds. J.G.-S. has received funding from the European Respiratory Society and the European Union's H2020 research and innovation program under the Marie Sklodowska-Curie RESPIRE4 grant agreement no. 847462. This study has been supported by the Horizon-2020 under the MICROB-PREDICT program (M.M., no. 825694) and ISLET (M.M., no. 874839), by the Max Planck Society for Advancement of Science (M.M.), by the Chan Zuckerberg Initiative (M.M., CZF2019-002448) and by grants from the Novo Nordisk Foundation, Denmark (M.M., grant agreements NNF14CC0001 and NNF15CC0001). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

F.A.R., M.T. and M.M. conceptualized the scDVP workflow. M.T., F.A.R., P.S. and M.W. acquired MS data. F.A.R., M.T.S., L.S., A. Metousis and K.M. performed data analysis. M.T.S. trained the ML model. C.A. developed quantification software. S.C.M. optimized the high-content microscopy pipeline. F.A.R., M.T., L.S., A. Metousis, S.C.M., E.R., T.M.N. and A. Mund developed and optimized the experimental scDVP workflow. J.G.-S., A.S. and H.B.S. provided mouse samples. F.A.R., M.T.S., P.S. and T.M.N. curated data. M.M. and F.A.R. supervised the project. F.A.R. and M.M. wrote the original manuscript draft. All authors read, revised and approved the manuscript.

Funding

Open access funding provided by Max Planck Society

Competing interests

M.M. is an indirect investor in Evosep. All other authors declare no competing interests.

Additional information

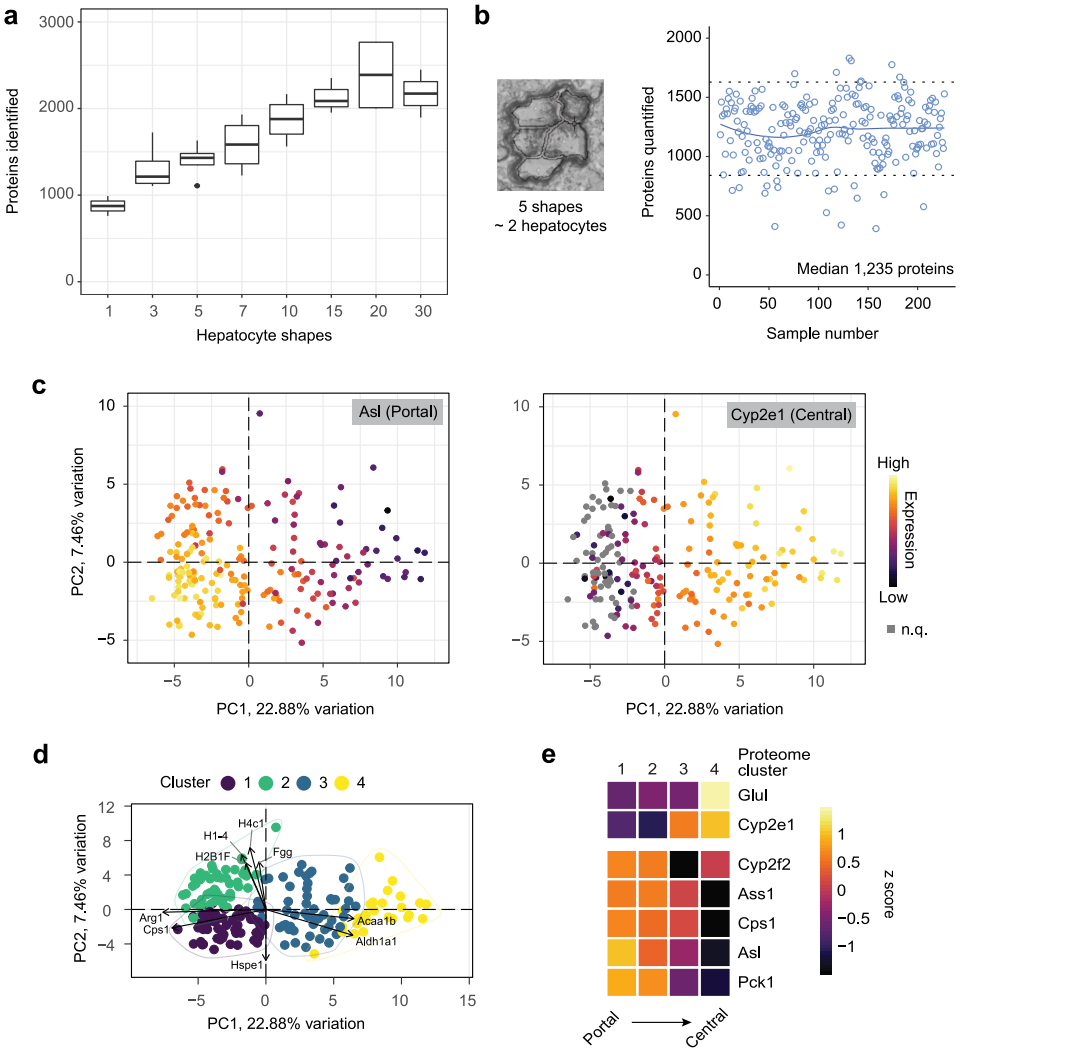
Extended data is available for this paper at https://doi.org/10.1038/s41592-023-02007-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-02007-6.

Correspondence and requests for materials should be addressed to Matthias Mann.

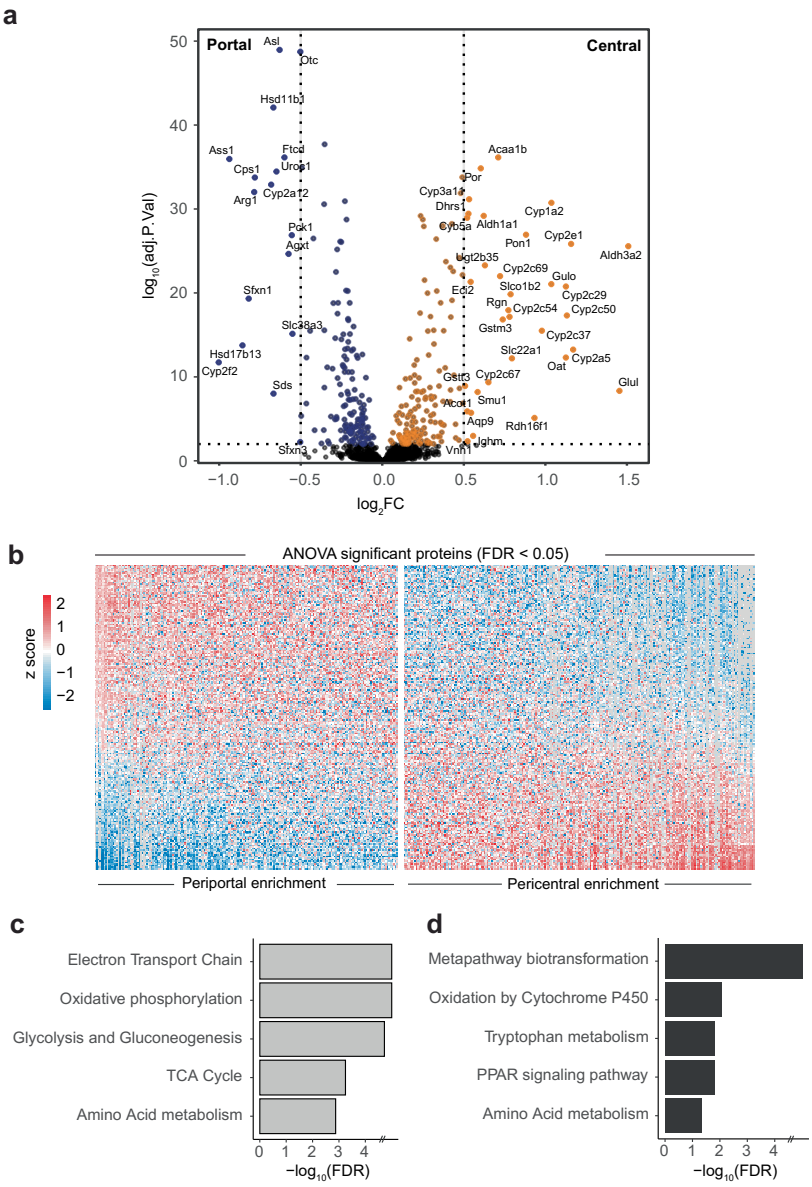
Peer review information *Nature Methods* thanks Albert Heck, Shalev Itzkovitz, and Sinem Saka for their contribution to the peer review of this work. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

Article https://doi.org/10.1038/s41592-023-02007-6



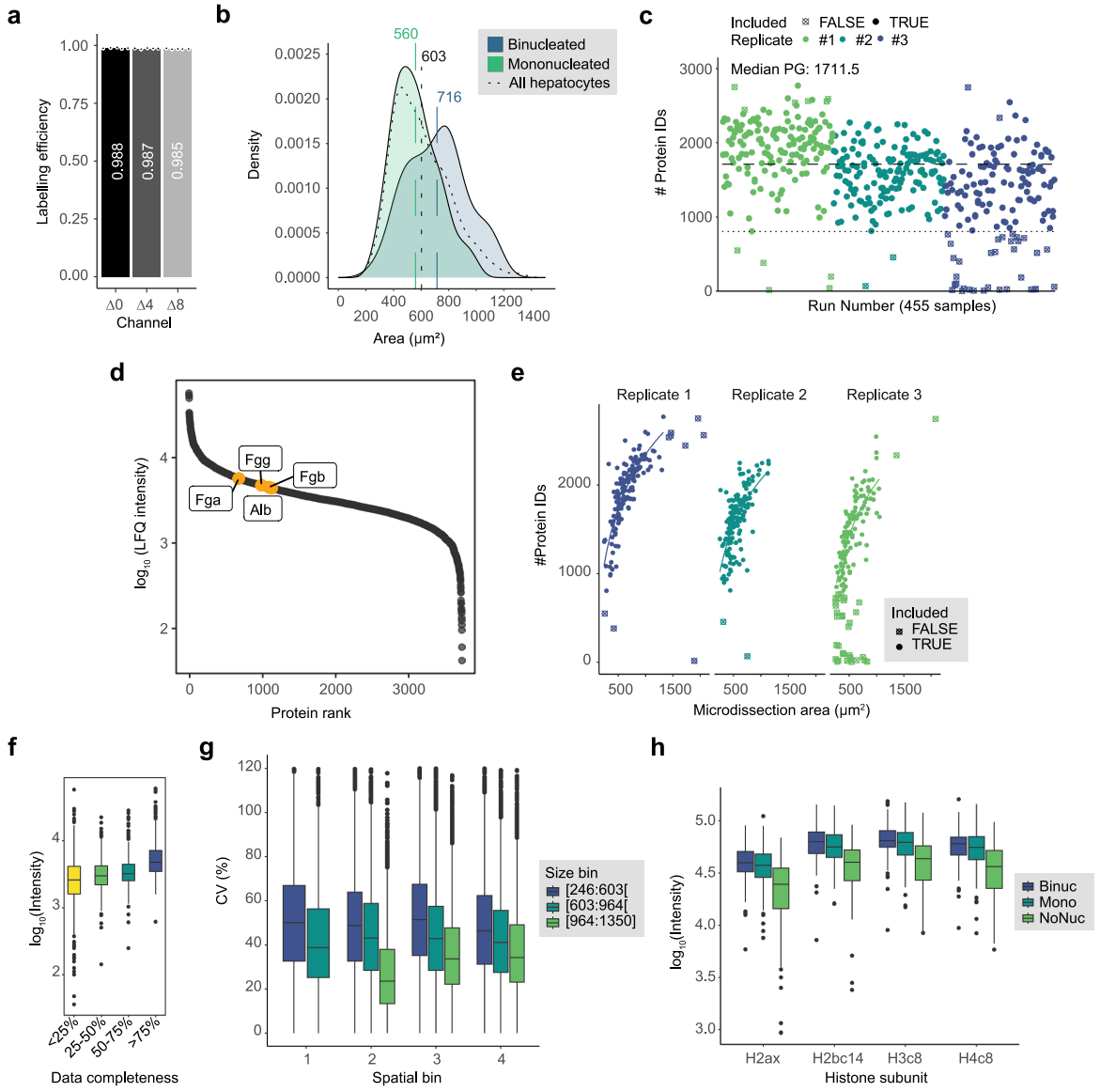
Extended Data Fig. 1 | Five shape proteomes resolve liver zonation. **a**, Titration of number of shapes (10 µm thick) versus proteome depth achieved (n = 3), and measured with the original protocol (single shape, 44 min Evosep gradient, 15 cm column at 500 nL/min, dia-PASEF 27 without optimized windows, library-dependent search in DIA-NN 30). Boxes are first and third quartile, the thick line is median, whiskers are ± 1.5 interquartile range, and outliers are indicated as individual points. **b**, Protein numbers per five shapes across 230 samples. Line is a smoothing curve. **c**, Principal component analyses with a color overlay of two indicated zonation markers; n.q. not quantified. **d**, Unbiased k means clustering of all samples into four bins. Labeled arrows are the top driver proteins of separation. **e**, Marker expression sorted by central (top) or portal (bottom) markers in the indicated k means clusters in d, expressed as z-score of log2 transformed protein abundances, and sorted according to summed zonal probability across all markers.

Article https://doi.org/10.1038/s41592-023-02007-6



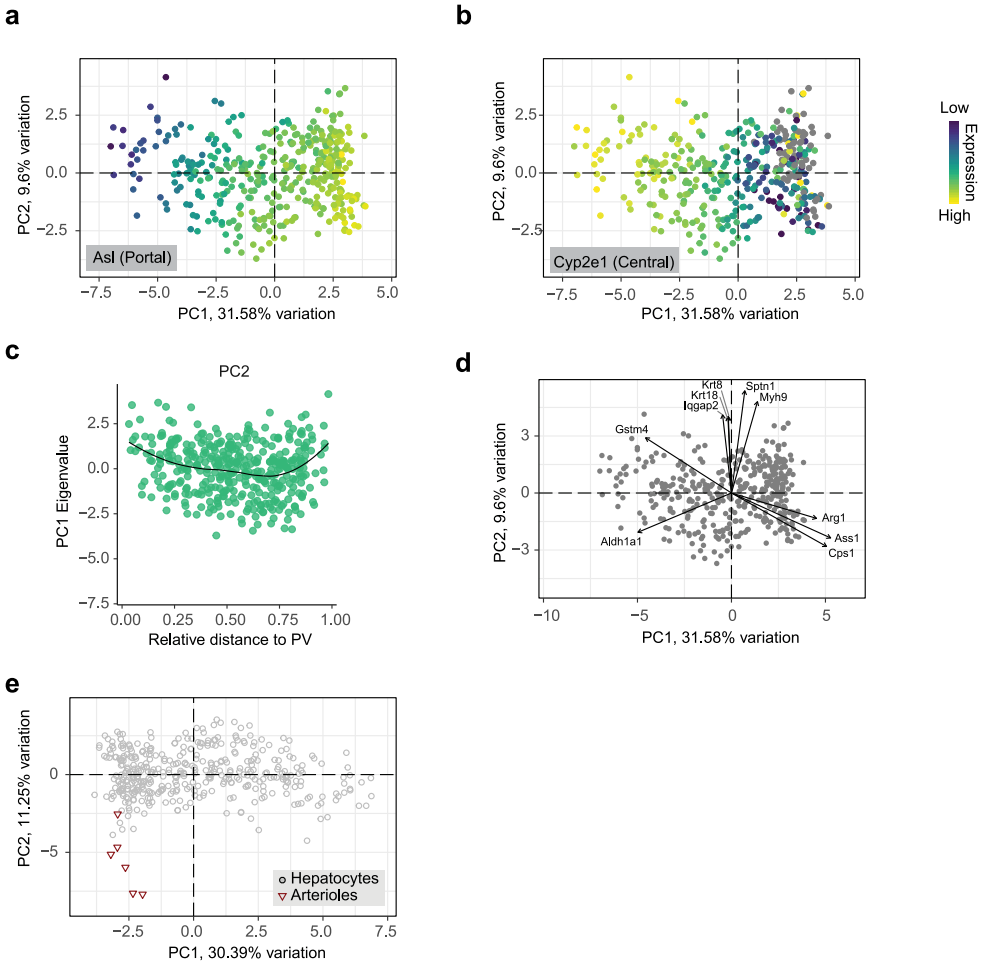
Extended Data Fig. 2 | Statistical analysis of five-shape proteomes. a, Volcano plot after an ANOVA over four sorted k means clusters (see Extended Date Fig. 1d). Statistically significant proteins (FDR < 0.05, n = 333 of 1652) with an absolute fold change of more than 0.5 are labeled. Colors indicate upregulation towards portal, or central zones. **b**, Heatmapping of statistically significant proteins in a. The blocks are separate by negative, or positive fold change. Protein expression as z-score of log2 transformed protein abundances. **c**,**d**, Five top significant terms (FDR < 0.05) after over-representation analysis enriched in peri-portal (**c**) or peri-central regions (**d**). See Supplementary table S2 for further reference.

Article https://doi.org/10.1038/s41592-023-02007-6



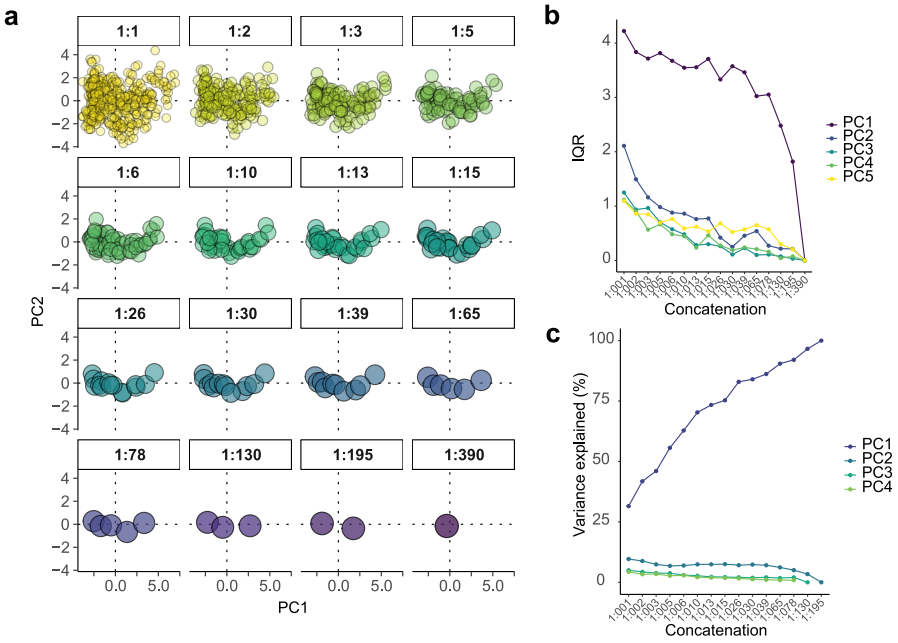
Extended Data Fig. 3 | Performance overview of single-shape proteomes. **a**, Labeling efficiency of 10 ng mouse liver peptide samples. Mean efficiency by intensity is stated in the bar (n = 5, mean and individual measurements). **b**, Density distribution of shape areas across all measured and included hepatocytes, split by visually distinguished mono- (N = 191) and binucleated (N = 99) hepatocytes. Vertical lines and numbers above are mean sizes in the respective group. **c**, Number of proteins per sample (N = 455). The dotted line is the median, the fine pricked line is the sample exclusion cutoff of median minus 1.5 standard deviations. Samples were measured from left to right. Shape type indicates whether the samples was included for the final analysis. **d**, Levels of plasma proteins in the scDVP dataset. Hba, Hbb and Hbd were not detected. **e**, Association between the area of the cut shape, and number of proteins. Line is a log10 regression curve. Symbols indicate whether sample was included or discarded for analysis, for exclusion criteria see Methods section. **f**, Percentage of proteins quantified, binned into four groups, versus log10 transformed median intensities in the respective bin. Data completeness is defined as percentage of samples across all samples in which a particular protein was quantified in. **g**, Coefficient of variation (CV) in bins of similarly sized shapes (color coded), and spatial bins with similar distance ratio to portal and central vein, that is similar zonation profile. **h**, Levels of four histone proteins shown in Fig. 2b by number of nuclei in the isolated shapes. Binuc: binucleated (N = 99); Mono: mononucleated (N = 191); NoNuc: no nucleus (N = 101). Boxes are first and third quartile, the thick line is median, whiskers are ± 1.5 interquartile range, and outliers are indicated as individual points.

Article https://doi.org/10.1038/s41592-023-02007-6



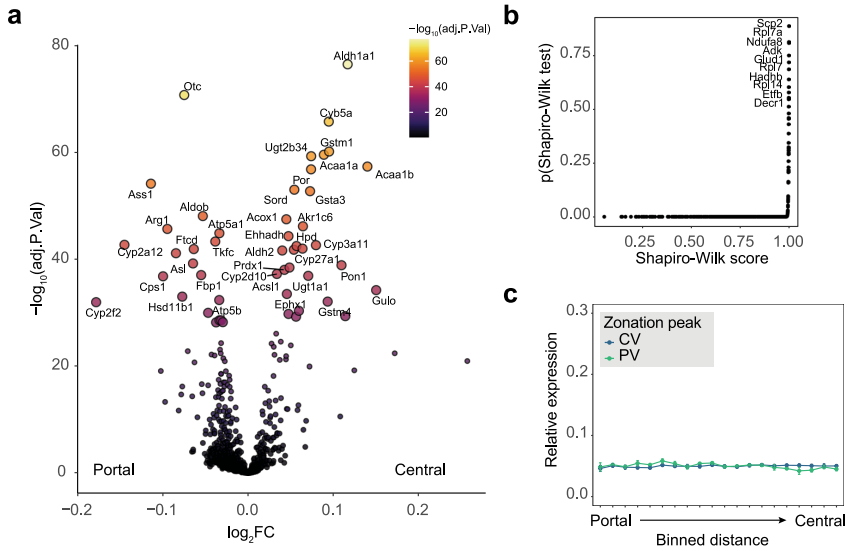
Extended Data Fig. 4 | Dimensionality reduction of single shape data. **a**, Color overlay is expression level of the portal marker Asl, or **b**, the central marker Cyp2e1. **c**, PC2 versus measured distance ratio portal over central vein for all shapes. **d**, Top 10-leading edges as Eigenvectors (arrows) with proteins. **e**, Arterioles were cut as quality controls (see Methods section), and separate from hepatocytes on PC2 (n = 6).

Article https://doi.org/10.1038/s41592-023-02007-6



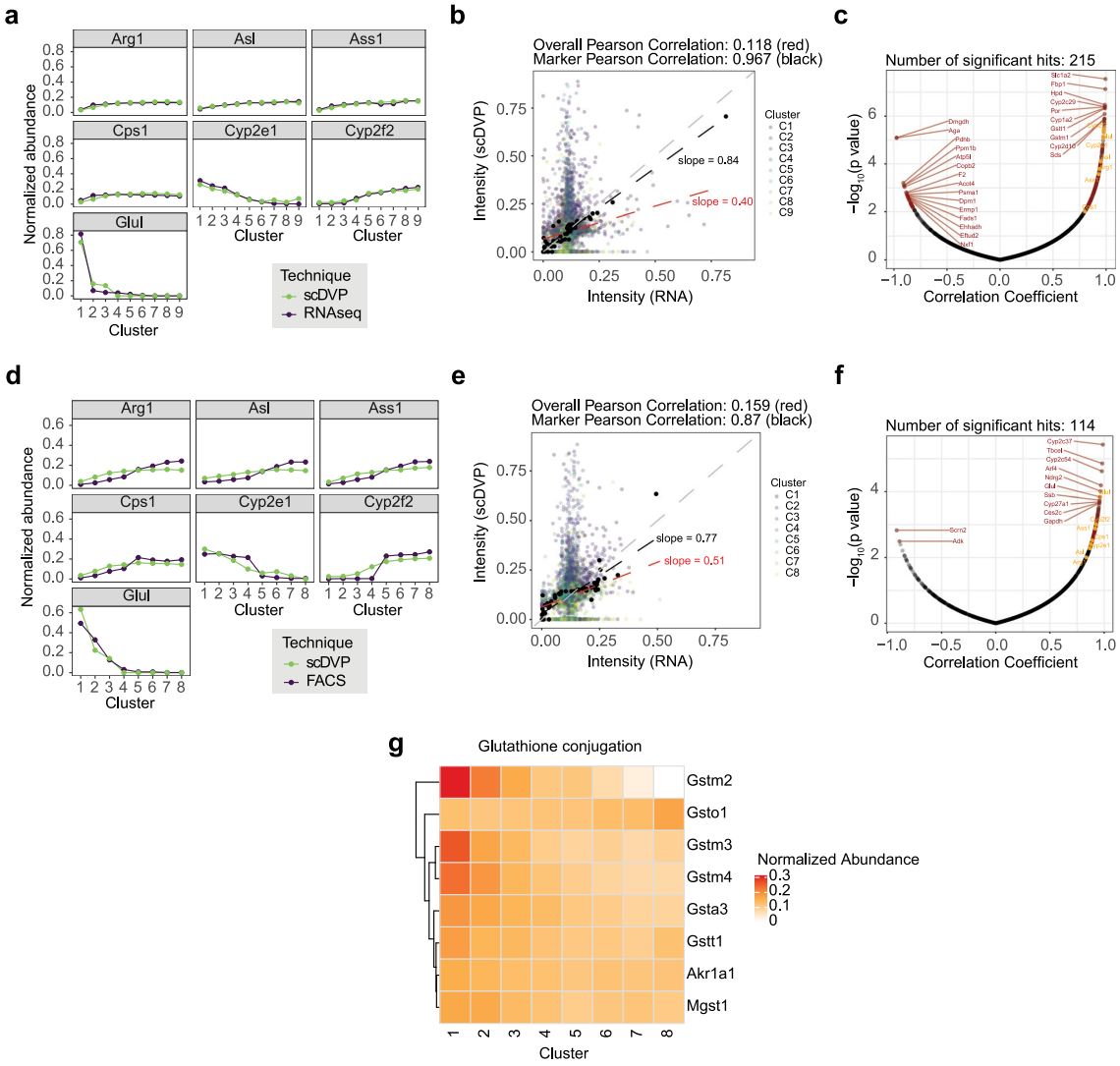
Extended Data Fig. 5 | Information aggregation from single shapes. **a**, Principal component analysis after averaging of close-by cells, as measured by relative position along the portal to central vein zonation axis. Ratios over every sub-plot indicate concatenation ratio (1:x averages x cells). **b**, Interquartile range (IQR) of principal components 1 – 5 at given concatenation ratio. **c**, Variance explained by the indicated principal component at given concatenation ratio.

Article https://doi.org/10.1038/s41592-023-02007-6



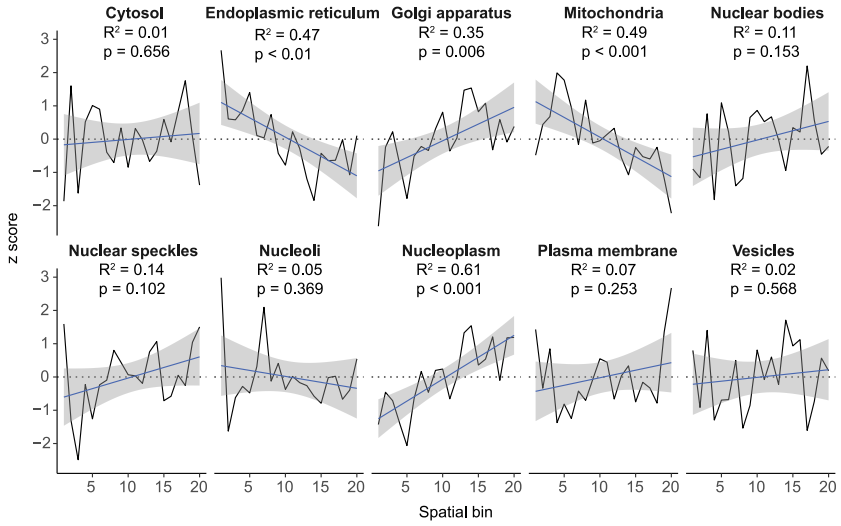
Extended Data Fig. 6 | Functional analysis of single shape data. a, Volcano plot after ANOVA across 20 spatially guided bins. Color overlay specifies adjusted p value, the top 40 significant proteins are labeled. **b**, Score and multiple testing-adjusted p value of a Shapiro-Wilk normality test. Lowest proteins are labeled. **c**, Relative expression normalized to 1 for each contributing protein (n = 10) of the least significant Shapiro-Wilk hits in b, from portal to central distance-guided bins.

Article https://doi.org/10.1038/s41592-023-02007-6



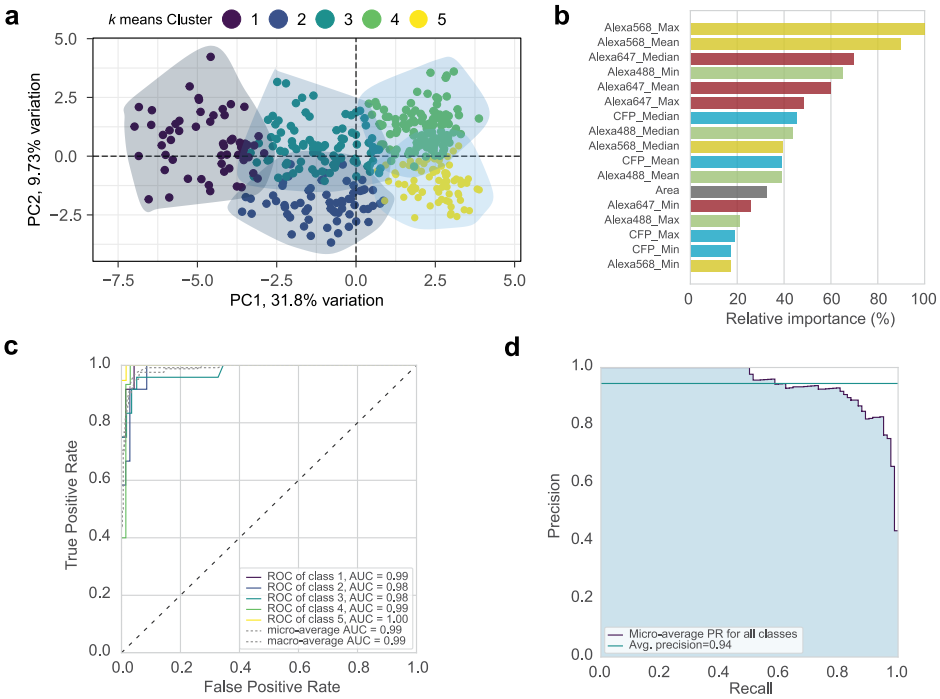
Extended Data Fig. 7 | Comparison of scDVP to existing scRNAseq data (a-c) and FACS-based proteomics data (d-g). a, Abundance normalized to 1 across 9 bins in Halpern et al. 6 (marker expression-guided bins), and this scDVP data (spatial bins). **b**, Intensity correlation of all hits (opaque dots, color according to cluster) and markers (black dots). Linear regression as dashed line, with Pearson correlation coefficient given over the figure. Grey line is the 45 degree line. **c**, Correlation coefficient for targets across all bins, with multiple testing adjusted p value. Top hits on either side are labeled in dark red, and marker proteins in orange. **d**, Abundance normalized to 1 across 8 bins in Ben-Moshe et al. 8 (marker expression-guided bins), and this scDVP data (spatial bins). **e**, Intensity correlation of all hits (opaque dots, color according to cluster) and markers (black dots). Linear regression as dashed line, with Pearson correlation coefficient given over the figure. Grey line is the 45 degree line. **f**, Correlation coefficient for targets across all bins, with multiple testing adjusted p value. Top hits on either side are labeled in dark red, and marker proteins in orange. **g**, A significant hit after gene set enrichment analysis on Pearson correlation coefficients, with normalized abundance of protein levels as heatmap colors.

Article https://doi.org/10.1038/s41592-023-02007-6



Extended Data Fig. 8 | Changes in subcellular compartment composition across space. Spatial bins are mean single shape data in 20 equidistant bins from portal to central vein. Ordinate values are z-transformed proportions of summed signal intensities per compartment. Pearson's R was calculated on z scores from a linear model. Blue line is the linear regression line with the 95% confidence interval in grey.

Article https://doi.org/10.1038/s41592-023-02007-6



Extended Data Fig. 9 | Machine learning (ML) accurately predicts proteome class. a, *k* means clustering, dividing all samples into five classes that inform the ML. **b**, Feature importance of the ML model, relative to the highest contributor. **c**, Receiver-Operating-Characteristics for each class. The individual Area Under the Curve (AUC) is given in the graph. **d**, Precision-recall-curve for the five classes.

4.3 SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes

Pooled forward genetic screens can characterise the function of individual genes in an organism's genome in an unbiased manner: After generating a library of mutations, those with interesting phenotypes are isolated and their genotype determined. These screens can be scaled up to cover the entire genome of an organism, allowing for the unbiased determination of all genes relevant for a specific biological process.

On the level of individual cells, mutations covering all protein-coding genes could be introduced via CRISPR/Cas systems. Each cell in the generated pool then carries a different mutation. While these types of screens scale to millions of cells to allow for genome-scale throughput, so far they have only been possible on comparatively simple phenotypes with limited descriptive power. Here we developed spatially resolved CRISPR screening (SPARCS), a technology to perform genome-scale CRISPR screens on microscopy-based phenotypes. In SPARCS, the generated mutant cell library is plated on polyphenylene sulfide (PPS) slides and imaged with a light microscope. This generates hundreds of millions of images describing different aspects of the spatial organisation of individual cells. Using deep learning, cells with interesting spatial phenotypes can be identified. By fully automating laser microdissection, we enabled the rapid and specific isolation of individual mutant cells from a library of millions of cells for subsequent genotyping. SPARCS is compatible with any light microscopy setup which makes it easily adoptable without the need for specialised imaging equipment. In our first SPARCS screen, we screened for the spatial distribution of autophagosomes in human U2OS cells. Autophagosomes are double-membraned vesicles that form, to help cells recycle their components, a process known as "autophagy" - for example in response to starvation. These vesicles are covalently decorated with proteins from the LC3 family. To follow the generation of autophagosomes we engineered U2OS cells to express an LC3 protein fused to the fluorescent protein mCherry. After building a mutant library covering the entire human genome in which one protein-coding gene is knocked out in each cell, we induced a starvation response and assessed their ability to form autophagosomes with a confocal microscope. With a deep learning classifier that we had trained to recognise autophagy-deficient cells, we were able to identify individual mutant cells with impaired autophagosome biogenesis. Excising these cells and sequencing their genetic knockout, we robustly identified almost all known autophagy regulators, demonstrating the power of SPARCS. We also identified a gene

called *EI24* that had so far only been described to cause the formation of spontaneous autophagosomes when knocked out, rather than a decrease in autophagosomes. The discovery of a novel phenotype for *EI24* in our screen is especially interesting as it establishes that even supervised deep learning approaches like the one developed here are capable of identifying previously undescribed phenotypes. Ultimately, SPARCS enables a new type of cell-based forward genetic screen on complex spatial phenotypes defined by microscopy.

The following research article was originally published here:

Schmacke, N. A., Mädler, S. C., et al. (2023). "SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes". In: *bioRxiv*, p. 2023.06.01.542416. DOI: 10.1101/2023.06.01.542416

SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes

Niklas A. Schmacke[1]†, Sophia C. Mädler[2]†, Georg Wallmann[1,2], Andreas Metousis[2], Marleen Bérouti[1], Hartmann Harz[3], Heinrich Leonhardt[3], Matthias Mann[2]*‡, Veit Hornung[1]*‡

[1]Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, 81377 Munich, Germany.

[2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany.

[3]Faculty of Biology, Human Biology and BioImaging, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany.

†, ‡ These authors contributed equally to this work

*Corresponding authors. Email: mmann@biochem.mpg.de, hornung@genzentrum.lmu.de

Abstract

Forward genetic screening associates phenotypes with genotypes by randomly inducing mutations and then identifying those that result in phenotypic changes of interest. Here we present spatially resolved CRISPR screening (SPARCS), a platform for microscopy-based genetic screening for spatial cellular phenotypes. SPARCS uses automated high-speed laser microdissection to physically isolate phenotypic variants *in situ* from virtually unlimited library sizes. We demonstrate the potential of SPARCS in a genome-wide CRISPR-KO screen on autophagosome formation in 40 million cells. Coupled to deep learning image analysis, SPARCS recovered almost all known macroautophagy genes in a single experiment and discovered a role for the ER-resident protein EI24 in autophagosome biogenesis. Harnessing the full power of advanced imaging technologies, SPARCS enables genome-wide forward genetic screening for diverse spatial phenotypes *in situ*.

Introduction

Genetic screens offer a powerful approach to dissecting the complexity inherent in biological systems. Within this space, forward genetic screening is an unbiased way to map phenotypic changes to changes in the genome: From a library of genetic variants generated by random mutagenesis, mutants with interesting phenotypes are selected and their genotypes determined. This approach has led to groundbreaking discoveries in a variety of model organisms (*1-3*). Now, with the ability to specifically target mutagenesis to exonic regions of interest and disrupt both alleles of a given genetic locus, CRISPR-based genome editing technologies (*4*) have enabled the generation of large mutant libraries in which a single gene is knocked out in each cell (*5*). Individual genetically perturbed cells can now be profiled for their transcriptome (*6-10*), protein expression (*11*), spatial composition (*12*) and chromatin landscape (*13*). However, genome-wide screening libraries typically contain tens of millions of cells, a scale with which most of these techniques are currently incompatible. To overcome this limitation, only those cells with an interesting phenotype are typically isolated from the library and subsequently genotyped. This paradigm has largely limited cell-based genome-wide screens to three types of easily selectable phenotypes: a difference in proliferation rate, an inhibition of cell death, or a change in fluorescence intensity compatible with fluorescence-activated cell sorting (FACS) (*14-17*).

Increasingly powerful microscopic imaging provides information-rich data on diverse cellular phenotypes (*18*) and would therefore be an ideal technology to read out biological phenotypes of interest, particularly if combined with recent advances in deep learning. However, its application in genome-wide forward genetic screening has been hampered by a lack of scalability and other limitations: 'in situ sequencing by synthesis', a technology originally developed to profile the cellular transcriptome in tissue samples, has been adapted to sequencing short perturbation-encoding barcodes on the DNA level (*19, 20*). This method separates genotyping and image collection, resulting in complete image datasets for unbiased identification of new phenotypes. However, by design it does not include an enrichment step for selected phenotypes, requiring all cells in a mutant library to be sequenced irrespectively of whether they show a phenotype. In addition, the genotype can only be determined for a fraction of cells due to low sequencing fidelity even in low-complexity libraries (*20*), which in combination with the technology's high costs has limited its applicability for screening genome-wide libraries at sufficient coverage (*21*). Image-based flow cytometers with sorting capabilities have recently enabled the investigation of spatial phenotypes at high throughput (*22, 23*). These devices rely on low-resolution flow-based microscopy of detached cells, preventing the identification of complex phenotypes. In addition, this technology makes sorting decisions in real time, restricting it to the identification of predefined phenotypes and preventing reanalysis of past screens. A method originally proposed for the transcriptomic characterization of B-cell populations (*24*) photoactivates fluorophores to mark cells for subsequent isolation by FACS (*25-27*). This approach can only separate few different phenotypes by fluorophore brightness (*28*). It also requires a real time

decision on which cells to isolate, preventing whole-dataset analysis to discover unexpected new phenotypes and has not been demonstrated to be compatible with cell fixation, which is necessary for antibody-based staining of intracellular targets.

To enable robust genome-wide high-throughput screening for spatial cellular phenotypes, we set out to develop a technology that meets four key requirements: First, it should work on cells *in situ* and utilize state-of-the-art microscopy techniques. Second, it should accommodate large screening libraries to ensure adequate representation of rare phenotypes. Third, it should be compatible with the unbiased identification of previously unknown phenotypes from entire complex image datasets rather than single images in real time. Fourth, it should allow for reanalysis and reselection of cells for genotyping from previous archived screens. Importantly, the latter feature would allow the application of novel image analysis methods to previously performed screens as they become available.

Results

Spatial genotyping by laser microdissection

To analyze the spatial composition of tissues and clinical samples by mass spectrometry, we have been advancing workflows based on laser microdissection (LMD), a technique that uses a focused UV laser to cut out and collect arbitrary shapes from tissue sections (*29, 30*). In a most recent development, deep visual proteomics (DVP), we use LMD to excise defined tissue regions for subsequent proteomic characterization of individual cell types or extracellular zones by mass spectrometry (*31, 32*). We reasoned that the isolation of single phenotypically interesting cells from a pooled library by LMD would provide an ideal basis for a forward genetic screening technology for spatial phenotypes.

LMD requires samples to be present on a membrane that can be cut by a UV laser, so we first tested whether cells could be grown and imaged directly on such polymer membranes. Indeed, on polyphenylene sulfate (PPS) membranes, spinning disk confocal microscopy produced high-quality images that showed normal cellular morphology (fig. 1A). By segmenting these images into individual cells, we generated multi-channel perturbation image datasets from which we aimed to identify cells with phenotypes of interest for genotyping (fig. 1B). We then developed a rapid cutting protocol for LMD that is compatible with subsequent genotyping by minimizing autofocus time and optimizing the trade-off between laser speed and accuracy. Compressing the cutting path and leveraging the fact that it is sufficient to isolate nuclei to determine a cell's CRISPR perturbation by sequencing, we ultimately reached a speed of 1,000 nuclei per hour.

Counting of excised membrane regions collected in a microwell plate using this protocol showed a yield of approximately 80 % (fig. 1C). We then tested the genotyping of excised nuclei by generating a pool of U2OS cells each expressing one of 77,441 unique sgRNAs in the Brunello CRISPR library (*33*),

plated these cells onto PPS membranes and imaged them. To register membrane slides between imaging microscopes and the LMD microscope, we marked the membrane with calibration crosses as landmarks that define a coordinate system across each slide, allowing us to find the positions of cells to excise. We stitched individual field of view images of a slide into one whole slide image (WSI), segmented nuclei based on a DNA stain, generated a cutting map using our newly developed open-source python library py-lmd (fig. S1) and then excised and lysed 1,000 nuclei. Sequencing identified 549 unique sgRNAs on average in these lysates, demonstrating that isolating individual nuclei for subsequent CRISPR genotyping is feasible with LMD (fig. 1D). Comparing the number of unique sgRNAs in the LMD lysate with a lysate of cells from the same library isolated by FACS revealed that both techniques recovered an sgRNA from approximately 50 % of cells (fig. 1D). From these data we concluded that potential DNA damage induced by laser microdissection does not hamper sgRNA recovery. In summary, our results show that it is possible to employ LMD to recover genetic information from imaged cells at a throughput compatible with genetic screening (fig. 1E). We call this approach spatially resolved CRISPR screening (SPARCS).

Validating SPARCS for genetic screening

To further develop SPARCS we applied it to screen for regulators of starvation-induced macroautophagy (hereafter referred to as autophagy), a fundamental process for cellular energy management (*34, 35*). The signature of autophagy is the formation of vesicles called autophagosomes. These are covalently decorated with proteins from the ATG8 family, including the well-studied human protein LC3B. During a key event in autophagosome biogenesis LC3B is conjugated to the head group of the lipid phosphatidylethanolamine (PE) through a series of ubiquitin ligation-like reactions. A critical component of this cascade is the protein ATG5 that forms an E3-like complex with ATG12 to mediate the covalent attachment of LC3B to PE. To follow the formation of autophagosomes during starvation we stably expressed LC3B tagged with mCherry in U2OS cells, because – unlike GFP – mCherry remains fluorescent upon fusion with the lysosome. We then treated these cells with the mTOR inhibitor Torin-1 to mimic starvation, which induces autophagy. Cells treated this way began to accumulate mCherry-LC3-positive puncta over the course of 14 hours (fig. 2A).

In a screen, those cells containing sgRNAs against essential regulators of autophagy are unable to form these puncta. To identify these cells we trained a deep learning-based image classifier to differentiate between cells with or without autophagosomes (fig. 2B, fig. S2A). The training dataset was composed of segmented single cell images of mCherry-LC3 expressing U2OS cells that were treated with Torin-1 (autophagy-on class) or left untreated (autophagy-off class). As an additional group we introduced cells treated with Torin-1, yet deficient in ATG5 (autophagy-off class). We used images from several biological replicates to avoid overfitting of our classifier to batch-specific characteristics such as staining intensity or cell density (table S1). To evaluate the performance of this classifier 1.0, we

generated a new test dataset of images from unstimulated and Torin-1 stimulated wildtype and *ATG5*$^{-/-}$ mCherry-LC3-expressing U2OS cells that had not been part of the training set and as such had never been seen by the classifier before. Classifier 1.0 achieved a false discovery rate (FDR) of < 1% (fig. S2B) at the chosen threshold, meaning that less than 1% of cells classified as potential hits with an autophagy-off phenotype were instead false positives that actually came from the autophagy-on class.

We then validated SPARCS by performing a small pilot screen on autophagosome formation in 1.2 million Torin-1-stimulated mCherry-LC3 U2OS cells transduced with the Brunello CRISPR knockout (KO) library (fig. 2C). From this library we isolated the top 0.1 % of cells classified as autophagy-off by classifier 1.0 with a score > 0.94, corresponding to a test set FDR of 0.38 %. Compared to the entire library, we found sgRNAs targeting *ATG5* to be highly enriched among isolated cells (median 200-fold) (fig. 2D). sgRNAs targeting other autophagy-related genes had a median of 60-fold enrichment with the most strongly enriched sgRNAs even exceeding 700-fold (fig. 2D). Control sgRNAs not targeting any human genes ('non targeting controls' (NTCs)) were rare among isolated cells with a median enrichment of 10-fold (fig. 2D). These results confirm that the SPARCS protocol stitches and registers WSI with sufficient accuracy for the isolation of the nuclei of interest. They also demonstrate that assessing autophagosome formation based on images is feasible with a deep learning classifier, and that in SPARCS, this classifier can be used to screen for autophagosome formation.

Accurate detection of autophagy defects in single cell images

A classifiers' Receiver Operating Characteristic (ROC) curve visualizes the tradeoff between true positive rate (the fraction of all autophagy-off cells that are correctly identified) and false positive rate (the fraction of autophagy-on cells incorrectly predicted as autophagy-off). The ROC curve of our classifier 1.0 confirmed its overall accuracy with an area under the curve (AUC) of > 0.92 (fig. S2C). However, at the precision (the fraction of predicted autophagy-off cells that are actually autophagy-off, 1-FDR) corresponding to 1 % FDR, the recall (= true positive rate) of classifier 1.0 was below 26% (fig. S2D). Closer analysis of the different categories of cells in the test dataset revealed that the classifier excelled at identifying autophagy-on cells, but performed poorly at recognizing autophagy-off cells (fig. S2E, F).

To improve classification of autophagy-off cells we refined our staining and imaging protocol and then trained a new version of our classifier. For this version 2.0 we decided to use a more streamlined multilayer perceptron (MLP) head with fewer trainable parameters, add another linear layer and increase the number of cells and biological replicates in the training dataset to capture as much biological variance as possible (fig. 3A, B, table S1). We also prefiltered the unstimulated and Torin-1 stimulated images for autophagy-off and -on cells to minimize the number of mislabeled training examples (table S1). To evaluate classifier 2.0, we first used parametric UMAP (*36*) to investigate if layers of the CNN had learned to differentiate between images of autophagy-on and autophagy-off cells.

This revealed that wildtype cells stimulated with Torin-1, unstimulated wildtype cells and *ATG5*$^{-/-}$ cells clustered separately in representations of lower layers, most prominently in the 8th of 9 convolutional layers (fig. 3C). These results suggested that our CNN had now learned to featurize images of LC3 distribution in a way that enables accurate classification of cells undergoing autophagy. Of note, the network of classifier 2.0 was capable of discriminating between *ATG5*$^{-/-}$ and unstimulated wildtype cells despite those cells being in the same training class (fig. 3C), a clear improvement over classifier 1.0 (fig. S2G). Its ROC curve was also drastically improved with an AUC of > 0.999 (fig. 3D). Remarkably, in the binary classification output almost all cells were correctly classified according to their autophagy status even with a simple classification score threshold of 0.5 (fig. 3E, F). With classifier 2.0, classification thresholds > 0.98 produced FDRs of < 1 %, with higher thresholds reducing the FDR further without yet diminishing the excellent recall of nearly 100 % (fig. 3G, fig. S2H). Thus, for a complex biological process such as autophagy, training a CNN-based classifier on images from comparatively few biological replicates achieves excellent performance.

Genome-wide autophagy screen with SPARCS

Encouraged by these results we used SPARCS to conduct a genome-wide screen on autophagosome formation. We screened a library of 40 million mCherry-LC3 expressing U2OS cells at a median coverage of 1,818 cells per gene in the human genome in batches of 5 and 35 million cells (fig. S3). Classifying autophagy based on the distribution of LC3 within the first batch showed that 0.56% of cells had a score > 0.98. We regarded these cells as potential autophagy-defective hits and, upon examining the 8th CNN layer featurization of their LC3 distribution using parametric UMAP, found them to cluster separately from autophagy-on cells in the library with a classification score < 0.02 (fig. S4A, B). For genotyping we divided the hits into six bins according to their classification score (fig. S4C): The top bin represented a cutoff at which we found *ATG5*$^{-/-}$ to be strongly enriched in our test dataset, whereas the second bin corresponded to unstimulated wildtype cells. Bins 3 – 6 contained the remaining potential hits with a roughly equal number of cells per bin. Zooming in on the 8th CNN layer featurization of the LC3 distribution in the potential hits revealed that cells in bins 1 & 2 clustered separately from bins 3 - 6 (fig. S4D). This indicated that they contained different phenotypic variants with regard to their LC3 distribution, potentially corresponding to stronger defects in autophagosome formation. Indeed, we observed the fewest LC3 puncta in cells from bins 1 & 2 (fig. S4E).

For the second genome-wide screen batch we refined our classifier by more stringently selecting training examples of autophagy-on and -off cells, thereby further improving its overall performance (fig. S5). Using this new classifier 2.1 we obtained similar results from the second batch compared to 2.0 on the first screen batch: 1.40% of cells were classified as autophagy-off with a score above 0.98 and in their 8th CNN layer featurization these cells again clustered separately from their autophagy-on counterparts with a score < 0.02 (fig. 4A, B). We therefore applied the same binning strategy to these

images (fig. 4C), and, upon zooming in on their 8th CNN layer featurization using parametric UMAP, found the separation of cells in bin 1 & 2 to be even more apparent than in the first batch (fig. 4D, E). We then isolated a total of 395,173 nuclei across both screen batches and sequenced their sgRNAs. Given their similarity on the phenotypic level we analyzed the genetic data of both batches together. All bins showed a marked enrichment of targeting over non-targeting sgRNAs and enrichment scores up to 600-fold, promising the identification of autophagy relevant genes (fig. 4F). In line with our FDR calculation (fig. 3G, fig. S5B) and our conclusions from the featurization of individual images (fig. 4D, fig. S4D), sgRNAs targeting genes known to be involved in autophagosome formation were most strongly enriched in bins 1 & 2 (fig. 4F). On the gene level *ATG5*, which our classifier was trained to identify, was among the most highly enriched genes in several bins, validating our supervised classification approach in the context of this large-scale screen (fig. 4G). The Brunello library targets each gene in the human genome with four sgRNAs. While in the higher score bins 1 & 2, genes with a high mean enrichment score had several sgRNAs enriched, in lower bins genes with a relatively high mean enrichment score often only had a single highly enriched sgRNA, indicating potential off-target effects. This prompted us to evaluate the number of sgRNAs enriched per gene as an alternative metric to score screening hits. Here we again found the strongest hits to contain mainly autophagy-related genes (fig. 4H). Taken together, these results establish that SPARCS is highly effective for large scale genetic screens on spatial phenotypes. Furthermore, despite the inherent complexity of image-based phenotypes, our supervised classifier facilitated the enrichment of a very small subset of individual cells with a genetically defined phenotype from a diverse genome-wide library of 40 million cells.

EI24 reorganizes membranes for autophagy

The power of SPARCS became even more apparent when we evaluated our screen from the perspective of the investigated biological process: Remarkably, this single screen recovered almost all known essential genes of the starvation-induced macroautophagy pathway. This included the complete ULK1 complex and LC3 lipidation cascade (fig. 5A). Closer inspection of individual hits revealed that the most strongly enriched gene that is not a canonical macroautophagy gene was *EI24*, a gene coding for an ER-resident transmembrane protein (*37*) (fig. 5B). *EI24*$^{-/-}$ cells have previously been described as autophagy-defective, but with a phenotype resulting in spontaneous LC3-puncta formation (*38*). This finding is not in line with the 82-fold enrichment of *EI24* in our screen, given that our classifier was trained to recognize cells with impaired rather than increased autophagosome formation. To investigate why we found EI24 KOs enriched among cells classified as autophagy-off, we generated individual *EI24*$^{-/-}$ clones. Consistent with the previously reported spontaneous LC3 puncta formation, EI24-deficient cells have been described to exhibit increased lipidation of LC3 under steady state conditions (*38*), a phenotype we confirmed in *EI24*$^{-/-}$ clones (fig. 5C). However, in contrast to previous results we found LC3 puncta formation in response to Torin-1 to be largely abolished in EI24-deficient cells (fig. 5D). Instead, these cells formed a single mCherry-LC3-positive speck that became more pronounced

with Torin-1 stimulation, indicating a general defect in membrane traffic or autophagosome formation (fig. 5E). These results explain why our classifier picked up *EI24* knockouts and demonstrate again that even supervised image classification is capable of identifying previously undescribed phenotypes. Our results further indicate that *EI24* is required for autophagosome formation and has a function beyond its recently described LC3 puncta promoting role in maintaining Ca^{2+} homeostasis across the ER membrane (*39*) that remains to be investigated.

Discussion

We present SPARCS, a platform that enables unbiased exploration of the genetic basis of subcellular spatial features in forward genetic screens. At the core of the SPARCS methodology, we have adapted and refined laser microdissection technology to unprecedented throughput to facilitate genetic screening applications. We have improved the precision and efficiency of isolating single nuclei from cell cultures, while automating the extraction of several hundred thousand nuclei into distinct bins. By integrating a deep learning-based classifier, our genome-wide SPARCS screen successfully identified nearly all known genes related to macroautophagy and revealed a novel phenotype associated with the *EI24* gene.

SPARCS offers a unique combination of features (table S2) that make it a powerful forward genetic screening platform. It can be seamlessly integrated with any state-of-the-art microscope for *in situ* cell imaging. The screening library size is not constrained, except by the imaging microscope's throughput. Consequently, microscopy-based genome-wide perturbation screens can now achieve exceptional coverage. Besides the method described here, which involves isolating cells based on predefined classes, SPARCS is also compatible with the identification of individual cells exhibiting entirely novel or unexpected phenotypes. This is achieved through unbiased clustering and anomaly detection applied to the entire image dataset. Furthermore, we discovered that samples can be stored long-term, allowing for the reanalysis of archived SPARCS screens using newer algorithms. This facilitates the exploration of new biological insights within existing data. To streamline the process of translating the identification of individual cells with subcellular spatial phenotypes into a cutting map for LMD we have developed py-lmd, an open-source Python library for laser microdissection on arbitrary sample types that is available on GitHub. We hope that the accessible design of SPARCS, compatible with standard microscopes and sequencing workflows, will encourage its adoption by the scientific community.

Our screen uncovered a potential role in macroautophagy for *EI24*. This gene had previously been implicated in autophagy based on a *C. elegans* screen in 2010, but its mechanism of action had remained unclear (*38*). Beyond the original observation that EI24 deficiency leads to pronounced formation of non-functional autophagosomes even under steady state conditions, it was recently suggested that spontaneous Ca^{2+} fluxes across the ER membrane initiated autophagy in EI24 deficient cells (*39*). How

spontaneous induction of autophagosome formation in *EI24*$^{-/-}$ cells can be reconciled with a defect in autophagy remained unclear. The results from our screen and the following live cell imaging experiments, in which we found *EI24*$^{-/-}$ cells to form fewer autophagosomes than wildtype cells, now suggest that EI24 plays a – potentially additional – role in autophagosome formation.

Systems biology is increasingly driven by large-scale artificial intelligence models that set new standards for reconstructing and predicting cellular behavior, but require enormous amounts of data to train. In light of this development, comprehensive, unbiased data acquisition approaches that can generate large datasets across modalities have become highly desirable. In this context, SPARCS, with its focus on open and accessible design and the ability to screen large libraries, can make a valuable contribution to understanding biology from the molecular to the organismic scale.

Acknowledgements

We would like to thank Larissa Hansbauer for outstanding technical support; Jochen Rech and the BioSysM Liquid Handling Unit for excellent support with robotics; Claudia Ludwig and the BioSysM FACS Core Facility for great support with cell sorting; Mario Oroshi and the computing centre of the Max Planck Institute of Biochemistry for computational support and IT infrastructure; the Imaging Facility of the MPI of Biochemistry and the Center for Advanced Light Microscopy (CALM) for support with light microscopy; Rin Ho Kim and the Sequencing Facility of the MPI of Biochemistry as well as Stefan Krebs and the Genomics unit of the Laboratory for Functional Genome Analysis (LAFUGA) for sequencing; and Falk Schlaudraff, Christoph Greb and Florian Hoffmann from Leica Microsystems for technical support.

Funding

S.C.M. is a PhD fellow of the Boehringer-Ingelheim Fonds. This study was supported by the Max-Planck Society for Advancement of Science. This project was funded by European Research Council grant ERC-2020-ADG ENGINES (101018672 to V.H.).

Author Contributions

Conceptualization: N.A.S., S.C.M.; Formal Analysis: N.A.S, S.C.M., G.W.; Funding Acquisition: M.M., V.H.; Investigation: N.A.S., S.C.M., G.W., A.M., M.B., H.H.; Resources: H.H., H.L., M.M., V.H.; Software: N.A.S., S.C.M., G.W.; Visualization: N.A.S., S.C.M., G.W.; Writing – original draft: N.A.S., S.C.M., M.M., V.H.; Writing – review & editing: N.A.S., S.C.M., G.W., A.M., M.B., H.H., H.L., M.M., V.H.

Competing Interests

The authors declare no competing interests.

Data and materials availability

Code to recreate the figure manuscripts is available on GitHub (https://github.com/MannLabs/SPARCS_pub_figures).

The code described in this manuscript is available from the following GitHub repositories:

The py-lmd python library provides code to direct excision of defined regions on a Leica LMD7 laser microdissection microscope (https://github.com/MannLabs/py-lmd).

The SPARCStools python library provides code to rename TIF image files generated by the PerkinElmer Harmony software and stitch these into WSIs (https://github.com/MannLabs/SPARCStools).

The SPARCSpy python library contains the autophagy classifiers, as well as code to segment and extract single-cell images from entire fields of view up to WSIs (https://github.com/MannLabs/SPARCSpy).

All other data are available from the authors upon request.

Methods

Cell culture

U2OS cells were cultured in DMEM supplemented with 10 % fetal calf serum (FCS), penicillin/streptomycin and 1 mM sodium pyruvate and split every 2-3 days. PPS membrane slides were sterilized for 30 minutes under the UV light of a cell culture hood with their cavity side down. Cells were then plated onto these slides cavity down in 4-well plates with 5 mL DMEM per well.

Genome engineering

U2OS cells stably expressing Cas9, mCherry-LC3 and mNeon tagged N-terminally with the lipidation sequence of Lck (LckLip-mNeon, the original plasmid was a gift from Dorus Gadella (Addgene plasmid # 98821, (*40*))) were generated via lentiviral transduction. Briefly, HEK-293T cells were transfected with transfer plasmids for Cas9 or mCherry-LC3 and 3rd generation lentiviral particle production plasmids pMDLg and pRSV as wells as a VSV G-protein pseudotyping plasmid 18 hrs after plating. Eight hrs later, the medium was exchanged and cells were washed once in PBS. After 48 hrs supernatants containing viral particles were harvested and transferred onto U2OS cells plated 18 hrs before. 48 hrs later U2OS cells were washed. Cells were selected for Blasticidin resistance with 10 µg/mL Blasticidin or FACS-sorted for high fluorescent protein expression and single clones generated by limiting dilution cloning. A bright clone with a visible reaction to 600 nM Torin-1 was selected, expanded and used for all experiments. Lentiviral particles for the expression of individual sgRNAs from LentiGUIDE-Puro were generated analogously. Cells were selected for sgRNA expression with 5 µg/mL puromycin for 48 hrs. Of note, the cell line used for the autophagy screens did not yet stably express Cas9 but was instead transduced with a LentiCRISPRv2, a vector driving expression of both Cas9 and an sgRNA.

Laser microdissection

Cutting paths for laser microdissection of selected cells were generated using our open-source python library py-lmd (https://github.com/MannLabs/py-lmd) with the configurations specified in the "screen config" file. Each shape was dilated to ensure that the cutting line did not go through or damage the nucleus. Laser microdissection was carried out on a Leica LMD7 at 40 x magnification using the software version 8.3.0.8275. The microscope was equipped with the Okolab LMD climate chamber (H201-ENCLOSURE-LMD and H201-LEICA-LMD) to ensure stable temperatures throughout the cutting process. Slides were equilibrated in the microscope to 34.5 °C before cutting to ensure focus stability. Cutting contours were imported from the XML files generated with py-lmd after reference point alignment and cut with the following settings: power 60, aperture 1, speed 100, head current 46 % - 51 %, pulse frequency 1128 and offset 185. Autofocus adjustment was performed every 30 shapes. Shapes were sorted into 48-well plates. During cutting a custom-built wind protection was used around

the collection plate to ensure collection of excised shapes into the center of the well and prevent wind disturbances. After cutting, samples were stored at 4 °C before lysis and library generation.

CNN-based image classifier training

Neural networks with 9 randomly initialized convolutional layers and 3 (classifiers 1, P) or 4 (classifiers 2.1 & 2.2) linear layers were trained to classify segmented single cell images as autophagy-on or autophagy-off (table S1). The training datasets were based on several biological replicates of mCherry-LC3 expressing U2OS cells with and without autophagosomes. The autophagy-off class consisted of images from unstimulated wildtype cells (pre-filtered to remove cells showing spontaneous autophagosome formation for 2.1 and 2.2) and two different *ATG5*$^{-/-}$ clones. Where applicable, pre-filtering was performed with classifier P. The autophagy-on class consisted of single-cell images from stimulated wildtype cells, where applicable pre-filtered with classifier P to remove non-responding cells. To increase variability captured in the training data, the training slides were plated at an angle to include varying cell densities on one slide. 500 k, 1 million or 1.2 million single-cell images respectively were randomly selected from each class for training while ensuring balanced sampling from each test slide. An additional 50 k cells from each class were used for testing and validation during training. Training data were augmented by Gaussian blur, addition of Gaussian noise and random rotations in 90° steps. Training was performed using single-gradient descent with a learning rate of 1×10^{-3}. Gradient clipping was set to 0.5. Training was performed over a total of 20, 30 or 40 epochs. Classifier performance was tested on a biologically independent set of unstimulated wildtype cells, Torin-1 stimulated wildtype cells and *ATG5*$^{-/-}$ cells. Models were built and trained using PyTorch (*41*).

Segmentation of individual cells

Images were flat-field corrected during image acquisition using the Perkin Elmer Harmony software (v4.9) and intensity rescaled to the 1 % and 99 % quantile. Extremely bright regions (pixel values greater than 40000) were set to 0 before determining the normalization quantiles. Stitching of image tiles was performed using the ashlar python API (*42*) in our open-source python library SPARCStools (https://github.com/MannLabs/SPARCStools).

Stitched whole slide images were segmented using our open-source SPARCSpy python library (https://github.com/MannLabs/SPARCSpy) with the parameters defined in “config_screen” or “config_training” respectively. A nucleus segmentation mask was generated using a local median thresholding approach and the cytosol segmentation mask was calculated using fast marching from nuclear centroids with WGA staining as a potential map.

Single cell images were extracted based on nuclear and cytosolic segmentation masks. The masked area was extended using a Gaussian filter with a sigma of 5 to extract information from each of the imaged channels and saved to hdf5 files as individual 128 × 128 px images.

Sample preparation and imaging of autophagy

After stimulation with 600nM Torin-1, PPS slides were washed 1x in PBS in a Coplin jar and then stained with 10 µg/mL WGA-Alexa488 in PBS for 10 minutes at 37 ºC. After washing 1 x with PBS slides were fixed for 10 minutes at room temperature in 4 % MeOH-free PFA in PBS in 4-well plates. After washing 3 x in PBS, slides were stained with 10 µg/mL Hoechst-33342 in PBS at 37 ºC for at least 30 minutes. After washing 3 x in PBS, slides were dried in a centrifuge at 3,400 g for 1 minute. Cells in ibidi microwell slides and plates were stained according to the same protocol but imaged in PBS. Imaging was done on a Nikon Eclipse Ti2 spinning disk confocal microscope or an Opera Phenix high-content imager as indicated.

Genetic screening for autophagy regulators

We conducted our screen in mCherry-LC3 expressing U2OS cells using the Brunello human CRISPR KO library in the LentiCRISPRv2 backbone. The Brunello library was a gift from David Root and John Doench (Addgene #73178) and amplified according to their protocol (*33*). U2OS cells were transduced with lentiviral particles produced as described above at an MOI of approximately 0.2. After 48 hrs, successfully transduced cells were selected with 5 µg/mL puromycin for two days and then expanded for three days. We then plated 50 million cells on a total of 109 slides in 4 well plates, and in addition included unstimulated and wildtype controls on separate slides with every screening batch for classifier training. The day after plating, cells were stimulated with 600 nM Torin-1 for 14 hrs. Slides were then prepared for microscopy as described above and imaged on an Opera Phenix high content imager at 20 x resolution. Where applicable slides were stored at -20 °C and brought to 4 °C the day before laser microdissection. Cells from each bin were excised into multiple wells. Nuclei were then lysed in 48-well plates using the arcturis PicoPure™ DNA extraction kit. 120 µL of lysis buffer was added to each well and incubated at 65 °C for 4 hrs. Proteinase K was inactivated at 95 °C for 15 mins. Cooled samples were transferred to PCR tubes and the emptied wells were rinsed with 40 µL of ddH2O. Amplification of sgRNAs was performed as described previously (*43*) but in a single step PCR (*33*) over 36 cycles with no added water. Sequencing was performed on an Illumina NextSeq with 500 reads per nucleus on average. An sgRNA read count table was generated for each sequencing library. Low quality sgRNAs were removed by applying a minimum number of reads per sgRNA threshold that was set based on the distribution of read counts per sgRNA in the sample. The non-targeting sgRNA with the sequence TACGTCATTAAGAGTTCAAC was excluded from sequencing results of approximately 40 % of cells from bins 3 - 6 of batch 2 of the genome-wide screen due to a contamination of the sequencing library leading to abnormally high read counts of this specific sequence. For further analysis only sgRNAs with at least a fraction of reads corresponding to a single cell per well were used. sgRNA read fractions of individual wells were then aggregated per bin by multiplying with the fraction of excised cells in that well over all excised cells in the bin. The per-bin aggregated sequencing results

were used for all further data analysis. Enrichment values were determined by normalizing the aggregated fraction of reads per sgRNA to the fraction of reads per sgRNA in the input cell library.

Immunoblotting

20,000 U2OS cells were plated per 96-well. 18 hrs after plating, cells were stimulated and then harvested in 1 x Lämmli buffer. 3 wells were pooled per condition. Lysates were boiled at 95 ºC for 5 min. and then run on 16 % TRIS-glycine polyacrylamide gels before immunoblotting onto 0.2 µm nitrocellulose membrane for 90 minutes. Membranes were blocked in 5 % milk in PBST for 1 hr and incubated with primary antibody at 4 ºC overnight. After washing 3 x in PBST for a total of 15 min. membranes were incubated with HRP-labelled secondary antibody for 2 hrs at room temperature. After washing 3 x in PBST for a total of 15 min. membranes were covered in luminescent HRP substrate and immediately imaged.

References

1. S. Brenner, The genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94 (1974).
2. C. Nusslein-Volhard, E. Wieschaus, Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795-801 (1980).
3. N. Kayagaki *et al.*, Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling. *Nature* **526**, 666-671 (2015).
4. J. Y. Wang, J. A. Doudna, CRISPR technology: A decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023).
5. O. Shalem, N. E. Sanjana, F. Zhang, High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16**, 299-311 (2015).
6. B. Adamson *et al.*, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821 (2016).
7. D. A. Jaitin *et al.*, Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896 e1815 (2016).
8. A. Dixit *et al.*, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866 e1817 (2016).
9. P. Datlinger *et al.*, Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* **14**, 297-301 (2017).
10. J. M. Replogle *et al.*, Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559-2575 e2528 (2022).
11. A. Wroblewska *et al.*, Protein Barcodes Enable High-Dimensional Single-Cell CRISPR Screens. *Cell* **175**, 1141-1155 e1116 (2018).
12. M. Lawson, J. Elf, Imaging-based screens of pool-synthesized cell libraries. *Nat Methods* **18**, 358-365 (2021).
13. A. J. Rubin *et al.*, Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361-376 e317 (2019).
14. O. Parnas *et al.*, A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686 (2015).
15. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).
16. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).
17. M. M. Gaidt *et al.*, The DNA Inflammasome in Human Myeloid Cells Is Initiated by a STING-Cell Death Program Upstream of NLRP3. *Cell* **171**, 1110-1124.e1118 (2017).
18. M. Boutros, F. Heigwer, C. Laufer, Microscopy-Based High-Content Screening. *Cell* **163**, 1314-1325 (2015).
19. D. Feldman *et al.*, Optical Pooled Screens in Human Cells. *Cell* **179**, 787-799 e717 (2019).
20. L. Funk *et al.*, The phenotypic landscape of essential human genes. *Cell* **185**, 4634-4653 e4622 (2022).
21. R. J. Carlson, M. D. Leiken, A. Guna, N. Hacohen, P. C. Blainey, A genome-wide optical pooled screen reveals regulators of cellular antiviral responses. *Proc Natl Acad Sci U S A* **120**, e2210623120 (2023).
22. N. Nitta *et al.*, Intelligent Image-Activated Cell Sorting. *Cell* **175**, 266-276 e213 (2018).

23. D. Schraivogel *et al.*, High-speed fluorescence image-enabled cell sorting. *Science* **375**, 315-320 (2022).

24. G. D. Victora *et al.*, Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* **143**, 592-605 (2010).

25. G. Kanfer *et al.*, Image-based pooled whole-genome CRISPRi screening for subcellular phenotypes. *J Cell Biol* **220**, (2021).

26. X. Yan *et al.*, High-content imaging-based pooled CRISPR screens in mammalian cells. *J Cell Biol* **220**, (2021).

27. J. Lee *et al.*, Versatile phenotype-activated cell sorting. *Sci Adv* **6**, (2020).

28. N. Hasle *et al.*, High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Mol Syst Biol* **16**, e9442 (2020).

29. L. F. Waanders *et al.*, Quantitative proteomic analysis of single pancreatic islets. *Proc Natl Acad Sci U S A* **106**, 18902-18907 (2009).

30. F. Coscia *et al.*, A streamlined mass spectrometry-based proteomics workflow for large-scale FFPE tissue analysis. *J Pathol* **251**, 100-112 (2020).

31. A. Mund *et al.*, Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* **40**, 1231-1240 (2022).

32. F. A. Rosenberger *et al.*, Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *bioRxiv*, 2022.2012.2003.518957 (2022).

33. J. G. Doench *et al.*, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature ...* **34**, 184-191 (2016).

34. I. Dikic, Z. Elazar, Mechanism and medical implications of mammalian autophagy. *Nat Rev Mol Cell Biol* **19**, 349-364 (2018).

35. H. Yamamoto, S. Zhang, N. Mizushima, Autophagy genes in biology and disease. *Nat Rev Genet*, 1-19 (2023).

36. T. Sainburg, L. McInnes, T. Q. Gentner, Parametric UMAP embeddings for representation and semi-supervised learning. 2020 (10.48550/arXiv.2009.12981).

37. P. J. Thul *et al.*, A subcellular map of the human proteome. *Science* **356**, (2017).

38. Y. Tian *et al.*, C. elegans screen identifies autophagy genes specific to multicellular organisms. *Cell* **141**, 1042-1055 (2010).

39. Q. Zheng *et al.*, Calcium transients on the ER surface trigger liquid-liquid phase separation of FIP200 to specify autophagosome initiation sites. *Cell* **185**, 4082-4098 e4022 (2022).

40. A. O. Chertkova *et al.*, Robust and Bright Genetically Encoded Fluorescent Markers for Highlighting Structures and Compartments in Mammalian Cells. *bioRxiv*, 160374 (2020).

41. A. Paszke *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019 (10.48550/arXiv.1912.01703).

42. J. L. Muhlich *et al.*, Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613-4621 (2022).

43. T. Schmidt, J. L. Schmid-Burgk, V. Hornung, Synthesis of an arrayed sgRNA library targeting the human genome. *Scientific Reports* **5**, 14987 (2015).

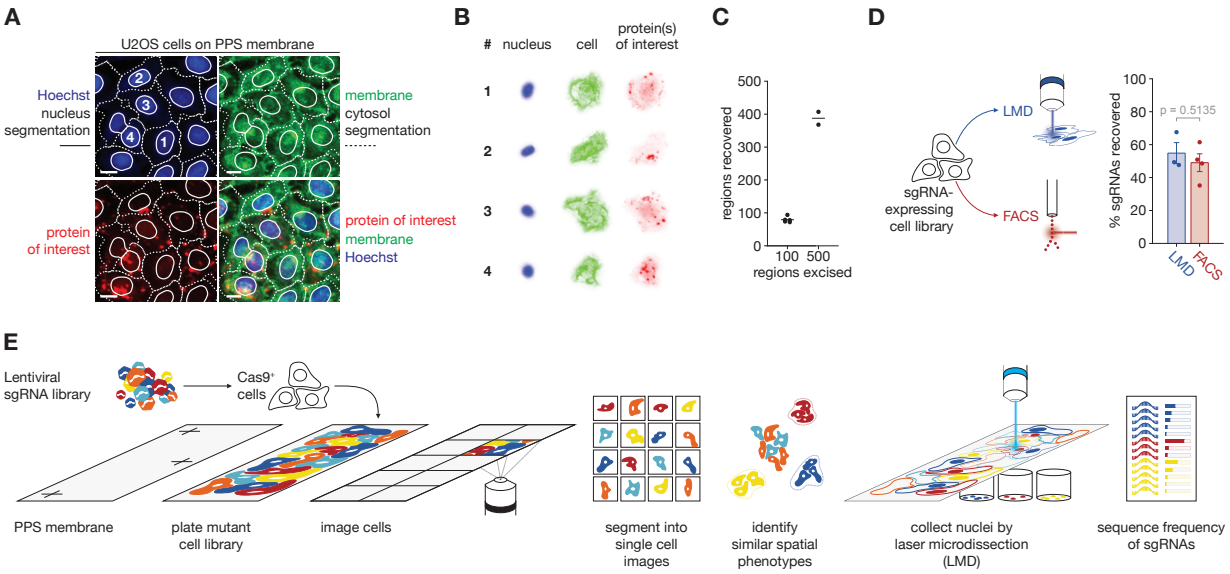


Figure 1

Figure 1 | SPARCS enables genome-wide CRISPR screening for spatial phenotypes in human cells

(A) Example images of U2OS cells on PPS membranes in several channels. Solid lines indicate nuclear segmentation based on Hoechst DNA staining; dotted lines indicate cytosol segmentation based on fast marching from nuclear centroids with wheat germ agglutinin (WGA)-Alexa 488 staining as a potential map. Numbers correspond to images of individual cells shown in (B). Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification. Scalebars represent 15 µm. PPS: polyphenylene sulfide.

(B) Post-segmentation images of individual mCherry-LC3 expressing U2OS cells. Numbers correspond to cells shown in (A).

(C) 100 or 500 regions were excised from U2OS cells grown on a PPS membrane slide and subsequently counted. Five and two technical replicates were excised from one slide, respectively.

(D) Comparison of sgRNA recovery after isolation of sgRNA-expressing fixed cells from one library either by laser microdissection (Leica LMD7, 1,000 nuclei per replicate, 3 independent biological replicates) or FACS (technical replicates). Bars indicate mean % sgRNAs recovered, error bars indicate SEM. p-value was calculated with an unpaired two-tailed t-test.

(E) Overview of genome-wide CRISPR screening for microscopy-based spatial phenotypes with the SPARCS pipeline. Laser microdissection of individual nuclei on a Leica LMD7 has been optimized to isolate 1,000 nuclei/hr. Instructions for laser microdissection of selected cells are generated using our open-source python library py-lmd. PPS: polyphenylene sulfide.

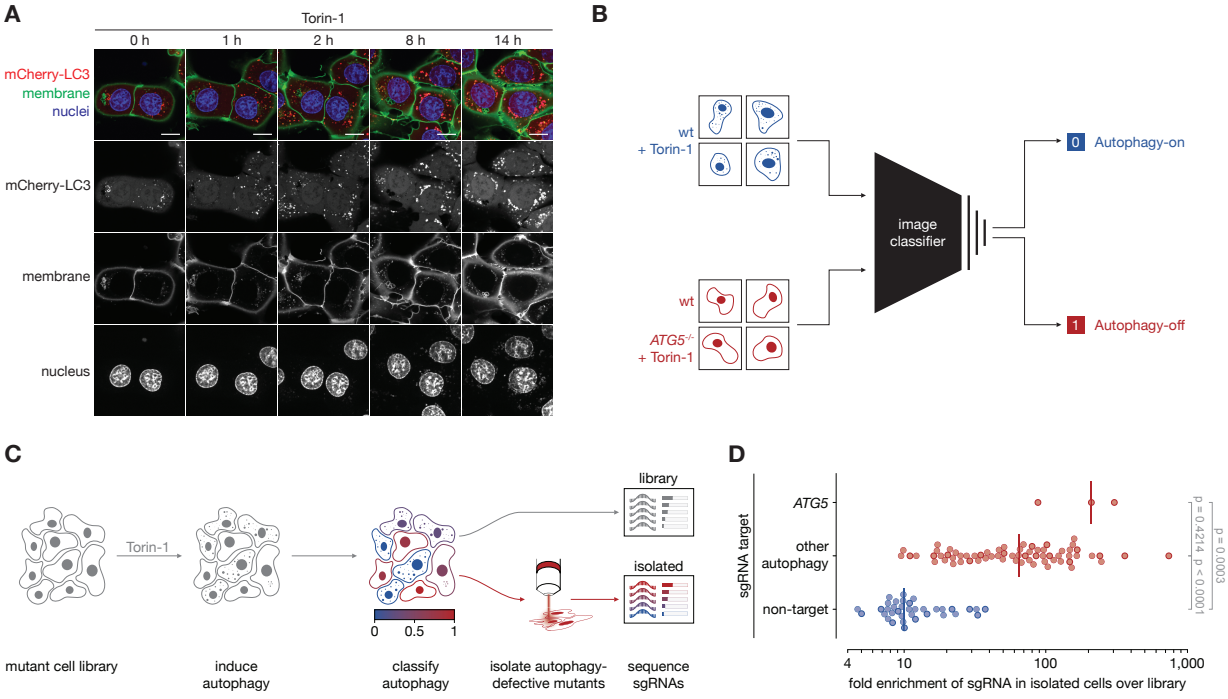


Figure 2

Figure 2 | SPARCS achieves strong enrichment of spatial phenotypes in a forward genetic screen

(A) U2OS cells expressing mCherry-LC3 and mNeon tagged with the lipidation signal of Lck at the N-terminus (membrane marker) were stimulated with Torin-1 and imaged live once per hour on a Nikon Eclipse Ti2 confocal microscope with 100 x magnification. Scalebars represent 15 µm. One representative of three independent experiments.

(B) Schematic describing the training of a convolutional neural network-based image classifier for the identification of individual autophagy-defective cells.

(C) Overview of SPARCS screening for autophagy.

(D) Results from a SPARCS screen for autophagosome formation on 1.2 million U2OS cells. The top 0.1 % of cells classified as autophagy-off with a score above 0.94 by classifier 1.0 were isolated by laser microdissection (LMD) and their sgRNAs sequenced to determine their enrichment relative to the input library. p-values were calculated with a Kruskal-Wallis test followed by Tukey's test.

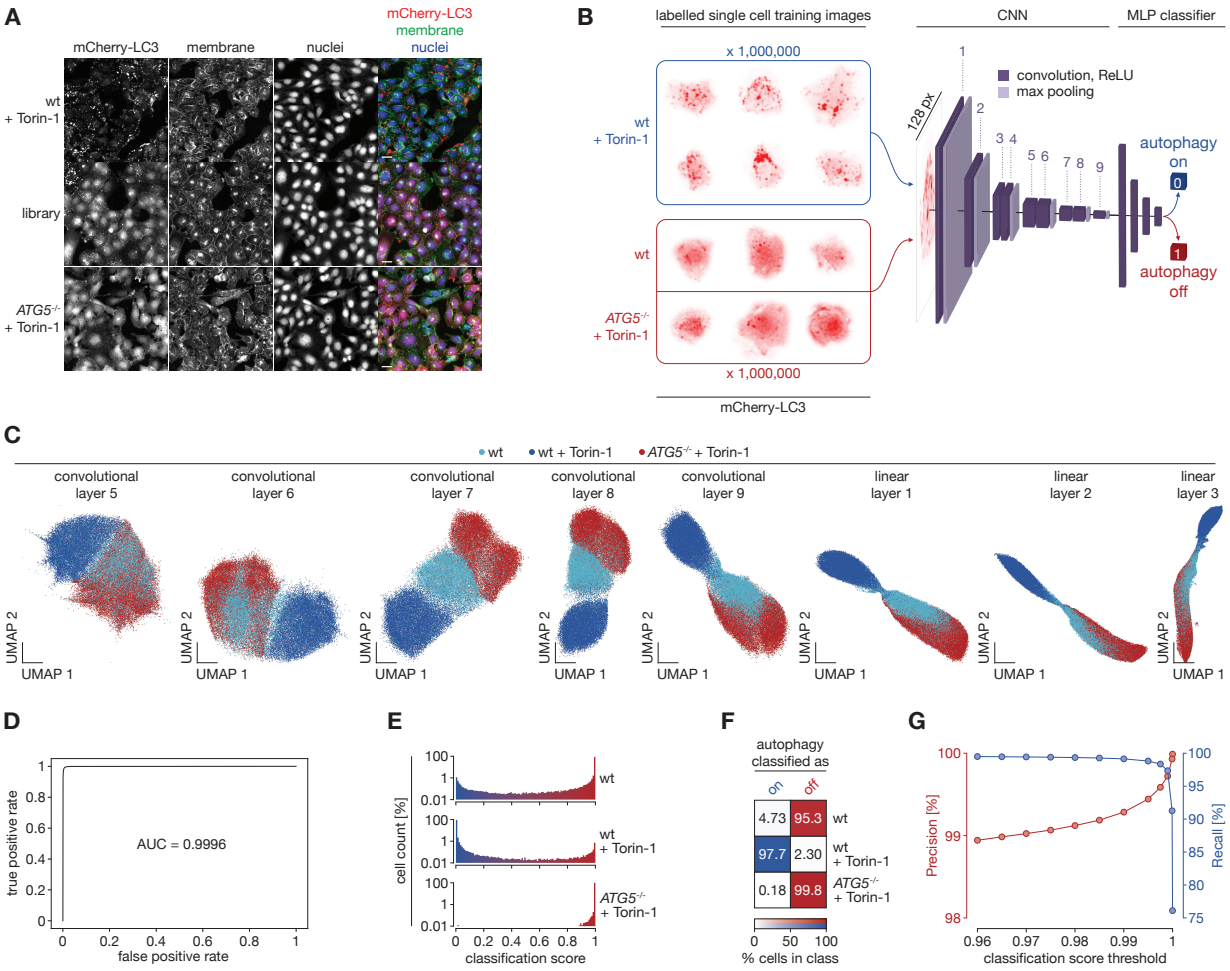


Figure 3

Figure 3 | Deep learning accurately identifies autophagy-defective cells

(A) Unsegmented images that were used for training autophagy classifier 2.0 after segmentation. Membranes were stained with WGA-Alexa488. "Library" refers to cells transduced with the Brunello CRISPR KO library. Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification. Scale bars represent 30 µm.

(B) Overview of the architecture and training paradigm of the convolutional neural network-based classifier 2.0 for autophagic or non-autophagic distribution of mCherry-LC3 in single U2OS cells. 1 million 128 × 128 px single cell images from several biological replicates were used in each training class. The autophagy-on class consisted of images of wildtype cells stimulated with Torin-1 pre-filtered for responsive cells. The autophagy-off class consisted of images of unstimulated wildtype cells pre-filtered to remove cells showing spontaneous autophagosome formation and images from two different *ATG5*$^{-/-}$ clones. Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification. CNN: convolutional neural network. MLP: multilayer perceptron.

(C) UMAPs of mCherry-LC3 images of single U2OS cells featurized through the autophagy classifier 2.0 illustrated in (B) up to the indicated layers. Colors depict the indicated genotypes and treatments. 20,000 cells are shown for each genotype and treatment.

(D) Receiver Operating Characteristic (ROC) curve of the autophagy classifier 2.0. AUC: Area under the curve.

(E) Histograms of images of mCherry-LC3 expressing U2OS cells of the indicated genotypes treated as indicated after autophagy classification with our classifier 2.0 as illustrated in (B).

(F) Heatmap showing the percentage of cells in e classified as autophagy-on or autophagy-off with a classification score threshold of 0.5.

(G) Precision (Percent *ATG5*$^{-/-}$ among cells classified as autophagy-off from an equal mix of Torin-1 stimulated wildtype cells and *ATG5*$^{-/-}$ cells) and recall (Percent *ATG5*$^{-/-}$ cells classified as autophagy-off) of our autophagy classifier at different thresholds for classifying cells as "autophagy-defective".

The data used for (C) – (G) come from an independent test dataset that was not used during training of the autophagy classifier.

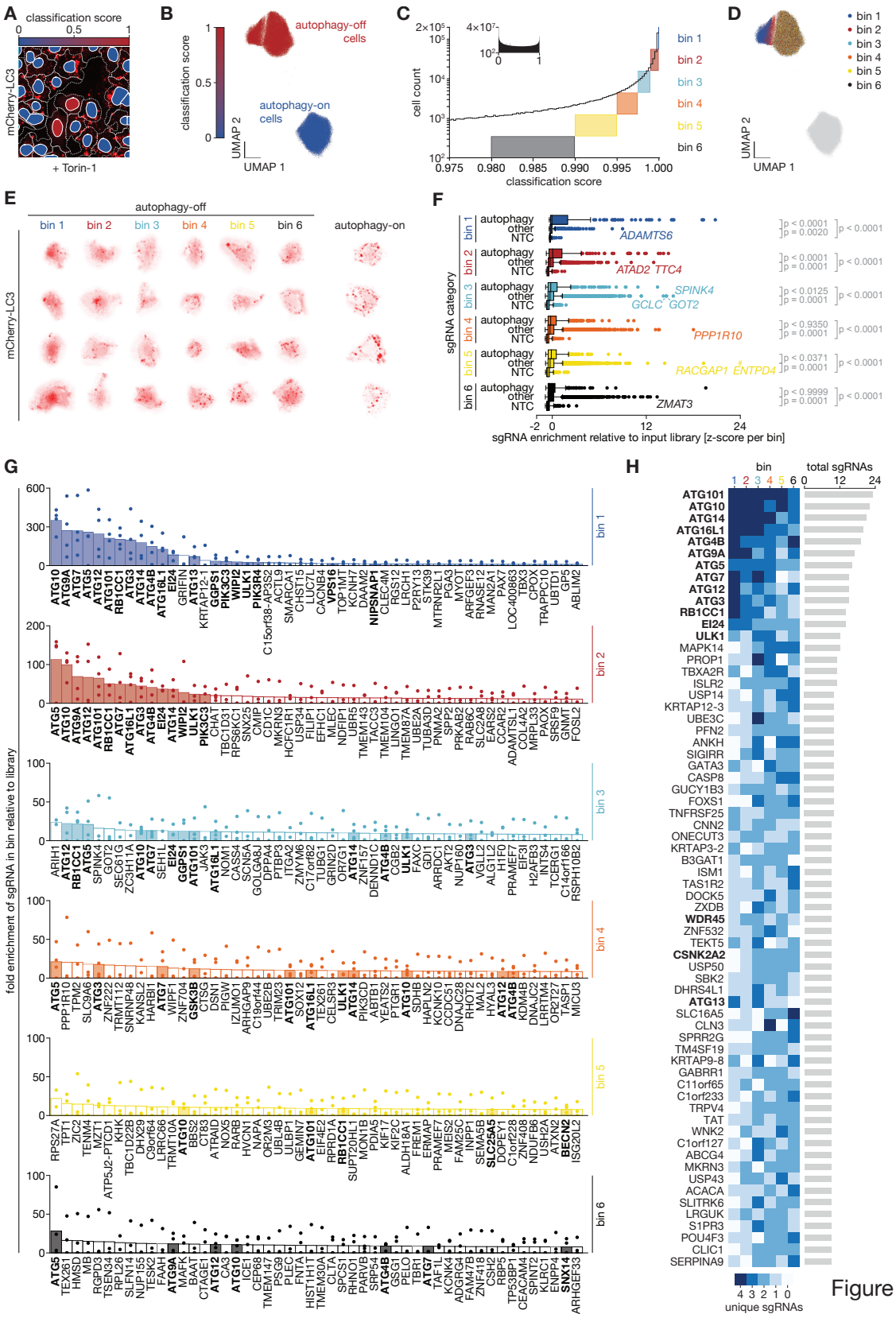


Figure 4

Figure 4 | Genome-wide CRISPR screening for autophagosome formation in 40 million U2OS cells using SPARCS

(A) Example region from a genome-wide SPARCS CRISPR knockout screen on autophagosome formation in mCherry-LC3 expressing U2OS cells after Torin-1 stimulation for 14 hrs. Colors in nuclei indicate the result of binary autophagy classification with the classifier 2.1, dotted lines indicate cytosol segmentation. Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification.

(B) Histogram of autophagy classification scores in the genome-wide CRISPR KO library batch 2 (inset) zoomed in on cells classified as autophagy-off with a score above 0.975. Colored boxes illustrate the binning strategy we used to isolate cells for sgRNA sequencing.

(C) UMAP representation of single cell images from all cells in screen batch 2 with a classification score ≥ 0.98 (dark blue) or < 0.02 (light blue) featurized through the first 8 convolutional layers of autophagy classifier 2.1. 91,320 images are depicted for each category.

(D) As C but colored according to our binning strategy along different autophagy classification thresholds as outlined in (B). 15,220 images are depicted per bin. Right panel shows a magnification of the UMAP region containing the putative screening hits.

(E) Images of individual cells from each bin in screen batch 2.

(F) z-scored enrichments of individual sgRNAs in each bin from batches 1 & 2. Vertical lines depict median, boxes depict interquartile range (IQR) and whiskers depict 1.5 × IQR. #: One sgRNA targeting the gene *ENTPD4* with a z-score of 42.1 in bin 5 is not depicted. p-values were calculated with a Kruskal-Wallis test followed by Dunn's test. NTC: non-targeting control.

(G) sgRNA sequencing results of the top 50 genes in each of the six bins filtered for genes for which we found at least two different sgRNAs in the respective bin in batches 1 & 2. Enrichment is calculated as the fraction of reads for an sgRNA in the respective bin divided by the fraction of reads of that sgRNA in the entire library. Bars indicate average enrichment per gene calculated from the enrichment of individual sgRNAs indicated as dots. Filled bars depict autophagy-related genes highlighted in bold.

(H) Number of different sgRNAs per gene in each bin for all genes with at least 9 total sgRNAs across all bins. sgRNAs were counted if they were sequenced with a read fraction in the top 50 % per bin.

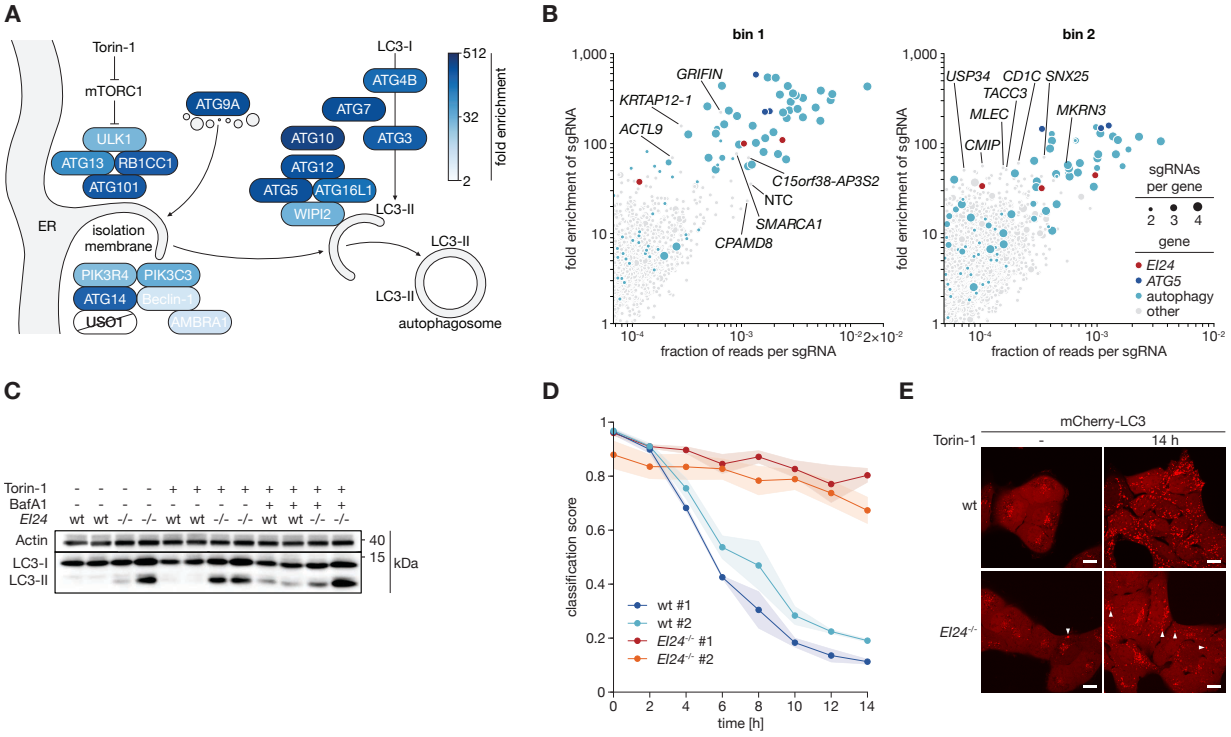


Figure 5

Figure 5 | Analysis of hits from genome-wide SPARCS screen

(A) Overview of the canonical macroautophagy pathway. Colors indicate the highest enrichment value we found for a given gene in any bin. *USO1* was not found with at least two different sgRNAs in any single bin.

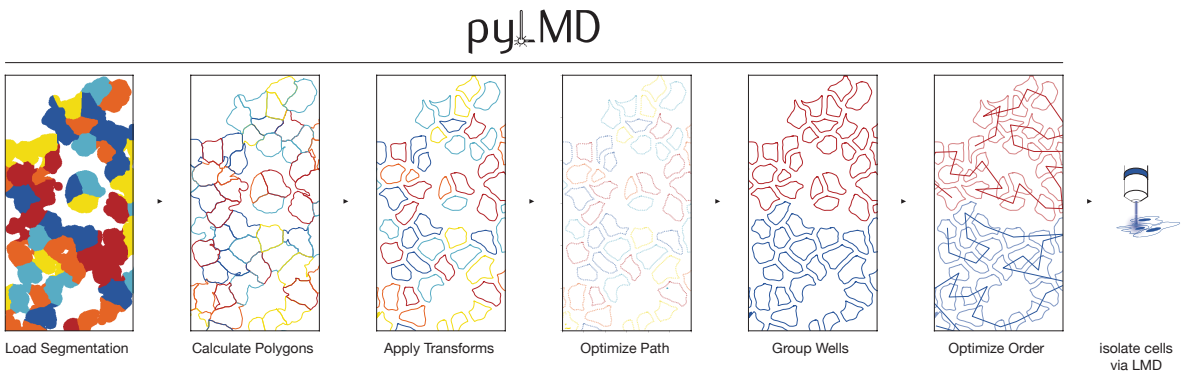
(B) Enrichment vs read count for individual sgRNAs in the top two bins for genes where we found at least two different sgRNAs in the respective bin. Circle sizes indicate the total number of different sgRNAs we found for a given gene, colors indicate different groups of genes. Individual sgRNAs from the "other" group are annotated. NTC: non-targeting control.

(C) Immunoblot of endogenous LC3 lipidation in wildtype and *EI24*$^{-/-}$ mCherry-LC3 and LckLip-mNeon expressing U2OS cell. Two clones are shown per genotype. One representative of three independent experiments.

(D) Time course analysis of autophagy classification in clones of wildtype and *EI24*$^{-/-}$ mCherry-LC3 and LckLip-mNeon expressing U2OS cells. Cells were treated with Torin-1 for up to 14 hrs. Dots represent average classifier scores from cells in 15 fields of view per timepoint and clone from three independent experiments, shaded areas represent SEM. Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification.

(E) Images of live mCherry-LC3 and LckLip-mNeon expressing wildtype and *EI24*$^{-/-}$ U2OS cells after 14 hrs of Torin-1 stimulation. Arrowheads indicate larger mCherry-LC3 aggregates. Images were acquired on a Nikon Eclipse Ti2 confocal microscope with 100 x magnification. Scalebars represent 15µm. One representative of three independent experiments.

A



B

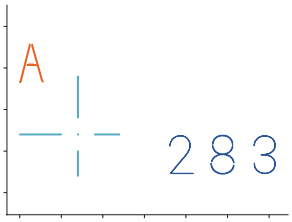
```
from lmd.lib import Collection
from lmd import tools

my_collection = Collection()

my_collection.join(
    tools.makeCross([20, 20], [50, 30, 30, 50], 1, 10)
)

my_collection.join(
    tools.glyph('A', offset=(-50, 130))
)

my_collection.join(
    tools.text('283', offset=(130, 20))
)
```



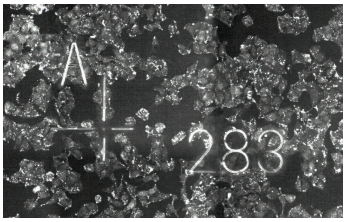


Figure S1

Figure S1 | The py-lmd python library generates cutting paths for automated laser microdissection

(A) Overview of cutting path generation with py-lmd.

(B) py-lmd allows the generation of arbitrary shapes such as calibration crosses.

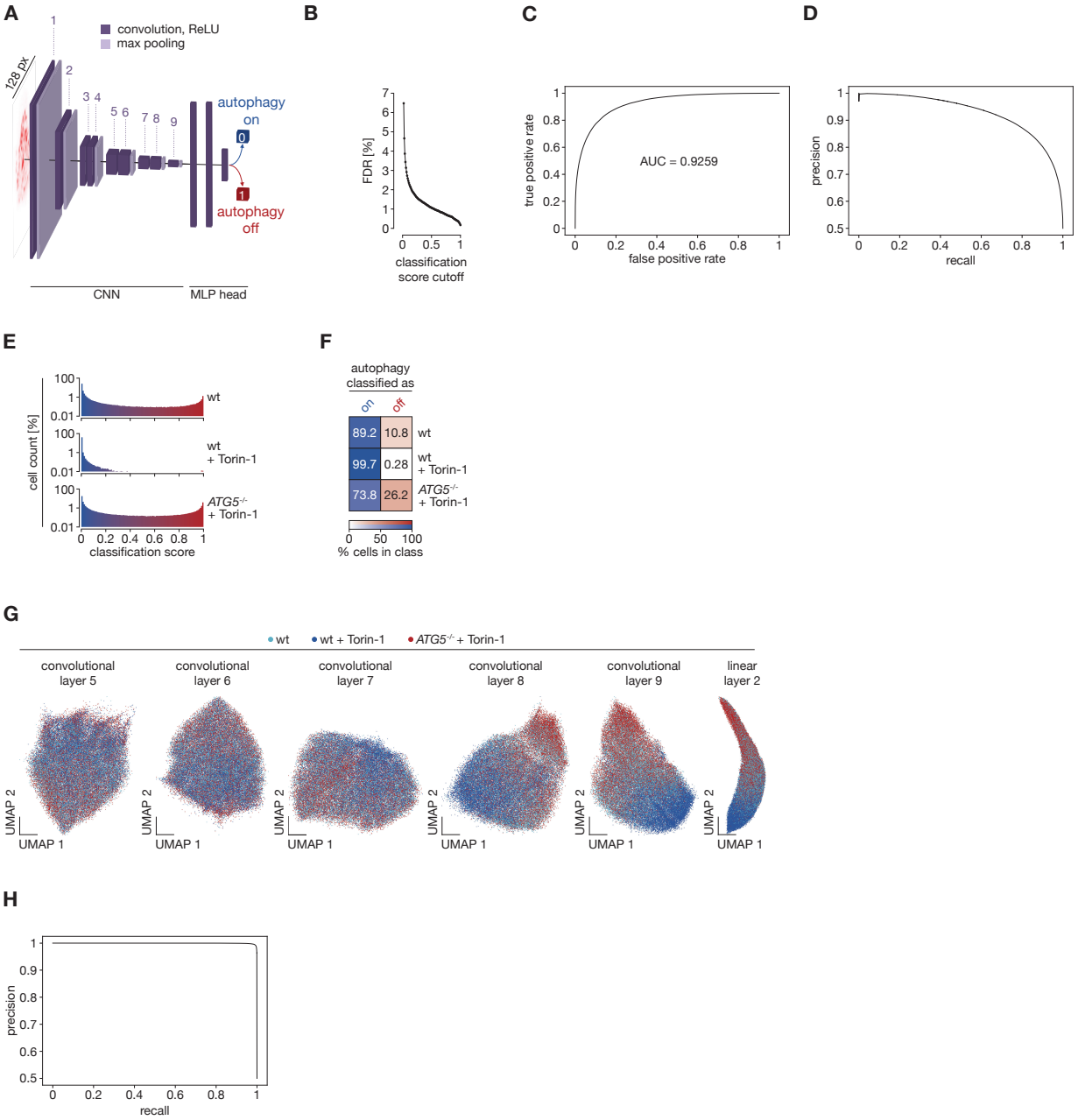


Figure S2

Figure S2 | Performance of LC3 image-based autophagy classifiers

(A) Overview of the architecture and training paradigm of our convolutional neural network-based classifier 1.0 for autophagic or non-autophagic distribution of mCherry-LC3 in single U2OS cells. 500,000 128 × 128 px single cell images from several biological replicates were used in each training class. The autophagy-on class consisted of images of wildtype cells stimulated with Torin-1. The autophagy-off class consisted of images of unstimulated wildtype cells and images from two different *ATG5*$^{-/-}$ clones. CNN: convolutional neural network. MLP: multilayer perceptron.

(B) False discovery rates (FDR) of the autophagy classifier 1.0 at different classification score cutoffs.

(C) Receiver Operating Characteristic (ROC) curve for autophagy classifier 1.0. AUC: area under the curve.

(D) Precision-Recall curve for our autophagy classifier 1.0.

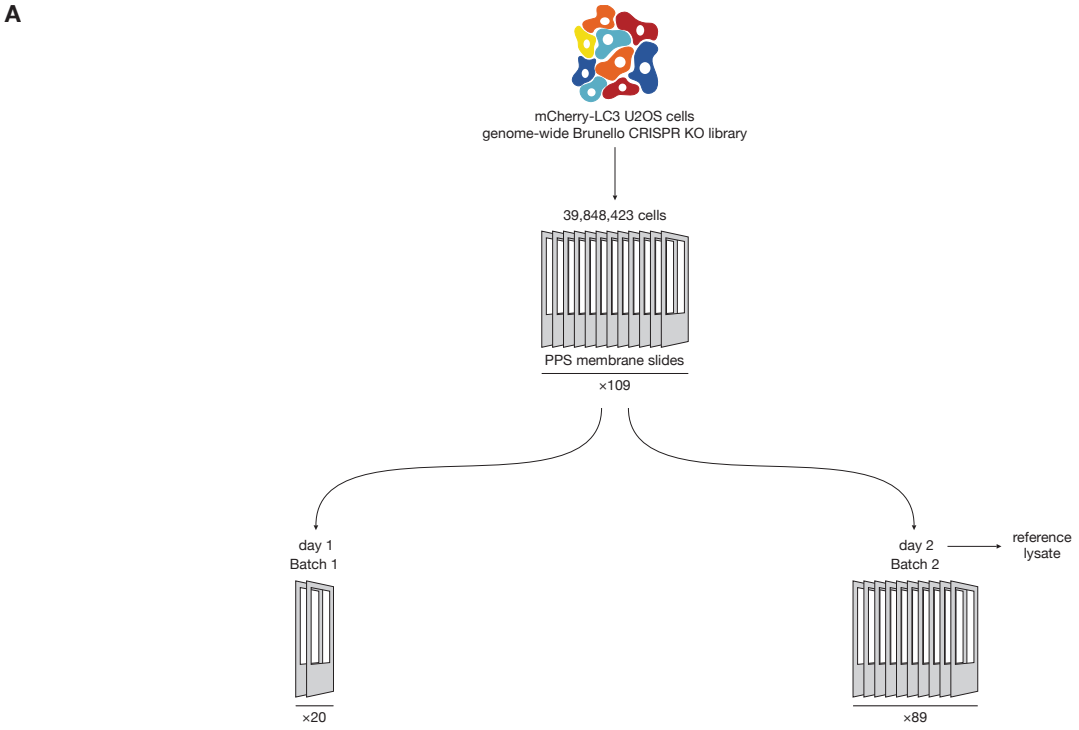
(E) Histograms of images of mCherry-LC3 expressing U2OS cells of the indicated genotypes treated as indicated after autophagy classification with classifier 1.0 as illustrated in (A).

(F) Heatmap showing the percentage of cells in d classified as autophagy-on or autophagy-off with a classification score threshold of 0.5.

(G) Parametric UMAPs of mCherry-LC3 images of single U2OS cells featurized through our autophagy classifier 1.0 illustrated in (A) up to the indicated layers. Colors depict the indicated genotypes and treatments. 20,000 cells are shown for each genotype and treatment.

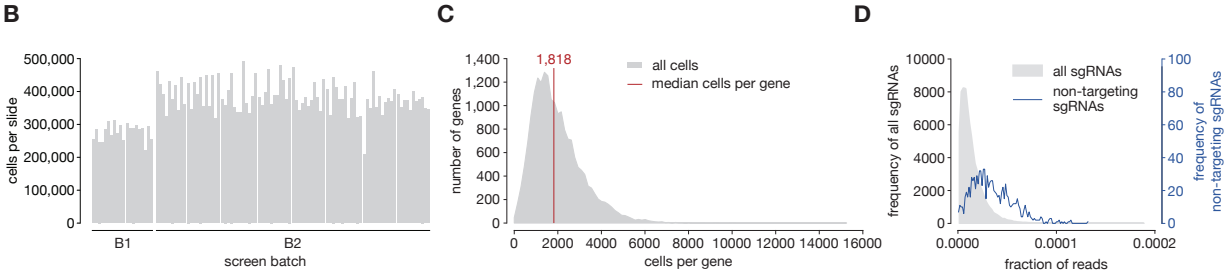
(H) Precision-Recall curve of classifier 2.0.

(B) – (H) were calculated on independent test datasets for the respective classifiers



bin	score [>]	cells imaged	nuclei excised
bin 1	0.99999	937	795
bin 2	0.999	10,794	9,113
bin 3	0.9975	5,021	4,387
bin 4	0.995	4,540	3,968
bin 5	0.99	4,839	4,270
bin 6	0.98	5,121	4,525
library		5,539,333	27,058

bin	score [>]	cells imaged	nuclei excised
bin 1	0.99999	19,516	14,009
bin 2	0.999	217,766	*179,898**
bin 3	0.9975	74,672	52,283
bin 4	0.995	56,478	41,061
bin 5	0.99	55,477	41,695
bin 6	0.98	56,102	43,189
library		34,309,090	372,135



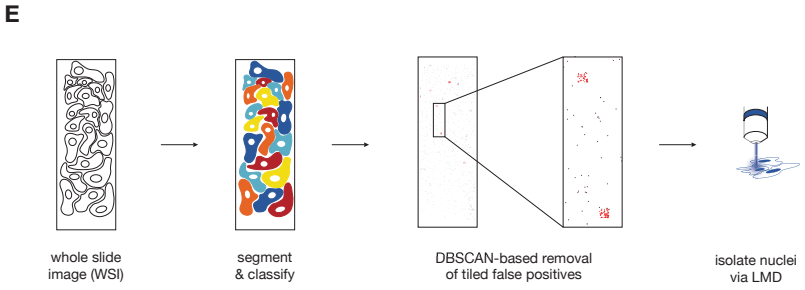


Figure S3

Figure S3 | Overview of genome-wide SPARCS screening for autophagy

(A) Batching and binning strategy for screening autophagosome formation in 40 million mCherry-LC3 expressing U2OS cells. We dissected fewer cells than we imaged for a given bin due to the quality control step outlined in e. *The efficiency of the PCR on bin 2 from batch 2 had decreased dramatically, presumably due to the high density of membrane fragments in the reaction, leading to a loss of sgRNAs for sequencing. PPS: polyphenylene sulfide.

(B) Number of cells segmented per screen slide.

(C) Distribution of human genes targeted in the screen across cells in the library as determined by deep sequencing.

(D) Distribution of non-targeting and targeting sgRNAs in the reference library as determined by deep sequencing.

(E) Quality control strategy for false positives arising from out-of-focus images. When we spatially clustered hits using DBSCAN, we found clusters above a certain size to correspond to entire out-of-focus imaging tiles and removed these clusters before nuclei excision.

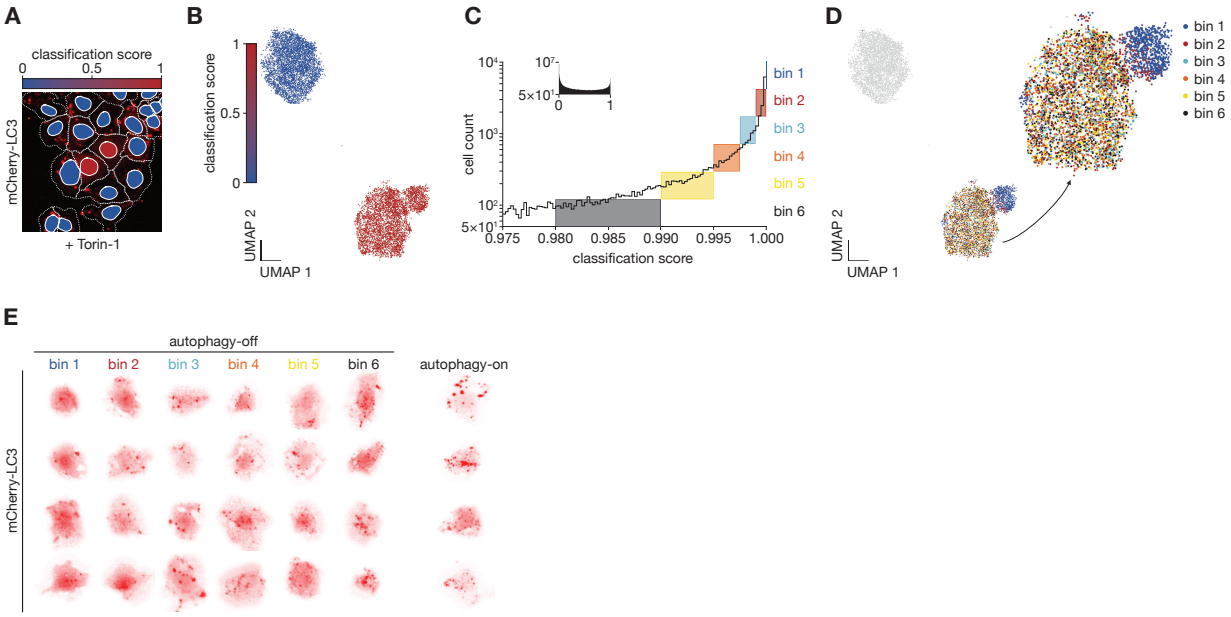


Figure S4

Figure S4 | Results from genome-wide autophagy screen batch 1

(A) Example region from a genome-wide SPARCS CRISPR knockout screen on autophagosome formation in mCherry-LC3 expressing U2OS cells after Torin-1 stimulation for 14 hrs. Colors in nuclei indicate the result of binary autophagy classification with classifier 2.0, dotted lines indicate cytosol segmentation. Images were acquired on an Opera Phenix microscope in confocal mode with 20 x magnification.

(B) UMAP representation of single cell images from all cells in screen batch 1 with a classification score ≥ 0.98 (dark blue) or < 0.02 (light blue) featurized through the first 8 convolutional layers of autophagy classifier 2.0. 4,806 cells are depicted per category.

(C) Histogram of autophagy classification scores in the genome-wide CRISPR KO library batch 1 (inset) zoomed in on cells classified as autophagy-off with a score above 0.975. Colored boxes illustrate the binning strategy we used to isolate cells for sgRNA sequencing.

(D) As (B) but colored by screening bin. 801 cells shown per bin.

(E) Images of individual cells from each bin in screen batch 1.

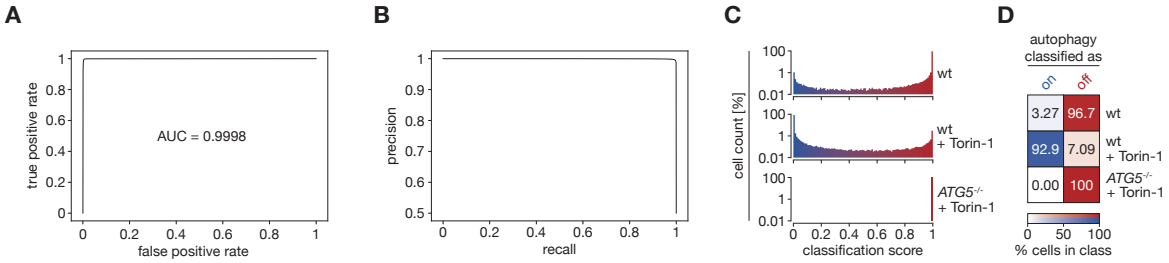


Figure S5

Figure S5 | Performance metrics of classifier 2.1

(A) Receiver Operating Characteristic (ROC) curve for autophagy classifier 2.1. AUC: area under the curve.

(B) Precision-Recall curve for our autophagy classifier 2.1.

(C) Histograms of images of mCherry-LC3 expressing U2OS cells of the indicated genotypes treated as indicated after autophagy classification with classifier 2.1.

(D) Heatmap showing the percentage of cells in (C) classified as autophagy-on or autophagy-off with a classification score threshold of 0.5.

(A) – (D) were calculated on an independent test dataset.

Table S1 | Overview of CNN-based image classifiers trained in this study

All classifiers were trained on 128 × 128 px single cell images using PyTorch lightning. Unstimulated control slides containing library cells plated in parallel with the screen slides were included during training to capture possible batch effects introduced during plating and staining of the screening library. Cells were pre-filtered according to their autophagy score using classifier version P where indicated.

Table S1

Classifier Version	Description	Architecture	Number of trainable parameters	Training data	Training Epochs	Number of cells per class	Fig.	Used to classify dataset	Independent Test Dataset	AUC ROC Curve
1.0	Trained on initial staining protocol. Used to classify pilot screen	As in Fig. 2b but classifier head only consists of 3 fully connected linear layers	17,882,244	2 slides unstimulated wt 2 slides wt + Torin-1 1 slide *ATG5*$^{-/-}$ clone 1 1 slide *ATG5*$^{-/-}$ clone 2	40	500,000	2, S2	A 1.2 million cells	1 slide *ATG5*$^{-/-}$ cells clone 1 1 slide unstimulated wt cells 1 slide wt cells + Torin-1	0.925879
P	Only used to pre-filter cells for training 2.0 & 2.1	As in first classifier	17,882,244	3 slides wt + Torin-1 3 slides *ATG5*$^{-/-}$ mixed clones 1 slide *ATG5*$^{-/-}$ clone 1 1 slide *ATG5*$^{-/-}$ clone 2	30	1,200,000			1 slide *ATG5*$^{-/-}$ cells clone 1 1 slide unstimulated wt cells 1 slide wt cells + Torin-1	
2.0	Used on initial batch of genome-wide screen	As shown in Fig. 2b	14,340,484	1 slide *ATG5*$^{-/-}$ clone 1 1 slide *ATG5*$^{-/-}$ clone 2 3 slides *ATG5*$^{-/-}$ mixed clones 2 slides unstimulated screen cells score > 0.9 (autophagy-off) 3 slides wt + Torin-1 score < 0.1 (autophagy-on)	20	1,000,000	3, S4	Screen batch 1 5 million cells	1 slide *ATG5*$^{-/-}$ cells clone 1 1 slide unstimulated wt cells 1 slide wt cells + Torin-1	0.999649

2.1	Refined for largest part of genome-wide screen	As shown in Fig. 2b	14,340,484	1 slide *ATG5*$^{-/-}$ clone 1 1 slide *ATG5*$^{-/-}$ clone 2 3 slides *ATG5*$^{-/-}$ mixed clones 2 slides unstimulated screen cells score > 0.999 (autophagy-off) 3 slides wt + Torin-1 score < 0.001 (autophagy-on)	20	1,000,000	4, S5	Screen batch 2 35 million cells	1 slide *ATG5*$^{-/-}$ cells clone 1 1 slide unstimulated wt cells 1 slide wt cells + Torin-1	0.999772

Table S2 | Comparison of high-throughput methods for combined spatial phenotyping and genotyping

Search space: library size that can be screened for phenotypes. Target space: Proportion of library that can be analyzed. Phenotypic variants that can be discriminated: The maximum number of different phenotypes that can be recovered from a single screen. Real time decision for genotyping necessary: Whether a decision has to be made for a given image in real time during screening ("yes") or whether entire single cell datasets can be analyzed after imaging before a decision on which cells to genotype has to be made ("no").

*A genome-wide screen using in situ-seq has recently been reported (ref 21) with a small library of 10 million cells in which the number of screened and successfully sequenced cells and sgRNA representation remain unclear.

°These technologies have low costs per screened cell, but require the use of instruments often provided by core facilities such as a laser dissection microscope for SPARCS, an imaging sorter device for imaging flow cytometry or an imaging setup equipped with a fluorescence recovery after photobleaching (FRAP) laser for pA-mCherry.

Table S2

	SPARCS	In situ seq	Imaging flow cytometry	pA-mCherry
References	This study	19, 20, 21	22, 23	25, 26, 27
Search space	large	medium	large	large
Target space	small	medium	large	small
Phenotypic variants that can be discriminated	microtiter plate	unlimited	microtiter plate	4 per fluorophore
Image quality	confocal	confocal	flow-based	confocal
Real time decision for genotyping necessary	No	No	Yes	Yes
Discovery of new phenotypes after screen by reanalysis with new computational model possible	Yes	Yes	No	No
Largest library size	40 million	31 million	12 million	12.6 million
Genes targeted	19,114	5,072*	18,408	18,905
Special equipment required	Laser microdissection microscope	Ultrafast imaging setup and precise stage for sequencing cycles	Imaging flow sorter	FRAP laser or equivalent
Cost per cell in library	Low˚	High	Low˚	Low˚

4.4 Deep Visual Proteomics maps proteotoxicity in a genetic liver disease

Alpha-1 antitrypsin deficiency (AATD), is a fibrogenic liver disease that is characterised by the misfolding and accumulation of alpha-1 antitrypsin (AAT) in hepatocytes. Despite having a homozygotic incidence of 1:2000, the progression mechanisms determining fibrogenesis or hepatocyte survival remain unclear which limits treatment options. Using our DVP technology, we characterised the proteomic makeup of hepatocytes with AAT accumulation to gain a better understanding of the underlying molecular mechanisms. By using a computer vision model that was pretrained on a large collection of natural images we were able to stratify cells according to their AAT aggregate morphology in an unbiased manner. Combining this unbiased AI-driven phenotyping with proteomics, resulted in the identification of a terminal hepatocyte state marked by globular protein aggregates with a distinct proteomic signature. The molecular targets identified through this analysis provide a valuable resource to better characterise AATD and perhaps intervene clinically. Furthermore, the approach established here of combining unbiased image featurisation with DVP, provides a robust framework for dissecting complex cellular processes in situ across a spectrum of proteotoxic diseases.

The following research article was originally published here:

Rosenberger, F. A., Mädler, S. C., et al. (2025). "Deep Visual Proteomics maps proteotoxicity in a genetic liver disease". In: *Nature* 642.8067, pp. 484–491. ISSN: 0028-0836. DOI: 10.1038/s41586-025-08885-4

Article

Deep Visual Proteomics maps proteotoxicity in a genetic liver disease

https://doi.org/10.1038/s41586-025-08885-4

Received: 26 August 2024

Accepted: 11 March 2025

Published online: 16 April 2025

Open access

Check for updates

Florian A. Rosenberger[1]✉, Sophia C. Mädler[1,13], Katrine Holtz Thorhauge[2,3,13], Sophia Steigerwald[1,13], Malin Fromme[4], Mikhail Lebedev[1], Caroline A. M. Weiss[1], Marc Oeller[1], Maria Wahle[1], Andreas Metousis[1], Maximilian Zwiebel[1], Niklas A. Schmacke[1,5], Sönke Detlefsen[3,6], Peter Boor[7], Ondřej Fabián[8,9], Soňa Fraňková[10], Aleksander Krag[2,3,11], Pavel Strnad[4] & Matthias Mann[1,12]✉

Protein misfolding diseases, including α1-antitrypsin deficiency (AATD), pose substantial health challenges, with their cellular progression still poorly understood[1–3]. We use spatial proteomics by mass spectrometry and machine learning to map AATD in human liver tissue. Combining Deep Visual Proteomics (DVP) with single-cell analysis[4,5], we probe intact patient biopsies to resolve molecular events during hepatocyte stress in pseudotime across fibrosis stages. We achieve proteome depth of up to 4,300 proteins from one-third of a single cell in formalin-fixed, paraffin-embedded tissue. This dataset reveals a potentially clinically actionable peroxisomal upregulation that precedes the canonical unfolded protein response. Our single-cell proteomics data show α1-antitrypsin accumulation is largely cell-intrinsic, with minimal stress propagation between hepatocytes. We integrated proteomic data with artificial intelligence-guided image-based phenotyping across several disease stages, revealing a late-stage hepatocyte phenotype characterized by globular protein aggregates and distinct proteomic signatures, notably including elevated TNFSF10 (also known as TRAIL) amounts. This phenotype may represent a critical disease progression stage. Our study offers new insights into AATD pathogenesis and introduces a powerful methodology for high-resolution, in situ proteomic analysis of complex tissues. This approach holds potential to unravel molecular mechanisms in various protein misfolding disorders, setting a new standard for understanding disease progression at the single-cell level in human tissue.

Spatial omics technologies are revolutionizing our ability to deconvolute molecular events at single-cell resolution within a tissue context. Whereas much focus has been placed on spatial genomics and transcriptomics, recent advances in multiplexed imaging and proteomics are beginning to shed light on the functional proteomic layer. Mass spectrometry (MS)-based proteomics has made significant strides towards biologically informative single-cell analysis, now enabling quantification of up to 5,000 proteins in cultured cells[6–8]. In the tissue context, we have recently introduced Deep Visual Proteomics (DVP), which integrates staining, artificial intelligence-guided cell segmentation and classification, laser microdissection of single-cell shapes and high-sensitivity MS[4,5]. DVP excels in digital pathology applications with pronounced spatial and visual components, providing simultaneous and deep proteomic characterization at the level of thousands of proteins[9].

We reasoned that these emerging technologies would be ideally suited to elucidate molecular events during the progressive worsening of proteotoxicity as it unfolds in patients. Proteotoxicity, characterized by the accumulation of misfolded and aggregated proteins leading to cell damage, is a hallmark of many diseases, including neurodegenerative pathologies such as Alzheimer's disease and Parkinson's disease[10–12]. The underlying cause of proteotoxicity is a disruption in protein homeostasis, resulting in an imbalance between protein synthesis, folding and clearance mechanisms[3].

To investigate proteotoxicity in a clinically relevant context, we focused on a disorder with unmet clinical need that exemplifies the challenges of protein misfolding and aggregation in a vital organ. The fibrogenic liver disease α1-antitrypsin (AAT) deficiency (AATD) is a genetic disorder caused by autosomal, codominant mutations in the *SERPINA1*

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Department of Gastroenterology and Hepatology, Centre for Liver Research, Odense, Denmark. [3]Department of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [4]Medical Clinic III, Gastroenterology, Metabolic Diseases and Intensive Care, University Hospital RWTH, AachenHealth Care Provider of the European Reference Network on Rare Liver Disorders (ERN RARE LIVER), Aachen, Germany. [5]Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, Munich, Germany. [6]Department of Pathology, Odense University Hospital, Odense, Denmark. [7]Institute of Pathology, University Hospital Aachen RWTH, Aachen University, Aachen, Germany. [8]Clinical and Transplant Pathology Centre, Institute for Clinical and Experimental Medicine, Prague, Czech Republic. [9]Department of Pathology and Molecular Medicine, Third Faculty of Medicine, Charles University and Thomayer Hospital, Prague, Czech Republic. [10]Department of Hepatogastroenterology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic. [11]Danish Institute of Advanced Study (DIAS), University of Southern Denmark, Odense, Denmark. [12]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [13]These authors contributed equally: Sophia C. Mädler, Katrine Holtz Thorhauge, Sophia Steigerwald. ✉e-mail: rosenberger@biochem.mpg.de; mmann@biochem.mpg.de

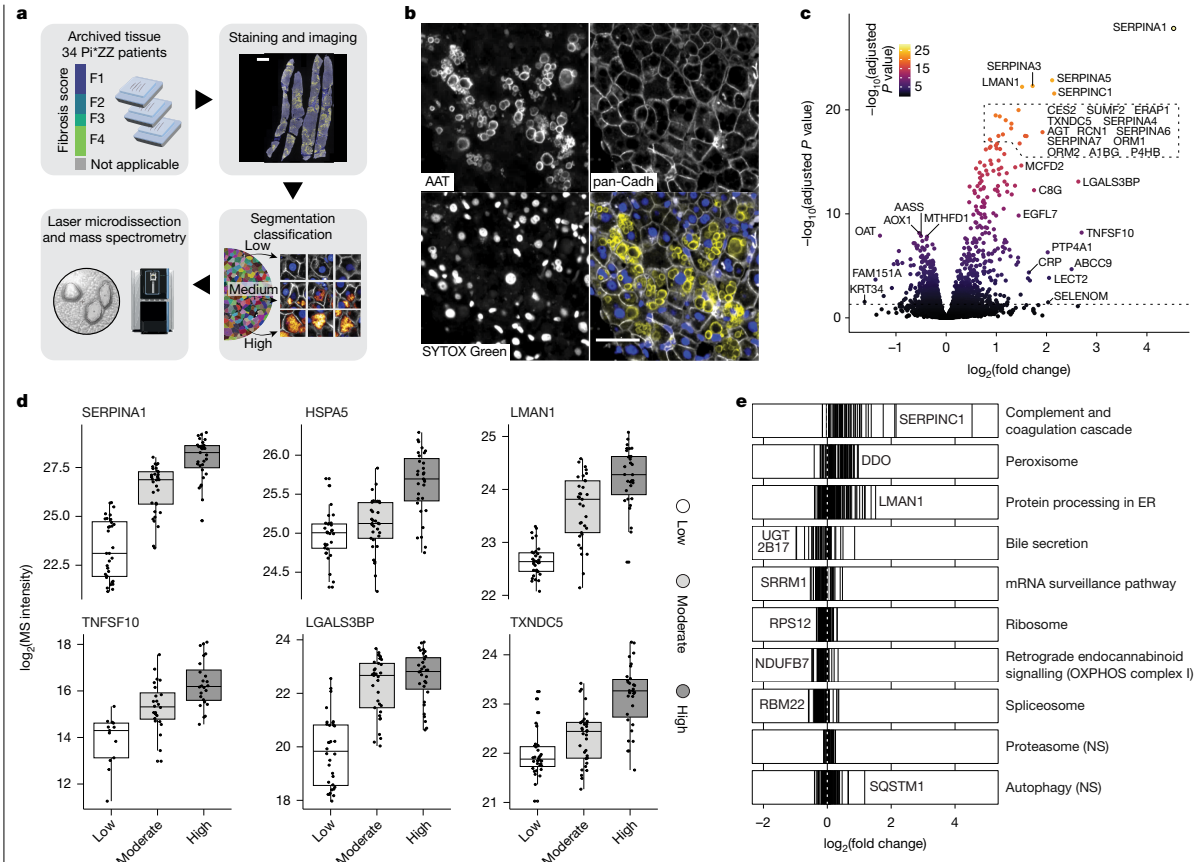


Fig. 1 | Proteomic mapping of hepatocyte stress response. a, Overview of the Deep Visual Proteomics workflow. Fibrosis stages are Kleiner scores. **b**, Immunofluorescence staining of AAT, the cell outline marker pan-cadherin (pan-Cadh), nucleus (SYTOX Green) and three-colour overlay. **c**, Proteomic changes in high versus moderate versus low AAT-accumulating cells. Enriched in high on the right side. Top significant and top changed hits are named (paired two-sided moderated *t*-test with load class as covariable, multiple testing corrected; $n = 96$ at 100 shapes per sample). **d**, MS intensity of selected proteins across three classes. One dot is one sample from a patient ($n = 34$). Boxplots show first and third quartiles (box), median (thick line) and whiskers (±1.5 interquartile range). **e**, Significantly (FDR < 0.05) enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways after GSEA. Each line is a member of the pathway. NS, not significant. Scale bars, 1 mm (**a**), 50 µm (**b**).

gene, resulting in misfolding and accumulation of AAT in hepatocytes. Most severe AATD cases are caused by a homozygous Z-variant (Pi*ZZ genotype) with a peak incidence of 1:2,000 in individuals of European descent[1,2,13,14]. Current hypotheses suggest that the severity of liver damage correlates with the amount of accumulated AAT[15–20]. However, the mechanisms driving fibrogenesis or hepatocyte survival versus death remain unclear, leaving potentially druggable targets unexplored.

To address this challenge, we curated a cohort of formalin-fixed paraffin-embedded (FFPE) biopsies and liver explants from patients homozygous for the pathogenic Z-variant, encompassing all fibrosis stages ($n = 34$; Extended Data Fig. 1a and Supplementary Table 1). Despite the same underlying disease-causing mutation at a similar median age (58 ± 10 (s.d.) years) and BMI (25.2 ± 4.0), fibrosis stages varied drastically, indicating unexplored molecular resilience or risk profiles.

Proteomic map of proteotoxic response

To elucidate the molecular basis of the observed clinical heterogeneity in patients with AATD, we implemented a comprehensive proteomic mapping approach to characterize hepatocyte responses to proteotoxic stress. We first laser microdissected 3-µm-thick FFPE sections from patient biopsies and analysed them with MS following our DVP workflow. After staining for cell outlines and AAT, we segmented and stratified cells into low, moderate and high aggregate load groups on the basis of their microscopy images (Fig. 1a,b). The proteome of 100 shapes—equivalent to the volume of 10–15 complete hepatocytes—was then acquired on the recently introduced Orbitrap Astral mass spectrometer, yielding a high-quality dataset with a mean proteomic depth exceeding 5,000 proteins per sample (Extended Data Fig. 1b,c and Supplementary Table 1). We observed a striking 23-fold difference in AAT levels between low- and high-load cells. The AAT load was captured on the second principal component, preceded only by the fibrosis stage on the first and second component (Extended Data Fig. 1d–f). Given the sparsity of AAT$^+$ cells in biopsy material, this validated our laser microdissection approach as it allowed the biological phenotype to emerge more clearly. Biopsies with a low fibrosis stage exhibited lower AAT baseline loading compared with high fibrosis stages on both proteomics and imaging data, in line with previous findings[15], whereas the maximum load remained fairly equal across

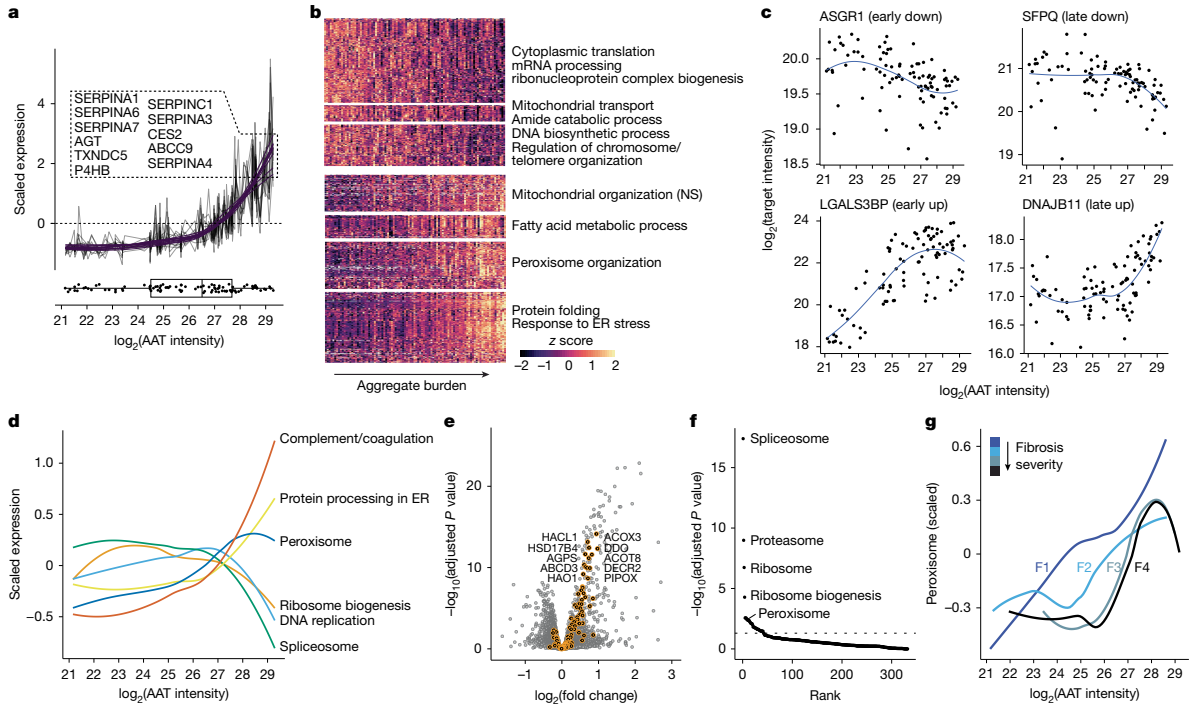


Fig. 2 | Early and late responses to proteotoxic stress. a, Expression profile of the top ten proteins correlating with AAT. All DVP samples are plotted, and values belonging to the same protein are on one line. Purple, polynomial fit (third order). Boxplot, distribution of AAT expression values along the *x* axis. **b**, Clustering of significantly (FDR < 0.01) changed proteins into early and late-responding genes to proteotoxic stress, ordered on *x* axis by AAT levels. The *y* axis was broken into seven groups to achieve good coverage of all response types. Significant KEGG terms per box are shown. **c**, Pseudotime expression of top early and late responders by directionality. **d**, Cumulative changes of indicated KEGG pathways expressed as *z* scores. **e**, Changes in protein levels across three AAT bins, highlighting peroxisomal proteins. Top ten significant hits are named (paired two-sided moderated *t*-test with load class as covariable, multiple testing corrected; $n = 96$). **f**, Top differential functional categories between F1 and F4 fibrotic samples during early AAT accumulation ($\log_2$ (AAT intensity) <25; two-sided Wilcoxon test, multiple testing corrected). **g**, Cumulative expression of peroxisomal proteins across four fibrosis stages.

all stages (Extended Data Fig. 1g,h). The proteomes of the three load classes differed markedly (17.4% significant hits at <5% false discovery rate (FDR), paired two-sided moderated *t*-test; Fig. 1c). Alongside AAT, several known markers of AATD liver pathology were highly enriched in aggregate-positive cells, such as a 1.6-fold increased endoplasmic reticulum (ER) chaperone HSPA5 and a 2.9-fold increased ER–Golgi cargo receptor LMAN1 (Fig. 1d)[21–23].

Among the most dysregulated hits, we identified other secretory proteins, including many unambiguous SERPINs, coagulation and complement factors (Fig. 1c and Extended Data Fig. 2a–d). This aligns with recent findings of SERPIN sequestration in AAT-inclusions, and supports the notion of crowding in the ER space[18,24], with potential systemic pathological implications due to accumulation of annotated plasma proteins in affected hepatocytes (Extended Data Fig. 2e). Galectin-3 binding protein LGALS3BP and the apoptotic inducer TNFSF10 had the most pronounced positive changes (Fig. 1c,d). LGALS3BP is a hepatocyte-produced protein targeted for secretion that is elevated in plasma from patients with liver disease[25]. Reports describing the immunomodulatory activity of LGALS3BP could explain the involvement of immune cells in AATD liver pathology[15,26,27].

Pathway enrichment analysis showed a strong elevation of proteins related to the three branches of unfolded protein response (UPR) mediated through ATF6, PERK and IRE1 along with a general upregulation of chaperones, accompanied by a reduction in the transcription and translation machinery. This occurred at the expense of physiological functions such as bile secretion (Fig. 1e). Many responses converged into a protective response to reactive oxygen species with upregulation of thioredoxins and glutaredoxins, including an atypical increase in the peroxisomal compartment and reduction of mitochondrial complex I (Fig. 1d and Extended Data Fig. 2a,b,f–j). Proteasomal and autophagy proteins remained largely unchanged, and neither did we detect disturbances of calcium homeostasis (Fig. 1e and Extended Data Fig. 2k).

Early and late-stage stress responses

Our experimental design, encompassing three aggregate load classes, should allow us to resolve the stepwise progression of molecular events. To determine the sequence in which molecular responses occur during AAT build-up, we first correlated AAT with other protein levels to identify 'followers' that tightly track AAT levels. Proteins of the ER were among the top ten hits, with many destined for secretion (Fig. 2a and Extended Data Fig. 3a). This included many structurally similar SERPINs, and the tight tracking of AAT levels suggests that these proteins accumulate in tandem with AAT rather than being coregulated.

We then categorized proteins into early and late responders to proteotoxic stress caused by AAT accumulation (Fig. 2b and Supplementary Table 2). We observed the most consistent relation with AAT load among coelevated proteins, with most (77.7%) manifesting as late responders and only a smaller fraction as early responders. The immunomodulatory marker LGALS3BP was most prominent among early responders, followed by the ER cargo receptor MCFD2 together with its co-binder LMAN1 (Fig. 2c). A strong peroxisomal biogenesis response

emerged early on, characterized by the peroxisomal proliferation factor PEX11B and other membrane-integral proteins, along with lipid metabolism and superoxide detoxifying proteins (Fig. 2d,e, Extended Data Figs. 3b–d and 4 and Supplementary Table 2). By contrast, most proteins of the core machinery of the UPR appeared later during AAT build-up, despite visual protein accumulation at earlier stages (Fig. 2d and Extended Data Fig. 3e,f). The crosstalk between UPR and peroxisomal activity remains poorly understood, yet lipid metabolism, cholesterol metabolism and reactive oxygen species detoxification intersect both pathways. Together, the data indicate a dominant increase of the ER oxidoreductase-1α (ERO1A)—a main peroxide producer (Fig. 1c and Extended Data Fig. 2f).

We then analysed samples at various fibrosis stages, revealing principal dysregulations with increasing fibrosis stage in proteotoxicity-responsive pathways (Fig. 2f and Extended Data Fig. 5). Notably, this included the peroxisomal response, which showed a gradually prolonged onset time relative to AAT load (Fig. 2g). Peroxisomal chaperones or chaperone-like proteins remained unaltered, suggesting that peroxisomes are unlikely to contribute to the clearance of unfolded proteins (Extended Data Fig. 3d).

Single-cell mapping in intact tissue

The accumulation of AAT in intact tissue exhibits a pronounced spatial component. Previous work has demonstrated that AAT accumulates unequally along the zonation gradient from portal to central vein axis in patients with AATD with the Pi*ZZ genotype[15,28,29]. Yet, sharp borders and the absence of gradual changes between neighbouring AAT^+ and AAT^- cells, as well as single positive cells, indicate a more complex picture (Fig. 3a). To map the spatial proteome in these regions, we built on our previous single-cell DVP workflow[5]. We isolated single shapes from selected regions in 10-µm-thick FFPE sections (equivalent to one-third to one-half of a complete hepatocyte) from six F1-stage biopsies. We selected early-stage (F1) biopsies to examine stress processes in a minimally fibrotic environment, reducing potential confounding effects from advanced disease. We quantified the proteome of these 'shapes' one at a time using the Orbitrap Astral mass spectrometer and a variable window precursor selection design (Extended Data Fig. 6a,b).

In this way, we quantified the proteome of 259 single shapes in three biopsies at a median depth of 2,785 proteins, and reaching up to 4,299 proteins (Fig. 3b, Extended Data Fig. 6c,d and Supplementary Table 3). The laser capturing proved highly precise, as evidenced by the complete separation of adjacent AAT^+ and AAT^- cells (Fig. 3a and Extended Data Fig. 6e–g). On comparing AAT^+ and AAT^- cells at border regions, we identified similar proteotoxic stress markers as before (Extended Data Figs. 6h and 7a,b). Interestingly, cells of the first or second row within a border region and within their respective AAT class displayed very similar proteomes (Fig. 3c). Consistent with this, the AAT accumulation markers LGALS3BP and ERO1A were markedly different between AAT^+ and AAT^- cells, but not among first- and second-order neighbours. Consequently, the data support an absence of dedicated stress propagation between neighbouring cells, suggesting that AAT-induced proteotoxic stress is a cell-intrinsic response.

AAT accumulation has been characterized previously as a periportal event[30]. However, our data indicate only partial or no dependence of AAT accumulation on zonation, as evidenced by no or little change in the expression levels of the portal markers ASS1, HAL and ARG1, or the central markers ADH1 and CYP2E1 at borders. We also did not observe any zonation effect in single AAT^+ cells compared with AAT^- direct neighbours (Extended Data Fig. 7c).

On mapping early- and late-responder markers back onto tissue, we found the expected pattern at border regions for SERPINC1 and LGALS3BP, which mirrored AAT levels early on. The late marker DNAJB11 remained unchanged in four of the six samples, indicating that we captured the accumulation event at an early to medium stage (Fig. 3d). However, we detected upregulation of the apoptotic inducer TNFSF10 in the border cells in two samples. Further inspection revealed that the aggregate morphology was markedly different, with a globular phenotype in contrast to amorphous AAT accumulation in the other two samples.

Globular aggregates mark apoptotic cells

Motivated by this observation, we enhanced our DVP workflow to connect morphological information with proteomic data acquisition. We obtained liver resection samples containing thousands of cells with various AAT aggregate morphologies on one slide. After staining and confocal imaging of 3-µm-thick sections of three biological and four technical samples, we segmented cells and transformed the AAT channel signal within cell boundaries into 2,048 features representing AAT morphology using the ConvNeXt convolutional neural network[31]. We projected these representations into a two-dimensional space using uniform manifold approximation and projection (UMAP) and determined 50 equally distributed centre points across the image information layer, from which we selected the 50 closest cells. These were isolated by laser microdissection and measured by MS, resulting in 250 morphology classes representing a total of 12,500 cells (Fig. 4a).

Using UMAP to project the representation of these microdissected cells into a two-dimensional space validated that the convolutional neural network used could indeed stratify cells by aggregate morphologies, with aggregate-devoid cells clustering on one end and globular and amorphous morphologies located at the opposite side and clearly separated from one another (Fig. 4b). We achieved a median proteomic depth of 5,970 proteins from the equivalent of five to ten complete hepatocytes (Extended Data Fig. 8a and Supplementary Table 4). The main drivers of our proteomic data were dynamic changes in keratins and AAT levels on principal components 1 and 2, respectively (Fig. 4c and Extended Data Fig. 8b–d). When grouping samples by proteome into clusters, patient samples were distributed equally across proteomic clusters without apparent genotypic or technical biases (Fig. 4d). As an inverse proof-of-principle, we mapped the proteomic clusters back onto the UMAP image space with clear dimensional separation (Extended Data Fig. 8e). Consistently, samples of one proteome cluster also exhibited the shortest distances to one another on a proteomic UMAP and *t*-distributed stochastic neighbour embedding plot (Extended Data Fig. 8f,g).

To better understand the molecular responses underlying morphology types, we comparatively analysed samples with clear globular versus amorphous aggregates (Fig. 4e). Contrary to expectation, markers that typically follow AAT levels, such as CES2 and ERO1A, were decreased in globular types. Conversely, the apoptotic inducer TNFSF10 and the inflammatory marker C-reactive protein (CRP) were positively enriched, indicating this to be a late-stage phenotype. We then mapped levels of marker proteins back onto the UMAP-derived image space. Intriguingly, ERO1A and TNFSF10 were localized in two distinct cell populations (Fig. 4f and Extended Data Fig. 9a–d). While ERO1A, indicative of an ongoing UPR response, was highly enriched in amorphous aggregate types, TNFSF10 was present mostly in cells with globular aggregates alongside innate immune system activators. In line with this, gene set enrichment analysis (GSEA) further identified processes related to cell death as upregulated in globular types (Extended Data Fig. 9e).

Given a rather linear response rate of CRP across the image UMAP space (Fig. 4f), we then sorted all samples in pseudotime by CRP expression levels. Across all four biological samples, we observed the emergence and disappearance of small corpuscular aggregates despite retained CRP signal. This was followed by a fulminant amorphous aggregation before condensation into globular aggregates as a late-stage feature before cell death and clearance (Fig. 4g).

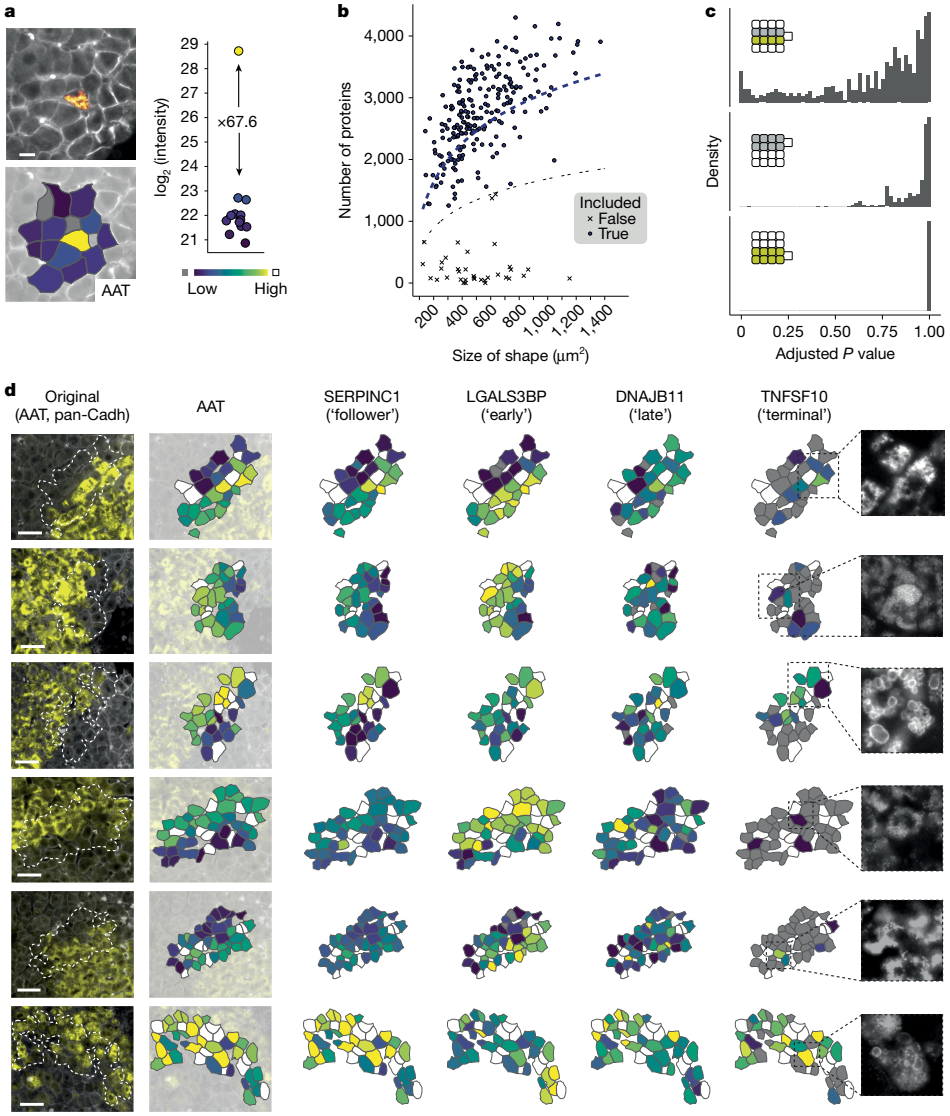


Fig. 3 | Mapping intact tissue at single-cell level. a, Enrichment efficiency of the workflow as shown by isolating adjacent cells from FFPE tissue. Proteome quantification of AAT mapped back onto tissue. Boxplot shows AAT expression enrichment. **b**, Number of proteins detected per single shape across all 259 runs against the area of the microdissected shape. Lower grey dotted line marks inclusion cutoff, upper blue dotted line is a logarithmic fit. **c**, Distribution of *P* values when comparing single cells at a border (top, $n = 107$), direct AAT$^-$ neighbours (middle, $n = 69$) and direct AAT$^+$ neighbours (bottom, $n = 111$; two-sided paired moderated *t*-test after multiple testing correction). **d**, Mapping of proteomic information onto the original microscopic image. Cut-out images show AAT staining only. Grey, protein not quantified; white, shape not captured and measured ($N = 6$, $n = 259$). Scale bars, 50 µm.

In addition to TNFSF10, we identified EGF-like domain-containing protein 7 (EGFL7) as a viable marker of this stage that appeared late in the AATD phenotype. Notably, EGFL7 is also upregulated in hepatocellular carcinoma, and high expression levels are associated with poor prognosis[32]. However, a potential link between globular phenotypes and hepatocellular carcinoma incidence in AATD remains unexplored. This late-stage phenotype was further characterized by a stagnating or even declining UPR in late stages, as evidenced by Calreticulin and ERO1A levels, whereas declining levels of proteins such as UGT2B17 suggest the termination of physiological functions in this hepatocyte subtype (Fig. 4h).

Discussion

We present a pseudotime-resolved proteome of individual hepatocytes undergoing proteotoxic stress due to AAT aggregation. Our findings, derived from FFPE biopsies and resections from patients, provide new insights into the progression and hepatic manifestation in AAT deficiency. Although there are several model systems in the field, including mouse models[33] and patient-derived induced pluripotent stem cells[34], our approach uniquely captures responses to proteotoxic stress directly in patients using human tissue specimens representing the full disease spectrum (stages F1–F4). Notably, our

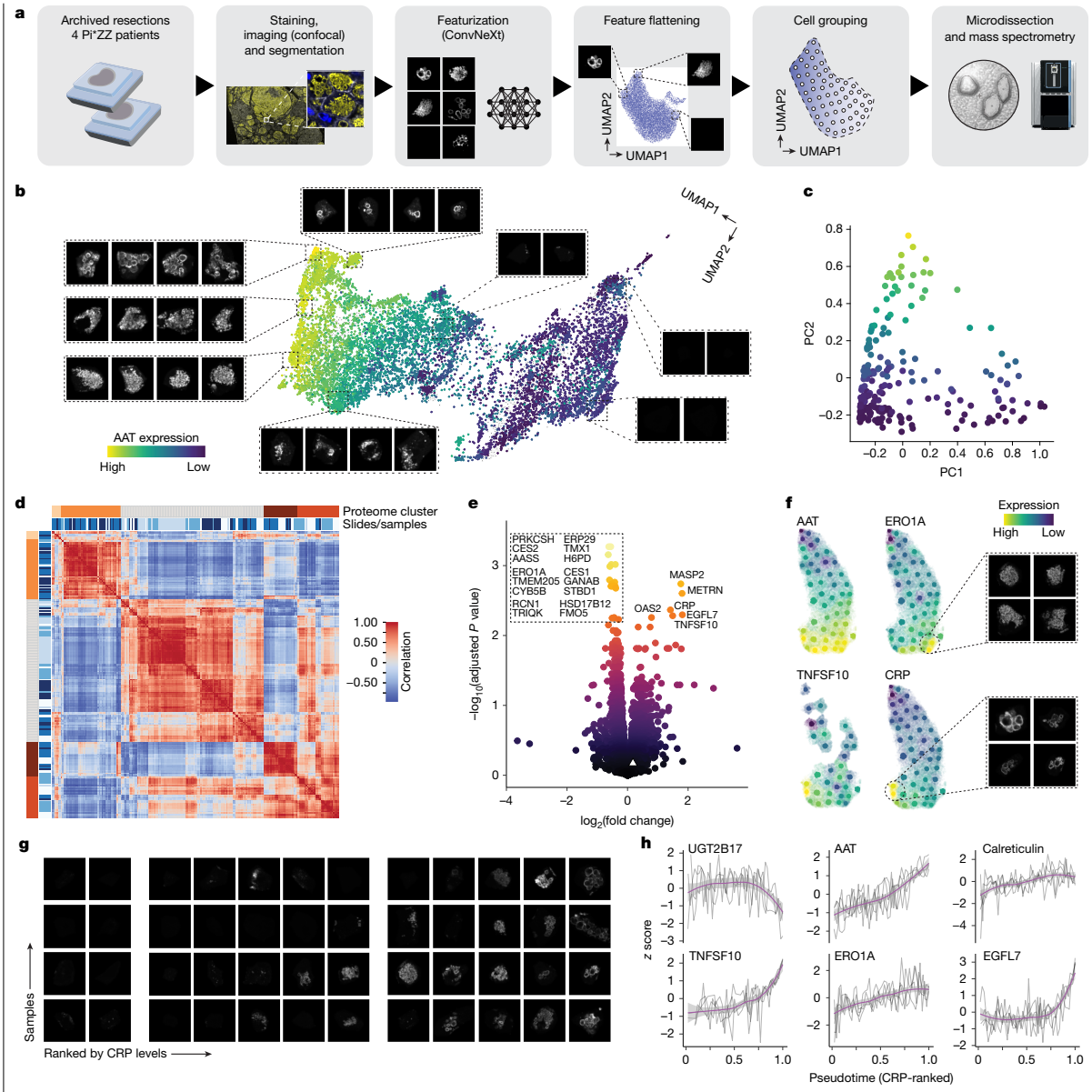


Fig. 4 | Morphology-guided DVP. a, Overview of the pipeline. **b**, Projection of all laser microdissected cells (12,500) and representative AAT images in indicated areas. Colour scheme refers to AAT expression level (proteomic). **c**, Proteomic data of 209 samples (after filtering) reduced by PCA ($n = 4$ tissue sections), coloured by AAT expression level. **d**, Proteomic sample correlation heatmap, indicating proteome clusters based on k-means clustering (five groups chosen manually) and sample slides. **e**, Comparison of proteomes from cells with globular versus amorphous aggregates after selecting for similar AAT levels (white triangle). Up in globular on the right, top hits annotated (paired two-sided moderated t-test after multiple testing correction). **f**, Projection of proteomics data onto image-based UMAP space of one representative sample, with representative images of indicated clusters. **g**, Pseudotime-sorted images of all four biological replicates. Groups mark inflection points of CRP. **h**, Expression levels of indicated proteins in CRP-ranked pseudotime. Each line is one sample, smoothing curve in purple with 95% confidence interval in grey.

data reveal that existing Pi*ZZ models do not accurately recapitulate the UPR, which manifests as a late but fulminant mode of action in our patient-derived samples[1,35]. This discrepancy extends to the globular phenotype, which we now identify as a late-stage cellular feature preceding cell death[16]. Our approach strikingly underlines the power of harnessing patient cohorts and tissues. As many potentially druggable targets and pathways are intrinsically more difficult to validate when appropriate model systems are not in place, this inverts the traditional biomedical discovery cycle. A limitation of this study is the low sample numbers due to limited availability of particularly low-grade fibrotic tissue. This prevents us from further disentangling confounding factors such as alcohol consumption. Nevertheless, the

cellular enrichment by DVP allows the biological phenotype to emerge more clearly, leading to statistically robust and actionable insights even at low sample numbers.

Here we developed a single-cell proteomics approach to generate high-resolution maps of adjacent hepatocytes in intact tissue, leveraging recent advancements in ultra-low-input MS[6,7,36]. Building on our previous work mapping zonation profiles in frozen mouse liver sections at single-cell resolution[5], we now quantify 50% more proteins and apply single-cell DVP (scDVP) to FFPE tissue using the Orbitrap Astral mass spectrometer with a variable window precursor selection scheme. This compatibility with FFPE tissue specimens—the gold standard in diagnostic pathology—expands access to cohorts of virtually any origin, age and size[37], broadening the potential applications of this technology. Spatial transcriptomics has become a powerful tool for spatial analyses in intact FFPE tissue, often approaching single-cell resolution[38]. By contrast, the scDVP approach provides orthogonal biological insights by directly measuring protein abundance with single-cell localization. This is particularly valuable when post-transcriptional regulation and protein accumulation are central to pathology, such as for understanding proteotoxic diseases[38]. Although the scDVP approach is currently limited in throughput compared with transcriptomics, its combination with the herein presented morphology-guided DVP allows efficient sampling of histologically heterogenous material. This could be expanded into morphology-based proteome prediction for large numbers of cells.

Our findings indicate that cells without aggregates are not directly affected or triggered by seeding-like mechanisms from adjacent aggregate-bearing cells. However, the presence of large patches of positive cells implies a propagation mechanism. Given the extensive metabolic perturbations observed, including alterations in fatty acid metabolism and detoxification pathways, AAT aggregate formation in one cell may lead to changes in the metabolic microenvironment, thereby inducing stress and proteostatic imbalance in adjacent cells. This hypothesis aligns with other reports in the AATD field, and similar mechanisms have been proposed in the context of neurodegenerative proteotoxic disorders, where it remains the subject of ongoing debate[39,40].

We present an integration of image featurization and DVP that enables characterization of the entire proteomic and phenotypic lifecycle of stressed hepatocytes in a proteotoxic and fibrogenic liver disease. This methodology establishes a robust framework for dissecting complex cellular processes in situ across a spectrum of proteotoxic diseases. This strategy—an example of digital pathology with quantitative and very deep proteomic readout—yielded exceptionally deep proteomes of 6,000 quantified proteins, sufficient to infer most of the functional proteome of a given cell type. Our datasets are large enough to generate robust models capable of predicting the proteome of a cell solely on the basis of its phenotype. This advancement paves the way for whole-slide proteomics in the future, representing a leap forward in our ability to comprehensively analyse tissue types by MS at exceptional molecular and spatial resolution.

The methods developed here recapitulate known disease progression markers while identifying hundreds of additional dysregulated proteins. The present study is necessarily limited in functional follow-ups, yet these new candidates clearly offer a valuable resource for biological and clinical validation. Of particular clinical relevance, we uncover an early upregulation of the peroxisomal compartment in samples from patients with low-grade liver fibrosis. This response is significantly delayed in high-grade fibrotic samples, suggesting a potential window for therapeutic intervention. Of note, a peroxisomal response is not significantly correlated with fibrotic stages in bulk liver proteomes of patients with alcohol-related liver disease, suggesting that it is specifically important to the AATD pathomechanism[25]. PPAR-α agonists, such as fibrates, which increase peroxisome load in the liver, may be promising candidates for treating patients with late-diagnosed advanced liver fibrosis due to AATD. Given their well-established safety profiles, we suggest that these drugs could be repurposed for AATD, potentially transforming the treatment landscape of this proteotoxic disorder.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-08885-4.

1. Greene, C. M. et al. α1-Antitrypsin deficiency. *Nat. Rev. Dis. Primers* **2**, 16051 (2016).
2. Strnad, P., McElvaney, N. G. & Lomas, D. A. Alpha1-antitrypsin deficiency. *N. Engl. J. Med.* **382**, 1443–1455 (2020).
3. Hipp, M. S., Park, S.-H. & Hartl, F. U. Proteostasis impairment in protein-misfolding and -aggregation diseases. *Trends Cell Biol.* **24**, 506–514 (2014).
4. Mund, A. et al. Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat. Biotechnol.* **40**, 1231–1240 (2022).
5. Rosenberger, F. A. et al. Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nat. Methods* **20**, 1530–1536 (2023).
6. Petrosius, V. et al. Evaluating the capabilities of the Astral mass analyzer for single-cell proteomics. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.06.543943 (2023).
7. Guzman, U. H. et al. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* **42**, 1855–1866 (2024).
8. Guo, T., Steen, J. A. & Mann, M. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature* **638**, 901–911 (2025).
9. Nordmann, T. M. et al. Spatial proteomics identifies JAKi as treatment for a lethal skin disease. *Nature* **635**, 1001–1009 (2024).
10. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
11. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 595–608 (2016).
12. Goedert, M., Jakes, R. & Spillantini, M. G. The synucleinopathies: twenty years on. *J. Parkinsons Dis.* **7**, S51–S69 (2017).
13. Lomas, D. A., Evans, D. L., Finch, J. T. & Carrell, R. W. The mechanism of Z alpha 1-antitrypsin accumulation in the liver. *Nature* **357**, 605–607 (1992).
14. Brantly, M., Nukiwa, T. & Crystal, R. G. Molecular basis of alpha-1-antitrypsin deficiency. *Am. J. Med.* **84**, 13–31 (1988).
15. Clark, V. C. et al. Clinical and histologic features of adults with alpha-1 antitrypsin deficiency in a non-cirrhotic cohort. *J. Hepatol.* **69**, 1357–1364 (2018).
16. Lindblad, D., Blomenkamp, K. & Teckman, J. Alpha-1-antitrypsin mutant Z protein content in individual hepatocytes correlates with cell death in a mouse model. *Hepatology* **46**, 1228–1235 (2007).
17. Rudnick, D. A. et al. Analyses of hepatocellular proliferation in a mouse model of alpha-1-antitrypsin deficiency. *Hepatology* **39**, 1048–1055 (2004).
18. Chambers, J. E. et al. Z-α1-antitrypsin polymers impose molecular filtration in the endoplasmic reticulum after undergoing phase transition to a solid state. *Sci. Adv.* **8**, eabm2094 (2022).
19. Segeritz, C.-P. et al. hiPSC hepatocyte model demonstrates the role of unfolded protein response and inflammatory networks in α1-antitrypsin deficiency. *J. Hepatol.* **69**, 851–860 (2018).
20. Fromme, M., Schneider, C. V., Trautwein, C., Brunetti-Pierri, N. & Strnad, P. Alpha-1 antitrypsin deficiency: a re-surfacing adult liver disorder. *J. Hepatol.* **76**, 946–958 (2022).
21. Zhang, Y. et al. LMAN1-MCFD2 complex is a cargo receptor for the ER-Golgi transport of α1-antitrypsin. *Biochem. J.* **479**, 839–855 (2022).
22. Schmidt, B. Z. & Perlmutter, D. H. Grp78, Grp94, and Grp170 interact with alpha1-antitrypsin mutants that are retained in the endoplasmic reticulum. *Am. J. Physiol. Gastrointest. Liver Physiol.* **289**, G444–G455 (2005).
23. Werder, R. B. et al. Adenine base editing reduces misfolded protein accumulation and toxicity in alpha-1 antitrypsin deficient patient iPSC-hepatocytes. *Mol. Ther.* **29**, 3219–3229 (2021).
24. Spivak, I. et al. Alpha-1 antitrypsin inclusions sequester GRP78 in a bile acid–inducible manner. *Liver Int.* **45**, e16207 (2025).
25. Niu, L. et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28**, 1277–1287 (2022).
26. Cho, S.-H. et al. Lgals3bp suppresses colon inflammation and tumorigenesis through the downregulation of TAK1-NF-κB signaling. *Cell Death Discov.* **7**, 65 (2021).
27. Khodayari, N. et al. Characterization of hepatic inflammatory changes in a C57BL/6J mouse model of alpha1-antitrypsin deficiency. *Am. J. Physiol. Gastrointest. Liver Physiol.* **323**, G594–G608 (2022).
28. Porat-Shliom, N. Compartmentalization, cooperation, and communication: the 3Cs of hepatocyte zonation. *Curr. Opin. Cell Biol.* **86**, 102292 (2024).
29. Piccolo, P. et al. Down-regulation of hepatocyte nuclear factor-4α and defective zonation in livers expressing mutant Z α1-antitrypsin. *Hepatology* **66**, 124 (2017).
30. Crowther, D. C. et al. Practical genetics: alpha-1-antitrypsin deficiency and the serpinopathies. *Eur. J. Hum. Genet.* **12**, 167–172 (2004).
31. Liu, Z. et al. A ConvNet for the 2020s. Preprint at https://arxiv.org/abs/2201.03545v2 (2022).
32. Yang, C. et al. Increased expression of epidermal growth factor-like domain-containing protein 7 is predictive of poor prognosis in patients with hepatocellular carcinoma. *J. Cancer Res. Ther.* **14**, 867–872 (2018).
33. Carlson, J. A. et al. Accumulation of PiZ alpha 1-antitrypsin causes liver damage in transgenic mice. *J. Clin. Invest.* **83**, 1183–1190 (1989).

34. Yusa, K. et al. Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**, 391–394 (2011).
35. Hidvegi, T., Schmidt, B. Z., Hale, P. & Perlmutter, D. H. Accumulation of mutant alpha1-antitrypsin Z in the endoplasmic reticulum activates caspases-4 and -12, NFkappaB, and BAP31 but not the unfolded protein response. *J. Biol. Chem.* **280**, 39002–39015 (2005).
36. Rosenberger, F. A., Thielert, M. & Mann, M. Making single-cell proteomics biologically relevant. *Nat. Methods* **20**, 320–323 (2023).
37. Coscia, F. et al. A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* **251**, 100–112 (2020).
38. Fan, R. Integrative spatial protein profiling with multi-omics. *Nat. Methods* **21**, 2223–2225 (2024).
39. Henrich, M. T. et al. Determinants of seeding and spreading of α-synuclein pathology in the brain. *Sci. Adv.* **6**, eabc2487 (2020).
40. Bassil, F. et al. Amyloid-beta (Aβ) plaques promote seeding and spreading of alpha-synuclein and tau in a mouse model of Lewy body disorders with Aβ pathology. *Neuron* **105**, 260–275.e6 (2020).

Article

Methods

Clinical cohorts and sample preparation

Patient biopsies and explant samples were obtained at two different sites, Odense University Hospital (OUH) and Aachen RWTH University Hospital (UKA). The sample origin is indicated in Supplementary Table 1. Following ethical guidelines, the clinical data provided here are deidentified by reporting only sample type, fibrosis score and site of origin.

OUH patient recruitment. Patients were recruited through the Danish patient organization (Alfa-1 Denmark) and clinical departments for liver and lung diseases as part of a cohort study. The cohort was designed to investigate liver health among nonpregnant adults (minimum age 18 years) diagnosed with AATD of any genotype and carrier status. This specific study includes 16 people diagnosed with Pi*ZZ who consented to undergo the procedure. The study was approved by the Danish Ethical Committee (S-20160187), and participants gave informed consent before enrolment. Participants without a history of liver transplant or decompensated cirrhosis were offered a percutaneous liver biopsy. The patients underwent liver core needle biopsies at OUH between 2017 and 2021. Liver core needle biopsies were taken during this period, stored in 4% formalin and embedded in paraffin. For the assessment of fibrosis stage, FFPE blocks were cut on a microtome into 3-µm-thick sections and mounted on FLEX IHC slides (Dako). Tissue sections were deparaffinized with xylene, rehydrated in serial dilutions of ethanol and stained with Sirius Red. A certified hepatopathologist (S.D.) assessed the Kleiner fibrosis stage (0–4) according to the Pathology Committee of the NASH Clinical Research Network (NAS-CRN).

UKA patient recruitment. The recruitment of patients is described in detail in ref. 41. Of this cohort, the present study includes 19 people diagnosed with Pi*ZZ, of whom 14 underwent liver core needle biopsies owing to medical indication and five received a liver transplant because of end-stage liver disease. One patient's sample was later removed owing to its outlier position on the proteome PCA (Supplementary Table 1). Samples were stored in 4% formalin and embedded in paraffin. Fibrosis stage was assessed after trichrome staining of 5-µm-thick sections by a certified hepatopathologist. Blocks were stored at room temperature. Ethical approval was provided by the institutional review board of Aachen University (EK 173/15). All participants provided written informed consent and were treated following the ethical guidelines of the Helsinki Declaration (Hong Kong Amendment) as well as Good Clinical Practice (European guidelines).

Staining

Polyethylene naphthalate membrane slides (2 µm; MicroDissect GmbH) were exposed to ultraviolet light (254 nm) for 1 h and then coated with Vectabond (Vector Laboratories; catalogue no. SP-1800-7) according to the manufacturer's protocol. FFPE sections (3-µm-thick, DVP, ML; 10-µm-thick, scDVP) were mounted onto these slides and dried at 37 °C overnight. Slides were stored at 4 °C until further processing, upon which slides were baked at 55 °C for 40 min and then deparaffinized and rehydrated (xylene 2 × 2 min, 100% ethanol 2 × 1 min, 90% ethanol 2 × 1 min, 75% ethanol 2 × 1 min, 30% ethanol 2 × 1 min, ddH_2O 2 × 1 min). Slides were transferred to prewarmed glycerol-supplemented antigen retrieval buffer (DAKO pH 9 S2367 + 10% glycerol) at 88 °C for 20 min, followed by a 20-min cooldown at room temperature (22 °C). After washing in water, sections were blocked with 5% bovine serum albumin (BSA) in PBS for 1 h, followed by an overnight incubation with primary antibodies in 1% BSA/PBS at 4 °C in a humid staining chamber (1:200 mouse IgG1 monoclonal AAT 2C1, Hycult catalogue no. HM2289; 1:200 rabbit recombinant anti-pan-Cadh (EPR1792Y), Abcam catalogue no. ab51034). After three washes in PBS for 2 min each, secondary antibodies (1:400 goat anti-mouse IgG1, Invitrogen catalogue no. A21127; 1:400 goat anti-rabbit AF647, Invitrogen catalogue no. A21245) in 1% BSA/PBS were applied for 90 min, followed by two 2-min washes in PBS, 15 min in SYTOX Green (1:40,000 in PBS, Invitrogen catalogue no. S7020), and three final 2-min washes in PBS. Excess liquid was removed and samples were coverslipped using SlowFade Diamond Antifade Mountant (Invitrogen, catalogue no. S36963).

Imaging

Widefield imaging. For DVP and scDVP experiments (Figs. 1–3), sections were imaged using a Zeiss Axioscan 7. For all excitation wavelengths (493 nm, 577 nm and 653 nm), 50% light source intensity was used. The illumination time was specified on one section and applied to all consecutive samples within one experimental group. Three *z*-stacks at an interval of 2 µm were recorded with a Plan-Apochromat ×20, 0.8 numerical aperture M27 objective and an Axiocam 712 camera at 14-bit, with a binning of 1 and a tile overlap of 10%, resulting in a scaling of 0.173 µm × 0.173 µm. Multiscene images were then split into single scenes, *z*-stacks combined into a single plane using extended depth of focus (variance method, standard settings) and stitched on the pan-Cadh channel using the proprietary Zeiss Zen Imaging software.

Confocal imaging. For experiments with downstream machine learning applications (Fig. 4), sections were imaged on a Perkin Elmer OperaPhenix high-content microscope, controlled with Harmony v.4.9 software, at ×40 magnification and 0.75 numerical aperture, with a binning of 1 and a per tile overlap of 10%. Only one *z*-plane was recorded, which was specified manually for each slide and channel. The three channels were imaged consecutively after deactivation of simultaneous recording to avoid any leakage between channels.

Cell selection with Biological Image Analysis Software

Images were imported as .czi files into the Biological Image Analysis Software (BIAS) using the packaged import tool[4]. Within BIAS, images were then retiled to 1,024 × 1,024 pixels with an overlap of 10%, and empty tiles were excluded from further analyses. Outlines of all cells per biopsy were identified in an unbiased way by using Cellpose v.2.0 with the default cyto2 model based on anti-pan-Cadh stains[42]. Masks were imported into BIAS, and duplicates, as well as cells touching the borders of a tile (0.1% on each side), were removed. Further filtering was applied to retain cells with a minimum size of 3,000 pixels, enriching for the hepatocyte population. For classification based on low, medium and high aggregate load, all cells per biopsy or explant tissue were divided into five classes using a multilayer perceptron with the following parameters: weight scale 0.01; momentum 0.01; maximum iterations 10,000; epsilon 0.0005 and five neurons in the hidden layer. Classification was based on the AAT maximum, median and mean intensity within the cell outline mask, involving no human intervention. The low class was attributed to the cells with the lowest normalized mean intensity, medium to the third highest and high to the highest normalized mean intensity; the other two intermediate classes were dropped. Reference points were selected on the basis of prominent nuclear and histological features; 100 cells were picked randomly for excision.

For single shape experiments, six characteristic low fibrosis samples (all F1) and regions were selected that presented with a clear border-like phenotype (that is, a row of AAT^+ cells in direct neighbourhood to AAT^- cells) or with single AAT^+ cells surrounded by AAT^- cells. The cells were selected manually in BIAS, starting from the innermost cell and moving spiral-like to the outermost cell, thus avoiding cross-contamination of consecutively cut material.

Single-cell image generation

Images were flat-field corrected during image acquisition using the Perkin Elmer Harmony software (v.4.9). Stitching of the flat-field corrected image tiles was performed using the Python library scPortrait (https://github.com/MannLabs/scPortrait). The stitched tile positions

were calculated using the anti-pan-Cadh stains imaged in the Alexa647 channel as a reference and then transferred to the other image channels. During stitching, the tile overlap was set to 0.1, the filter sigma parameter to 1 and the max shift parameter to 50.

The stitched images were then further processed in scPortrait. Cell outlines were identified on the basis of the seven times downsampled anti-pan-Cadh stains using Cellpose v.2.0 with the pretrained 'cyto' model[42]. Segmentation was performed in a tiled mode with a 100-pixel overlap. After resolving the cell outlines from overlapping regions, the resulting segmentation mask was upscaled to the original input dimensions during which the edges of the masks were smoothened by applying an erosion and dilation operation with a kernel size of 7.

Then, the generated segmentation mask was used to extract single-cell image datasets with a size of 280 pixels × 280 pixels. During extraction, the same single-cell image masks are used to obtain the pixel information from each channel for each cell. The resulting single-cell images were then rescaled to the [0,1] range while preserving relative signal intensities. The resulting single-cell image datasets were filtered to contain only cells from within manually annotated regions in the tissue section containing hepatocytes but not fibrotic tissue.

Cell selection with the convolutional neural network

The filtered single-cell image datasets produced by scPortrait were further filtered to remove any cells that fell outside the 5–97.5% size percentile. Representations of the remaining cells were generated by featurization using the natural image-pretrained ConvNext model[31]. For this, the single-cell images depicting the Alpha-1 channel were rescaled to the expected image dimensions of *N* pixels × *N* pixels and triplicated to generate a pseudo-rgb image. Inference was then performed using the huggingface transformers package v.4.26 (ref. 43).

The resulting 2,048 image features were projected into a two-dimensional space using the UMAP algorithm[44]. The UMAP dimensions were calculated on the basis of the first 50 principal components and the 15 nearest neighbours. Using the spectral clustering algorithm from scikit-learn[45], the resulting UMAP space was split into 50 clusters. The geometric centre of each cluster was calculated and the 50 cells with the smallest Euclidean distance to the cluster centre were selected for laser microdissection.

Contour outlines of the selected cells were generated in scPortrait using the py-lmd package[46], whereby the cell outlines were dilated with a kernel size of 3 and a smoothing filter of 25 was applied. Furthermore, the number of points defining each shape were compressed by a factor of 30 to improve laser microdissection cutting performance. The cutting path, that is, which cell is cut after one another, was optimized using the Hilbert algorithm (https://github.com/galtay/hilbertcurve).

Laser microdissection

After aligning the reference points, contour outlines were imported, and shapes were cut using the LMD7 (Leica) laser microdissection system in a semi-automated mode with the following settings: power 45; aperture 1; speed 40; middle pulse count 1; final pulse 0; head current 42–50%; pulse frequency 2,982 and offset 190. The microscope was operated with the LMD beta v.10 software, calibrated for the gravitational stage shift into 384-well plates (Eppendorf, catalogue no. 0030129547), leaving the outermost rows and columns empty. To prevent sorting errors, a 'wind shield' plate was placed on top of the sample stage. Plates were then sealed, centrifuged at 1,000*g* for 5 min, and subsequently frozen at −20 °C for further processing.

Peptide preparation and Evotip loading

Peptides were prepared as described previously using a BRAVO pipetting robot (Agilent)[47]. Briefly, 384-well plates were thawed, and shapes (both combined and individual) were rinsed from the walls into the bottom of the well with 28 µl of 100% acetonitrile (ACN). The wells were dried completely in a SpeedVac at 45 °C, followed by the addition of 6 µl of 60 mM triethylammonium bicarbonate (Supelco, catalogue no. 18597) (pH 8.5) supplemented with 0.013% n-dodecyl-beta-D-maltoside (Sigma-Aldrich, catalogue no. D5172). Plates were sealed and incubated at 95 °C for 1 h. After adjusting to 10% ACN, samples were incubated again at 75 °C for 1 h. Subsequently, 6 ng and 4 ng of trypsin and Lys-C protease, respectively, in 1 µl of 60 mM triethylammonium bicarbonate buffer were added to each sample, and proteins were digested for 16 h at 37 °C. The reaction was quenched by adding trifluoroacetic acid to a final concentration of 1%. Peptide samples were then frozen at −20 °C.

For loading, new Evotips were first soaked in 1-propanol for 1 min, then rinsed twice with 50 µl of buffer B (ACN with 0.1% formic acid). After another 1-propanol soaking step for 3 min, the tips were equilibrated with two washes of 50 µl buffer A (0.1% formic acid). Samples were loaded into 70 µl of preloaded buffer A. Following one additional buffer A wash, the peptide-containing C18 disk was overlaid with 150 µl buffer A and centrifuged briefly through the disk. All centrifugation steps were performed at 700*g* for 1 min. The final tips were stored in buffer A for a maximum of 4 days before liquid chromatography (LC)-MS.

LC-MS data acquisition

The peptide samples were analysed using an Evosep One LC system (Evosep) coupled to an Orbitrap Astral mass spectrometer (Thermo Fisher Scientific). Peptides were eluted from the Evotips with up to 35% ACN and separated using an Evosep low-flow 'Whisper' gradient for DVP samples, or an experimental Evosep 'Whisper Zoom' gradient for single shapes and DVP-machine learning samples, with a throughput of 40 samples per day on an Aurora Elite TS column of 15-cm length, 75-µm-internal diameter, packed with 1.7 µm C18 beads (IonOpticks). The column temperature was maintained at 50 °C using a column heater (IonOpticks).

The Orbitrap Astral mass spectrometer was equipped with a FAIMS Pro interface and an EASY-Spray source (both Thermo Fisher Scientific). A FAIMS compensation voltage of −40 V and a total carrier gas flow of 3.5 l min^{-1} were used. An electrospray voltage of 1,900 V was applied for ionization, and the radio frequency level was set to 40. Orbitrap MS1 spectra were acquired from 380 to 980 m/z at a resolution of 240,000 (at m/z 200) with a normalized automated gain control (AGC) target of 500% and a maximum injection time of 100 ms.

For the Astral MS/MS scans in data-independent acquisition (DIA) mode, we determined the optimal methods experimentally across the precursor selection range of 380–980 m/z: (1) for DVP samples, a window width of 5 Th, a maximum injection time of 10 ms and a normalized AGC target of 800% were used. (2) For DVP-machine learning samples, a window width of 6 Th, a maximum injection time of 13 ms and a normalized AGC target of 500% were applied. (3) For single shapes and other DIA scans, the window width was optimized on the basis of precursor density across the selection range of 380–980 m/z. A total of 45 variable-width DIA windows (Supplementary Table 3) were acquired with a maximum injection time of 28 ms and an AGC target of 800%. The isolated ions were fragmented using higher-energy collisional dissociation with 25% normalized collision energy. Detailed method descriptions are provided in a default format with each supplementary data table.

Spectral searches and normalization

The raw files were searched together with match-between run-in library-free mode within each experimental group with DIA-NN v.1.8.1 (ref. 48). A FASTA file containing only canonical sequences was obtained from Uniprot (20,404 entries, downloaded on 2 January 2023), and the disease-causing amino acid was changed manually (E342K). We allowed a missed cleavage rate of up to 1, and set mass accuracy to 8, MS1 accuracy to 4 and the scan window to 6. Proteins were inferred on the basis of genes, and the neural network classifier was set to 'single-pass mode'. For DVP and DVP-machine learning samples, precursor intensities in

the 'report.tsv' file were then normalized using the directLFQ GUI at standard settings including a minimum number of non-NaN ion intensities required to derive a protein intensity of 1 (ref. 49). The single shape data was additionally median-normalized to a set of proteins quantified across all samples (451 proteins quantified in 100% of included samples; Supplementary Table 3), thereby correcting for the dependence of protein numbers on shape size[5].

Data analysis and statistics

Data were analysed using R v.4.4.1. The directLFQ output file 'pg_matrix.tsv' was used for all subsequent data analysis, including the reported protein counts. Samples were included if the number of protein groups exceeded (1) the mean − 1.5 s.d. for DVP, resulting in 5.9% (6 of 102) dropouts; (2) the mean − 0.5 s.d. for DVP-machine learning samples; (3) a fitted logarithmic curve − 1.5 interquartile ranges for scDVP, taking the relation between size and proteomic depth into account, resulting in 15.4% (40 of 259) dropouts. The lower cutoffs were selected after manual inspection of the data distribution. Although some samples were collected in technical duplicates per patient biopsy, only the first replicate was used for statistical analyses and all reported measurements were taken from distinct samples. Coefficients of variation were calculated on nontransformed intensity values. For principal component analysis (PCA), the R package PCAtools v.2.16.0 was used on a complete data matrix, removing the lower 10% of variables based on variance. Statistical analyses were performed on proteins with at least 30% data completeness across samples, assuming normality using the limma package v.3.60.3 with two-sided moderated *t*-tests and 'fdr' as a multiple testing correction method. A per patient statistical pairing was applied for DVP and single shape experiments. Intensity and fold changes are reported as $\log_2$-transformed values unless indicated otherwise. GSEA was conducted using WebGestalt 2024 against the indicated databases, with an FDR of <0.05 considered significant[50]. Interaction networks were calculated with STRING database at standard settings[51]. Plasma proteins were retrieved from the Human Protein Atlas resource section with the search term 'sa_location:Secreted to blood AND tissue_category_rna:liver;Tissue enriched'[52]. The timing of responses ranked by the absolute difference between B values of limma's moderated *t*-test comparing three AAT load groups: low to moderate, and moderate to high. Only proteins with more than 70% data completeness and significance (FDR < 0.05) in either or both comparisons were considered. Differential pathway expression across fibrosis stages was calculated by fitting a linear model through $\log_2$-transformed intensity values of individual proteins in samples with $\log_2$(AAT)-intensity <25, and the slopes of proteins in a particular pathway were compared between F1 and F4 samples by a two-sided Wilcoxon rank test without assumption of normality. Indicated *P* values are corrected for multiple testing using the 'fdr' method. Spatial data was mapped using the 'simple features' package. Binned expression presented in supplementary tables was constructed by grouping AAT or CRP expression into ten equidistant bins and on median expression of proteins across samples in each bin.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium through the PRIDE[53] partner repository with the dataset identifier PXD054440. Imaging data of explant and morphological clusters have been deposited to BioStudies[54] with the identifier S-BIAD1523.

Code availability

The R and Python code used in this study are documented at https://github.com/MannLabs/Proteotoxicity with a readily deployable script to generate most of the figure panels.

41. Schneider, C. V. et al. Liver Phenotypes of European Adults Heterozygous or Homozygous for Pi∗Z Variant of AAT (Pi∗MZ vs Pi∗ZZ genotype) and Noncarriers. *Gastroenterology* **159**, 534–548.e11 (2020).
42. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
43. Wolf, T. et al. HuggingFace's transformers: state-of-the-art natural language processing. Preprint at https://arxiv.org/abs/1910.03771v5 (2020).
44. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426v3 (2020).
45. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
46. Schmacke, N. A. et al. SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.01.542416 (2023).
47. Thielert, M., Weiss, C. A. M., Mann, M. & Rosenberger, F. A. in *Mass Spectrometry Based Single Cell Proteomics* (eds. Vegvari, A., Teppo, J. & Zubarev, R. A.) 97–113 (Springer, 2024).
48. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
49. Ammar, C., Schessner, J. P., Willems, S., Michaelis, A. C. & Mann, M. Accurate label-free quantification by directLFQ to compare unlimited numbers of proteomes. *Mol. Cell. Proteomics* **22**, 100581 (2023).
50. Elizarraras, J. M. et al. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res.* **52**, W415–W421 (2024).
51. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
52. Uhlén, M. et al. The human secretome. *Sci. Signal.* **12**, eaaz0274 (2019).
53. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
54. Sarkans, U. et al. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2018).

Acknowledgements We thank our colleagues at the Department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry as well as our colleagues at the Centre for Proteome Research in Copenhagen for their input and support. We are particularly grateful for the technical assistance of D. Wischnewski (MPIB), for great scientific discussions with A. Wilson and J. Kaserman (Boston University School of Medicine) and for valuable input from T. Nordmann (MPIB), M. Thielert (MPIB), V. Brennsteiner (MPIB), L. Grauvogel (MPIB), A. Sinha (MPIB), K. Madden (MPIB) and S. Haber (UK Aachen). We thank the Computing Centre and the Imaging Facility of the MPI of Biochemistry for their support and resources. F.A.R. is an EMBO postdoctoral fellow (ALTF 399-2021). S.C.M. is a PhD fellow of the Boehringer Ingelheim Fonds. K.H.T. received a travel grant from the OUH Internationalisation Fund. This study has been supported by the Horizon-2020 under the MICROB-PREDICT programme (M.M., A.K., no. 825694); by the Max Planck Society for Advancement of Science (M.M.); by a grant from the Alpha-1 Foundation (F.A.R.) and Alfa-1 Liver Study (A.K.); by the Deutsche Forschungsgemeinschaft DFG through SFB 1382 (P.S., ID 403224013); P.S. holds a Heisenberg professorship (STR1095/6-1); P.B. is supported by the German Research Foundation (DFG, Project IDs 322900939 and 445703531), European Research Council (ERC Consolidator grant no. 101001791) and the Federal Ministry of Education and Research (BMBF, STOP-FSGS-01GM2202C).

Author contributions Conceptualization: F.A.R., K.H.T., S.C.M., P.S. and M.M. Project team leads: S.C.M., K.H.T. and S.S. Methodology: F.A.R., S.C.M. and S.S. Software: S.C.M., M.L. and N.A.S. Validation: F.A.R., C.A.M.W., M.O., M.W. and M.Z. Formal analysis: F.A.R., S.C.M. and M.L. Investigation: F.A.R., S.C.M., K.H.T., S.S., M.L., C.A.M.W., M.W., A.M. and M.Z. Resources: K.H.T., M.F., S.D., P.B., O.F., S.F., A.K., P.S. and M.M. Data curation: F.A.R. and S.C.M. Writing—original draft: F.A.R. and M.M. Writing—review and editing: all authors. Visualization: F.A.R., S.C.M. and M.L. Supervision: F.A.R., P.S. and M.M.

Funding Open access funding provided by Max Planck Society.

Competing interests M.M. is an indirect investor in Evosep. A patent for treatment of conditions related to α1-antitrypsin deficiency with PPARα agonists has been filed with the European Patent Office (application number EP24205578.8). The other authors declare no competing interests.

Additional information

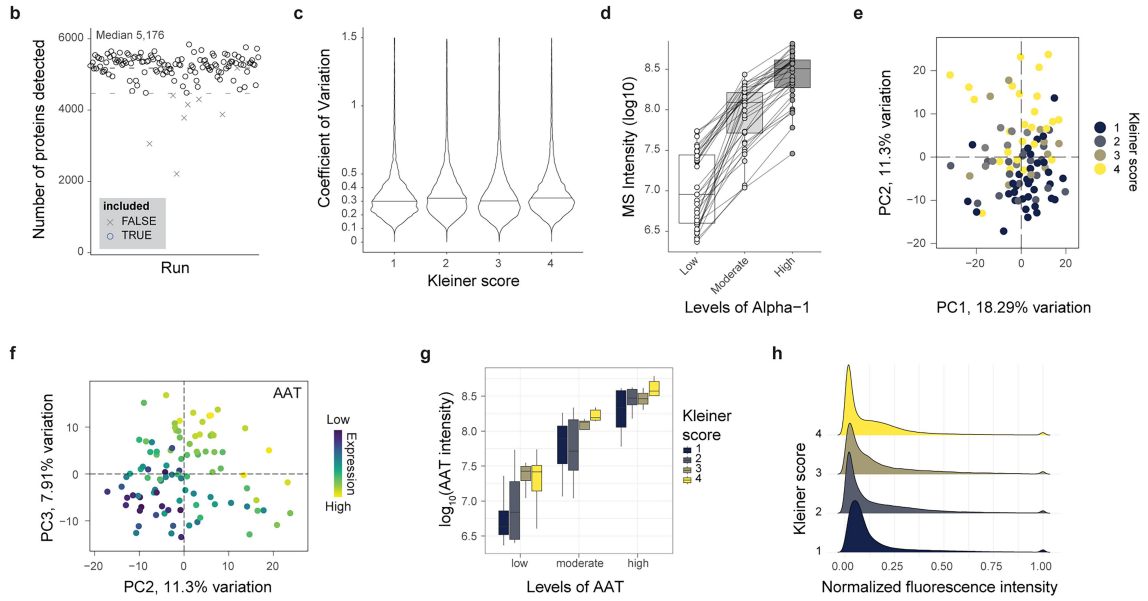
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-08885-4.

Correspondence and requests for materials should be addressed to Florian A. Rosenberger or Matthias Mann.

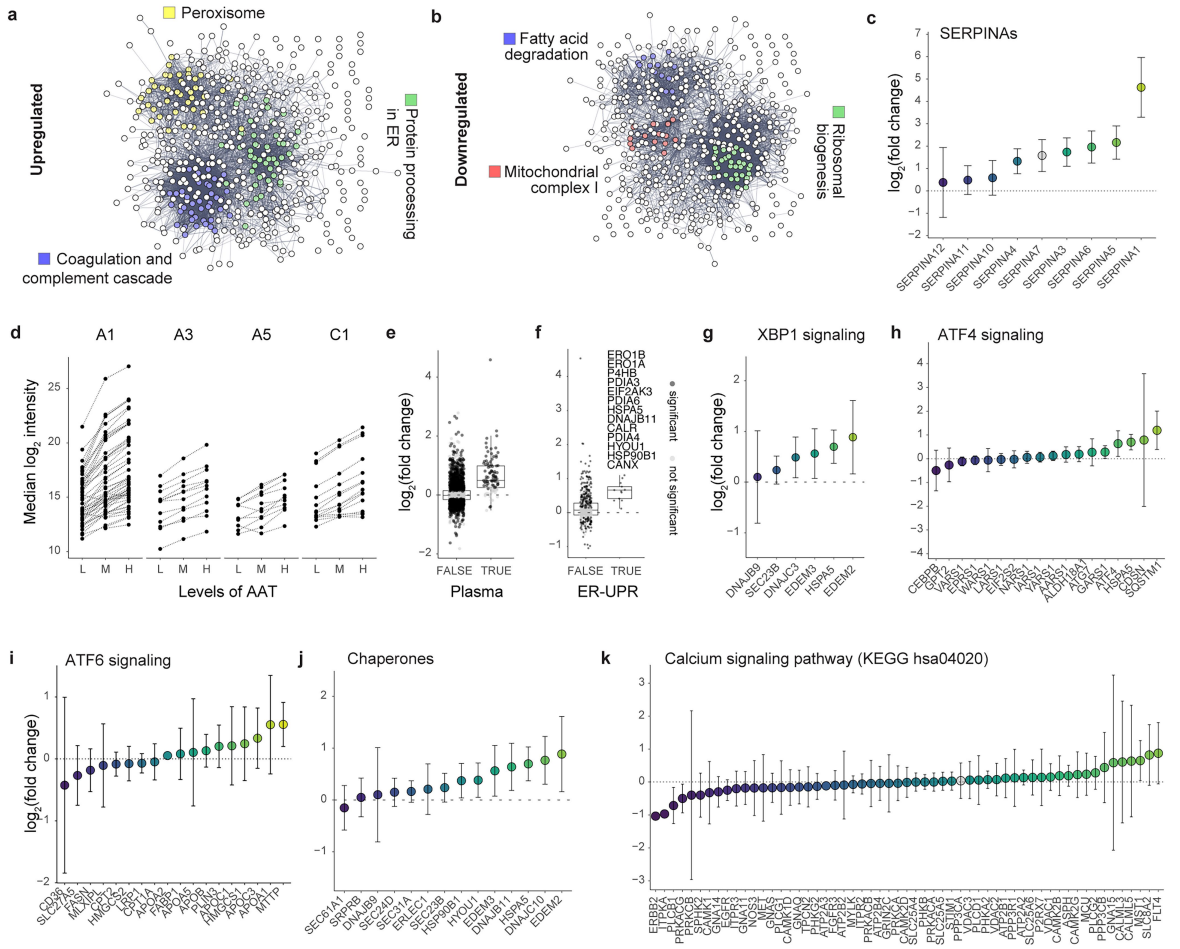
Peer review information *Nature* thanks David Lomas, Tiannan Guo and Stefan Marciniak for their contribution to the peer review of this work. Peer reviewer reports are available.

a

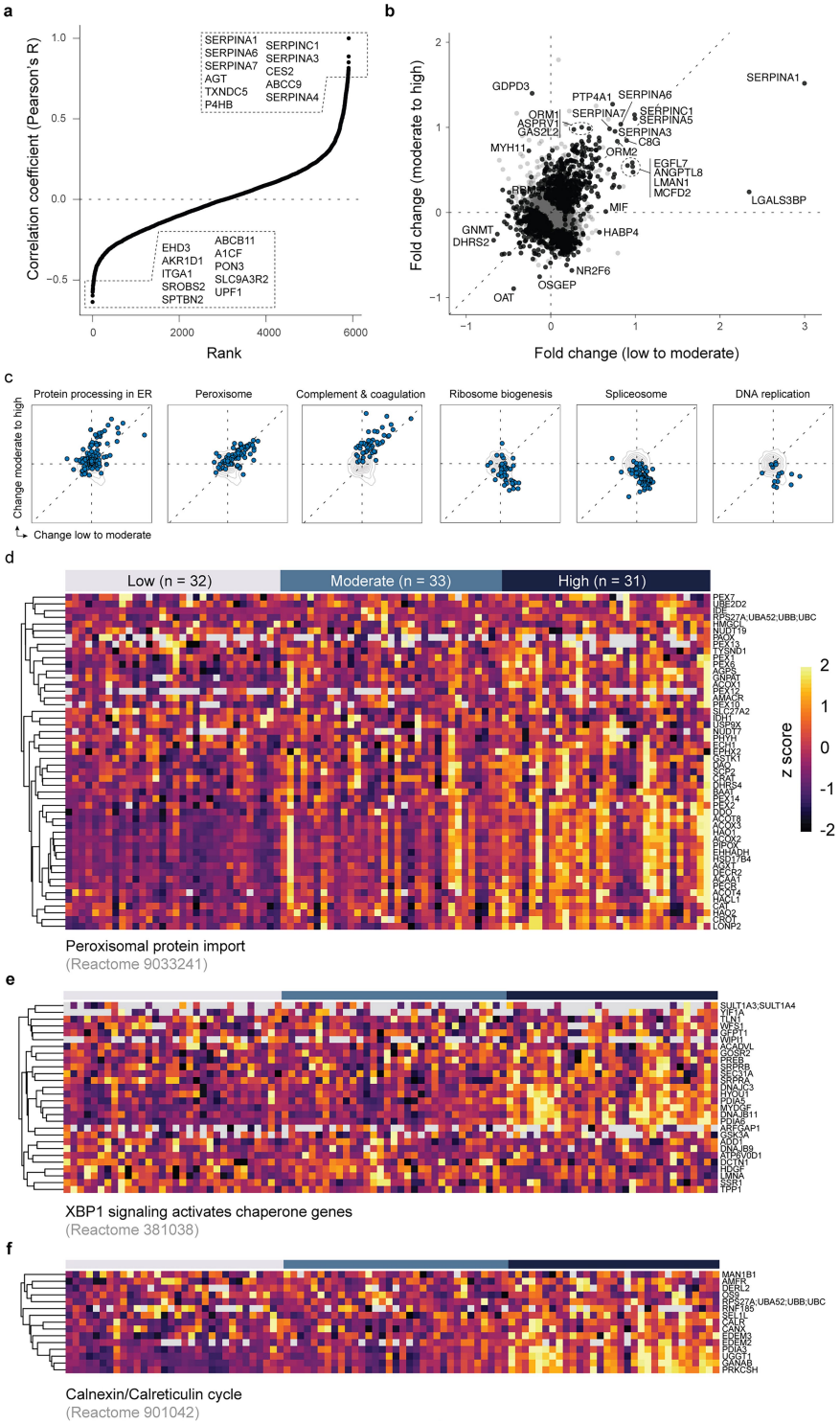
Kleiner score	F1	F2	F3	F4	Resections	unknown
Genotype		Pi*ZZ (100%)				
Number	11	7	3	5	5	3
Female sex	45.5% (5/11)	0% (0/7)	33.3% (1/3)	60.0% (3/5)	100% (3/3, 2 NA)	33.3% (1/3)
Age at biopsy (y)	58.0 (±11.1)	54.9 (±11.0)	64.0 (±10.9)	56.5 (±4.5)	56.4 (±7.5)	63.3 (±7.6)
Alcohol	18.1% (2/11)	33.3% (2/6, 1 NA)	0% (0/3)	50.0% (2/4, 1 NA)	0% (0/4, 1 NA)	33.3% (1/3)
Diabetes		3.0% (1/33, 1 NA)				
BMI	24.7 (±2.92)	25.6 (±3.1)	24.4 (±2.99)	25.5 (±1.88)	28.0 (±3.6)	31.3 (±8.7)



Extended Data Fig. 1 | Quality control of Deep Visual Proteomics data. **a**, Summary of clinical metadata shown as patient numbers or percentages (absolute numbers in brackets). Values reported as mean ± SD. **b**, Number of proteins detected across all runs before excluding technical replicates (n = 134). Upper dotted line: median number of protein groups. Lower dotted line: median − 1.5 SD. Excluded samples are marked with crosses. **c**, Coefficient of variation across fibrosis stages. **d**, MS intensity of alpha-1 antitrypsin in the three microdissected cell classes (N = 34 patients, n = 96 samples). **e**, Principal component analysis showing components 1 and 2 colored by fibrosis stage, and **f**, components 2 and 3 colored by alpha-1 antitrypsin level. Each dot represents one sample (n = 96). **g**, Alpha-1 antitrypsin levels by fibrosis stage across the three microdissected cell classes (N = 31 patients, n = 88 samples with known fibrosis status). **h**, AAT fluorescence intensity distribution across cells from biopsies with Kleiner scores (n = 31 biopsies, 2,967,275 cells total). Values were normalized to a 0–1 range after removing outliers (below 1st and above 99th percentile). Box plots show first and third quartiles (box), median (thick line), and whiskers (±1.5 interquartile range).

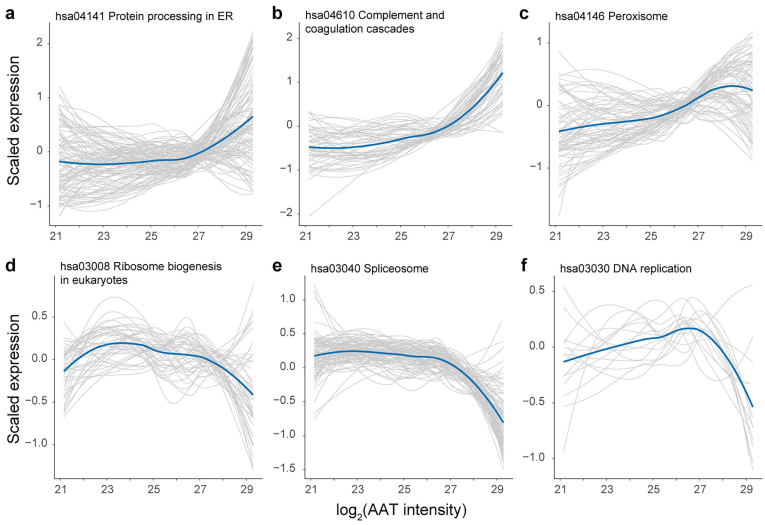


Extended Data Fig. 2 | Proteomics responses to proteotoxic stress. **a**, STRING interaction network of significantly upregulated proteins (FDR < 0.05) and **b**, downregulated proteins in cells (see Fig. 1c). **c**, Changes in SERPINA protein family members relative to baseline hepatocyte group. **d**, Changes in SERPIN protein family precursors across three hepatocyte classes. Lines connect the same precursor (defined as peptide by charge state). **e**, Changes in proteins targeted for plasma secretion relative to baseline hepatocyte group. Plasma protein dataset obtained from Human Protein Atlas using query "sa_location: Secreted to blood AND tissue_category_rna:liver;Tissue enriched" ('FALSE' includes 5806 protein, 'TRUE' includes 100 proteins). **f**, Changes in ER proteins (annotated as such in Uniprot, n = 677) relative to baseline hepatocyte group with a manually curated subset of ER-UPR components. **g-k**, Protein levels in indicated pathways comparing cells with and without aggregates. Circles show means, bars indicate SD across patient samples (n = 34). Proteins in panels i-m were manually selected; panel n shows proteins from KEGG database. Box plots show first and third quartiles (box), median (thick line), and whiskers (±1.5 interquartile range).



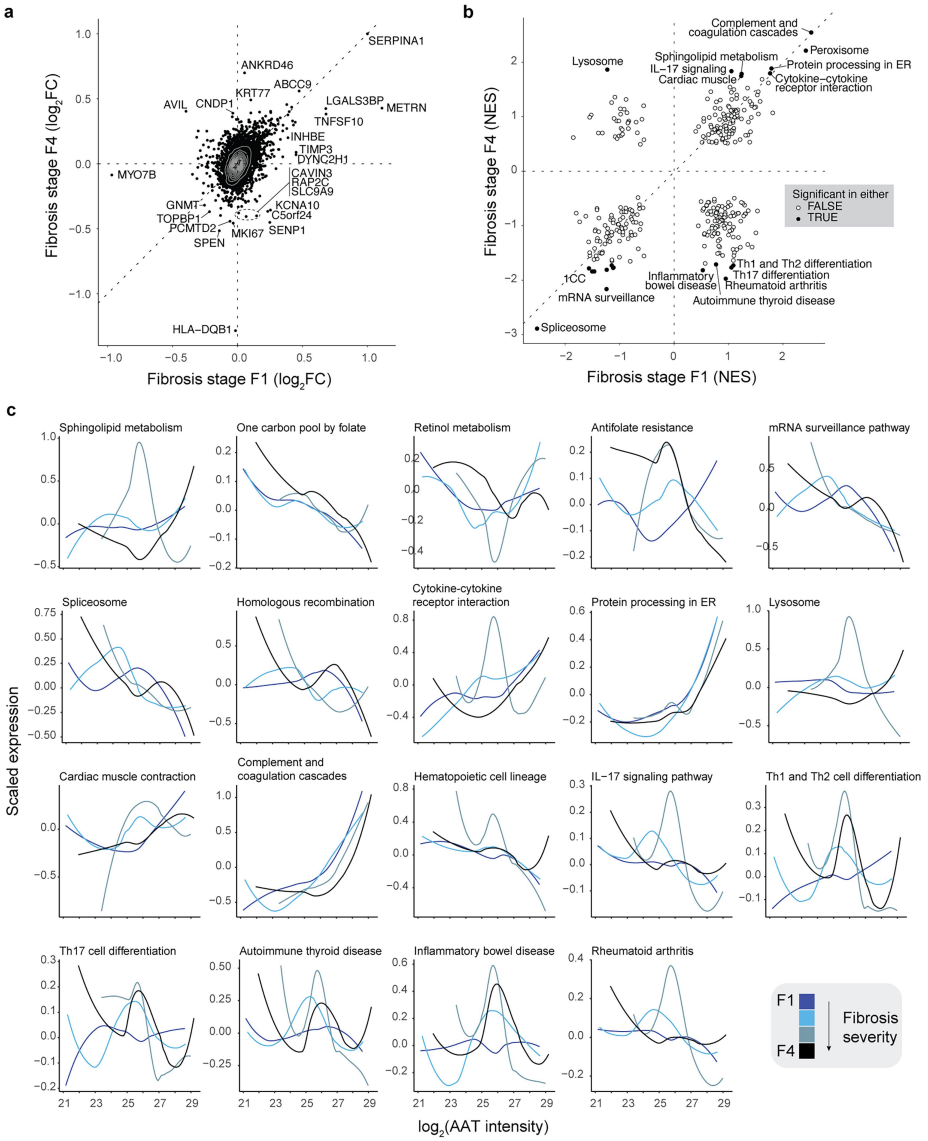
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Early and late responses to proteotoxic stress. **a**, Pearson's correlation coefficient (R) between each detected protein and alpha-1 antitrypsin levels per MS sample. Top and bottom 10 proteins are highlighted in boxes. **b**, Proteomic changes across high, moderate, and low AAT-accumulating cells with manually curated labels. **c**, Panel (b) overlaid with proteins from indicated KEGG pathways. Non-pathway proteins shown as density cloud. **d**, Expression levels of proteins involved in peroxisomal protein import, **e**, XBP1 signaling, and **f**, the Calnexin/Calreticulin cycle. Values shown as z-scores (assuming normal distribution) across samples split by load class. Database identifiers listed below each graph (n = 96 samples from 34 patients).

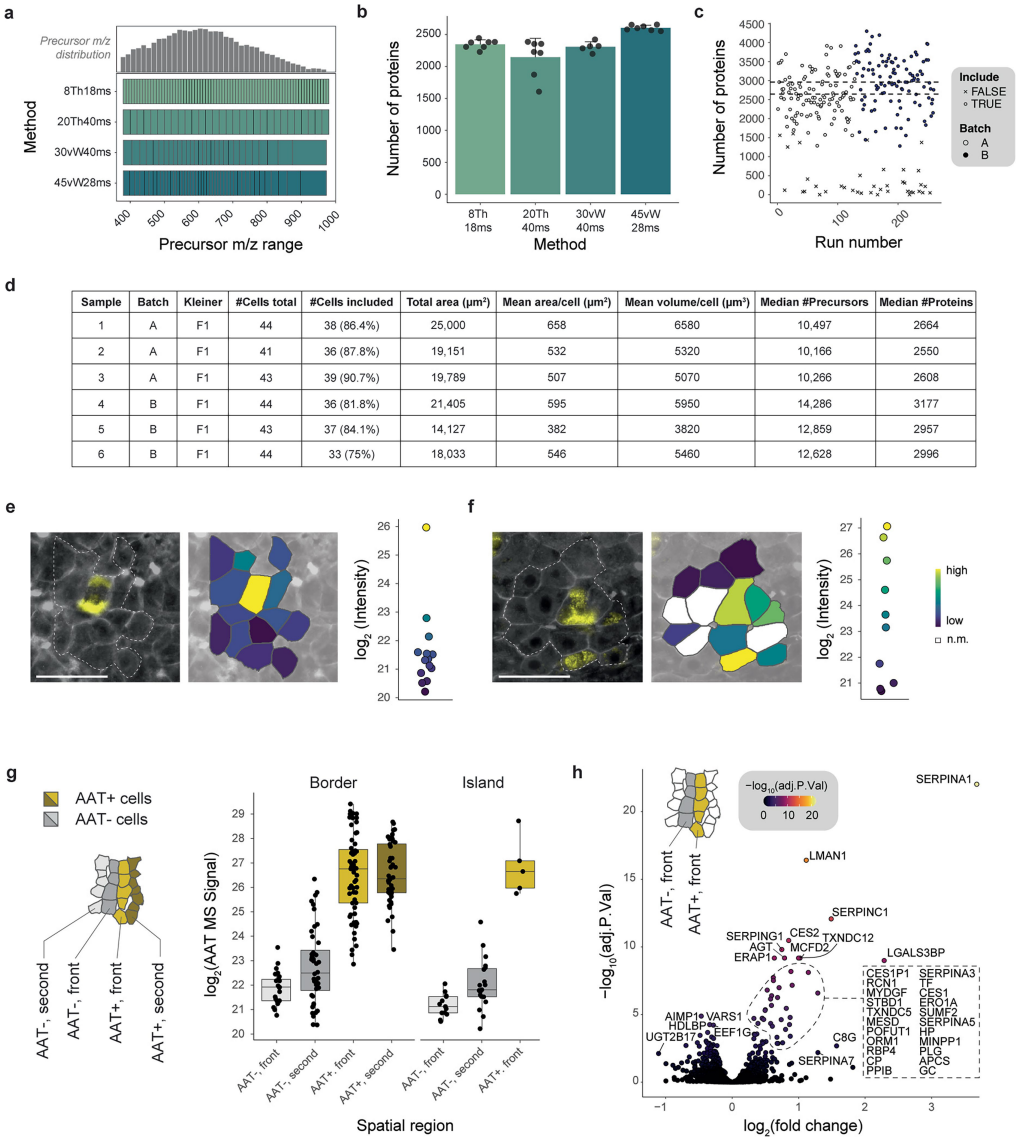


Extended Data Fig. 4 | Changes in functional pathways. a-f, Scaled protein intensities (z-scored) from indicated KEGG pathways plotted against AAT intensity. KEGG pathways identified by 'hsa00000' identifiers. Purple lines show local regression (span 0.75, degree 2).

Article



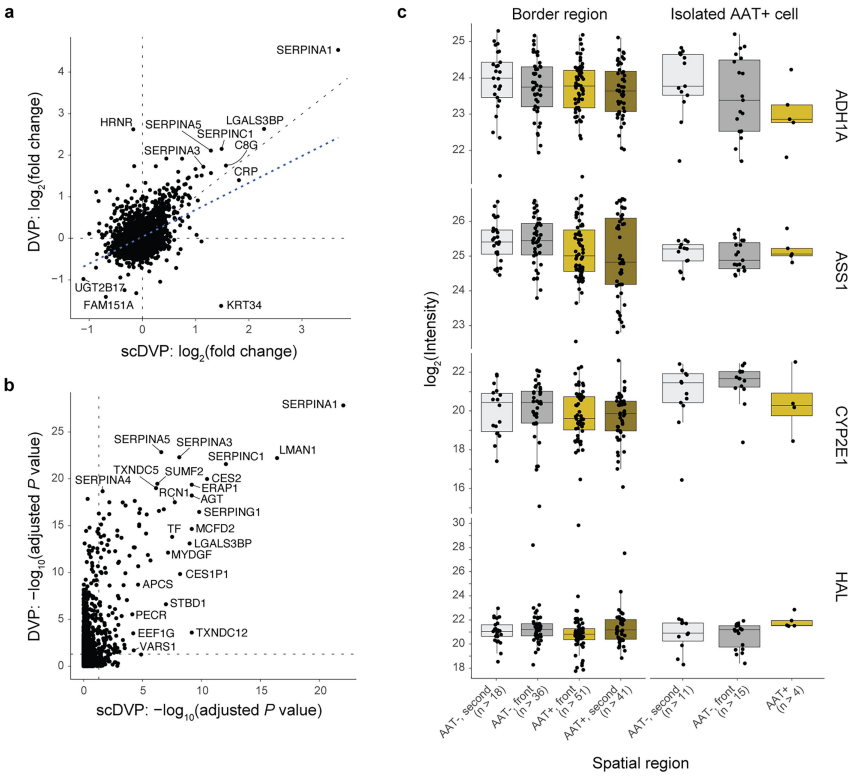
Extended Data Fig. 5 | Impact of fibrosis on functional pathways in relation to AAT load. a, Statistical comparison of three AAT load-defined hepatocyte classes, stratified by fibrosis grade (F1: n = 11, F4: n = 6; paired two-sided t-test with load class as covariate, multiple testing corrected). **b**, Gene Set Enrichment Analysis of fold changes shown in panel (a). Significant pathways are indicated by filled (black) or unfilled (white) markers; selected non-redundant terms are labeled. **c**, Scaled protein intensities (z-scored within fibrosis groups) for detected proteins in KEGG pathways plotted against AAT intensity. KEGG identifiers shown as 'hsa00000'. Purple line indicates local regression (span 0.75, degree 2). Legend for all panels shown in top right.



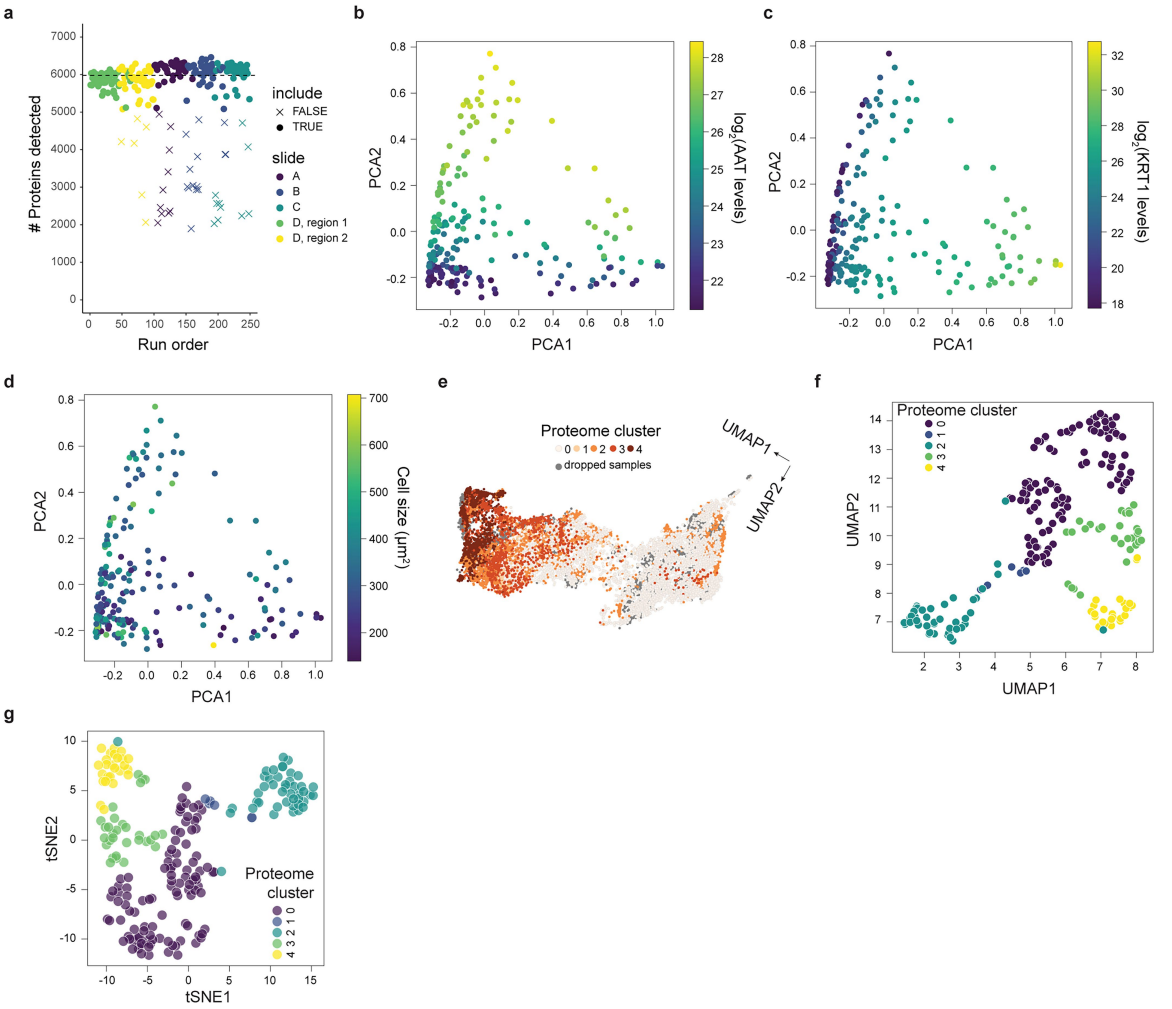
Sample	Batch	Kleiner	#Cells total	#Cells included	Total area (µm²)	Mean area/cell (µm²)	Mean volume/cell (µm³)	Median #Precursors	Median #Proteins
1	A	F1	44	38 (86.4%)	25,000	658	6580	10,497	2664
2	A	F1	41	36 (87.8%)	19,151	532	5320	10,166	2550
3	A	F1	43	39 (90.7%)	19,789	507	5070	10,266	2608
4	B	F1	44	36 (81.8%)	21,405	595	5950	14,286	3177
5	B	F1	43	37 (84.1%)	14,127	382	3820	12,859	2957
6	B	F1	44	33 (75%)	18,033	546	5460	12,628	2996

Extended Data Fig. 6 | The single-cell proteome. a, MS/MS acquisition design on the Orbitrap Astral mass spectrometer showing window width and injection time. "v" indicates variable windows (represented by box sizes). AAT expression shown by color in regions with single-positive cells. AAT levels for indicated shapes displayed in adjacent dot plots. **b**, Number of proteins quantified per acquisition strategy (n > 5) after exclusion of samples with less than 500 proteins. Error bars are positive standard deviations. **c**, Protein quantification across all hepatocyte shape runs (n = 259). Lower dashed line: median in batch A (2601 proteins); upper dashed line: median in batch B (3004 proteins). **d**, Summary statistics per sample. **e**, and **f**, AAT expression visualized by color in regions with single-positive cells. AAT levels for indicated shapes shown in adjacent dot plots. **g**, AAT expression in specified regions measured by immunofluorescence across all included samples (n = 219). Box plots show first and third quartiles (box), median (thick line), and whiskers (±1.5 interquartile range). **h**, Statistical comparison between AAT+ and AAT- cells in regions classified as 'borders' (paired two-sided t-test, multiple testing corrected, 63 AAT+ cells and 44 AAT- cells).

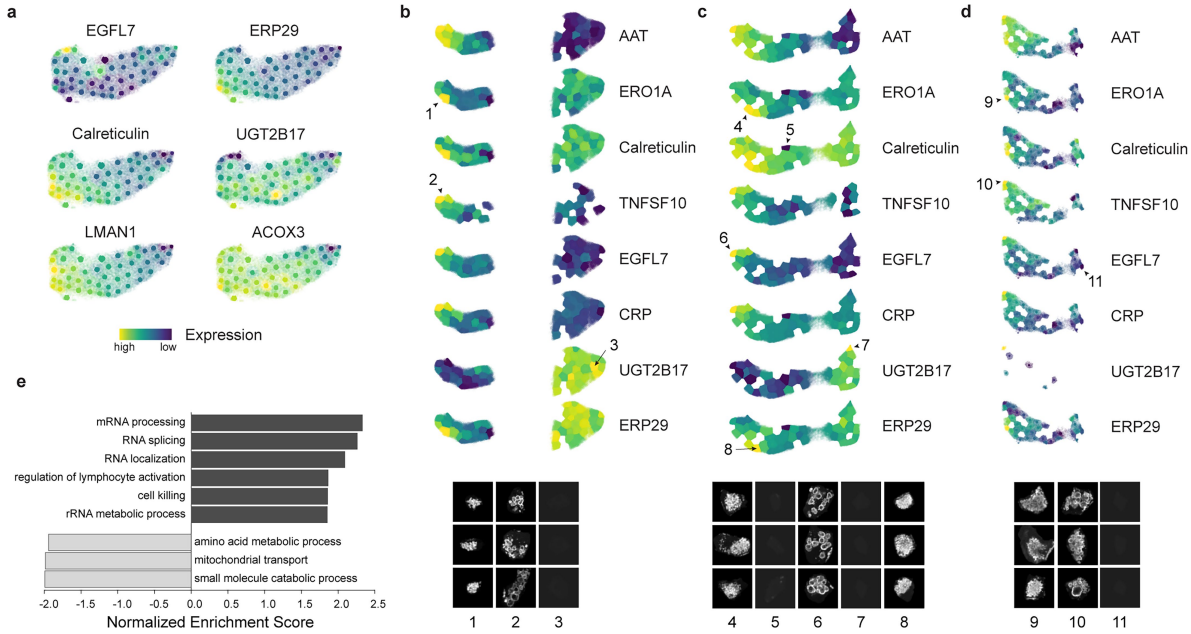
Article



Extended Data Fig. 7 | Comparison of the single-cell proteome with DVP class data and zonation. a, Comparison of $\log_2$(fold changes) and **b**, adjusted *P* values between AAT+ and AAT- single-cell comparisons (x-axis) versus cells along the accumulation gradient (y-axis; refer to Figs. 1 and 2). Statistics as in Extended Data Fig. 6h, and Fig. 1c. **c**, Protein expression in spatial regions for indicated markers. Periportal markers: ASS1 and HAL; pericentral markers: ALDH1A1 and CYP2E1. Box plots show first and third quartiles (box), median (thick line), and whiskers (±1.5 interquartile range).



Extended Data Fig. 8 | Quality control of morphology-guided DVP. a, Number of protein groups detected per sample. Each dot is one sample, the horizontal line indicates the mean across all included samples (n = 209 included, n = 41 excluded and marked with a cross). Exclusion criteria were that the number of detected proteins was smaller than mean minus 0.5 SD. **b**, Principal component analysis of all included samples with AAT, **c**, KRT1 expression levels, or **d**, shape size color coded (n = 209). **e**, Annotation of the proteome cluster in Fig. 4d onto the image space UMAP. Dropped samples are in grey (n = 12,500). **f**, Representation of individual samples color coded by proteome cluster in a proteomic UMAP, or **g**, tSNE space (n = 209).



Extended Data Fig. 9 | The proteome of cells with various aggregate morphologies. a-d, Protein expression across phenotypic UMAP space. Each panel represents one tissue section (n = 4). Notable clusters indicated by arrows and numbers, with representative images shown below. **e**, Gene Set Enrichment Analysis (GO: Biological Process noRedundant) comparing globular versus amorphous aggregate types.

4.5 scPortrait integrates single-cell images into multimodal modeling

Light microscopy is uniquely capable of assessing the spatial composition of cells down to a sub cellular resolution at high throughput. Furthermore, modern microscopes have the capacity to generate datasets of hundreds of millions of images describing various aspects of cellular composition and architecture in short time frames.

Machine learning powered by deep neural networks or deep learning, is a method that has recently emerged from computer vision research to evaluate these types of datasets. In deep learning, the many layers of the neural network are able to identify complex patterns in the input data and compress the contained information into a much smaller dimensional space. But for these types of models to be able to learn accurate representations, they need large amounts of training data. While modern microscopes have the capacity to generate datasets of hundreds of millions of images, to efficiently utilise this data for deep learning requires that it is properly organised.

For microscopy data, especially when integrating data from different sources, this task is very challenging. It requires the effective segmentation of individual cell outlines followed by the generation of single-cell images that can be provided to the model. Furthermore, this data needs to be stored in a format that can both handle terabyte-scale datasets, as well as quickly providing individual data instances on the fly, to allow for the efficient training of deep learning models.

Here, we built a computational pipeline which we have called *scPortrait*, which solves this task. *scPortrait* is built in python and is completely open source. Using out-of-core computation, it can efficiently handle very large image datasets that exceed available memory. By embracing open data formats and integrating into the *scverse* environment it is fully compatible with existing datasets and packages facilitating integrative workflows. Our package is widely applicable and is already being used across a variety of projects at multiple institutes.

We showcase *scPortrait*'s ability to not only generate single-cell image datasets, but to utilise these for cross-modality modelling by annotating fluorescently imaged tonsillitis samples with gene expression profiles generated in silico through flow matching. Additionally, we embed single-cell images into transcriptomic reference atlases and leverage morphological features to identify a specific subset of tumor-associated macrophages.

We anticipate that *scPortrait* will play a key role in future initiatives aimed at modelling cell behaviour using images and across multiple data modalities.

The following research article was originally published here:

Mädler, S. C. et al. (2025). "scPortrait integrates single-cell images into multimodal modeling". In: *bioRxiv*. DOI: 10.1101/2025.09.22.677590

scPortrait integrates single-cell images into multimodal modeling

Sophia C. Mädler[1,†,*], Niklas A. Schmacke[1,2,3,†,*], Alessandro Palma[2], Altana Namsaraeva[1,2,4], Ali Oğuz Can[2,6], Varvara Varlamova[3], Lukas Heumos[2], Mahima Arunkumar[2], Georg Wallmann[1], Veit Hornung[3], Fabian J. Theis[2,5,6,‡,*], Matthias Mann[1,‡,*]

[1] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[2] Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany

[3] Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, Munich, Germany

[4] Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA), Darmstadt, Germany

[5] School of Computing, Information and Technology, Technical University of Munich, Munich, Germany

[6] TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany

[†, ‡] These authors contributed equally

[*] Correspondence to: Sophia C. Mädler (maedler@biochem.mpg.de), Niklas A. Schmacke (niklas.schmacke@helmholtz-munich.de), Fabian J. Theis (fabian.theis@helmholtz-munich.de), Matthias Mann (mmann@biochem.mpg.de)

Abstract

Machine learning increasingly uncovers rules of biology directly from data, enabled by large, standardized datasets. Microscopy images provide rich information on cellular architecture and are accessible at scale across biological systems, making them an ideal foundation for modeling cell behavior. However, a standardized image format does not exist at the single-cell level. Here we present scPortrait, an scverse software package for generation, storage, and application of single-cell image datasets. scPortrait reads, stitches and segments raw fields of view with out-of-core computation scaling to larger-than-memory datasets. Parallelization enables rapid extraction of individual cells into a standardized single-cell image format with fast access to accelerate machine learning. scPortrait enables analysis across modalities including images, proteomics and transcriptomics, identifying cancer-associated macrophage subpopulations by morphology and embedding single-cell images into transcriptome atlases. scPortrait turns microscopy images into a reusable resource for integrative cell modeling, establishing single-cell images as a core modality in systems biology.

Introduction

The advent of machine learning has transformed multiple research fields including natural language processing[1], computer vision[2-4] and climate science[5]. Computational models can now detect patterns in complex datasets without external guidance. Given that biological datasets often contain entangled information on multiple underlying processes, and are therefore not straightforward to interpret, applying machine learning to biology research promises to uncover new mechanisms and generate new hypotheses[6,7].

A limitation of current machine learning approaches is their requirement for large datasets during training[2]. This has hampered their adoption in biology, where data acquisition is often costly and siloed. Fortunately, one of the most readily acquired modalities is imaging, a domain in which machine learning has recently shown tremendous success. Images can be acquired comparatively easily across biological scales, from whole ecosystems to subcellular structures, and they capture information on the spatial and temporal arrangement of a system's components in exquisite detail[8]. In cell biology, comprehensive image collections now describe cellular architecture[9], tissue structure[10,11] and perturbation responses[12]. In addition to increases in scale, recent advances have enabled the joint acquisition of images and other modalities such as genetic information[13,14,] protein abundances and transcriptomics[15]. Techniques like deep visual proteomics[16] and spatial transcriptomics[17] even enable paired collection of images and other modalities directly from tissue samples. The combination of spatially resolved imaging with the complementary and orthogonal molecular information from other modalities makes the resulting datasets a rich substrate for machine learning models[18].

Recent approaches to building comprehensive models of cellular activity, also called foundation models or virtual cells, critically depend on learning from multiple modalities[6,18,19]. Realizing the potential of the spatial resolution provided by images requires machine learning-compatible data structures that can support integration and large-scale modeling[20]. Storing data at the level of individual cells, the smallest functional units of life, is the de-facto standard in other modalities such as transcriptomics. However, existing image analysis software such as SPACEc[21], MCMICRO[22], spatiomic[23], QuPath[24], and CellProfiler[25,26] does not generate single-cell image datasets but instead runs analyses to generate and save collections of features. OME-NGFF, a recently proposed

storage format for biological images, currently does not provide a specification for saving single-cell images[27]. In addition, it saves images as collections of individual files in the zarr format, slowing down random access required when training machine learning models.

To address these limitations we present scPortrait, a software package and file format to process and store single-cell images (https://github.com/MannLabs/scPortrait). Through efficient parallelization and out-of-core computation, scPortrait accelerates the generation of standardized image datasets for machine learning from raw microscopy images on high-performance compute clusters. The resulting .h5sc files enable fast random access, reproducibility and integration with the scverse ecosystem, establishing images as a first-class modality for machine learning. We demonstrate the power of scPortrait for cross-modality modeling by annotating fluorescently imaged tonsillitis samples with gene expression data generated *in silico* using flow matching[28-30] and by embedding single cells into transcriptome atlases based on their images. We also use morphological information to identify a tumor-associated macrophage subset. We expect scPortrait to become an integral part of future efforts to model cell behavior across modalities.

Results

scPortrait generates single-cell image datasets at scale

scPortrait (https://github.com/MannLabs/scPortrait) is a software package and file format that transforms raw microscopy into standardized, analysis-ready single-cell images (Fig. 1). It ingests data from common sources, such as TIFF and zarr files, stitches individual fields-of-view[31], applies segmentation with built-in or external algorithms[32] and extracts individual cells into single-cell image datasets (Methods, Fig. S1a). These functions can be run in batches or included in workflow managing systems such as snakemake[33] or nextflow[34], using out-of-core computation to deal with larger-than-memory input images. Its open design allows individual processing steps in scPortrait to be run with integrated algorithms or to be offloaded to external pipelines. All intermediate results are saved as SpatialData[35] objects that can be inspected interactively[36]. CPU- and GPU parallelization speeds up processing. Using 64 threads, scPortrait extracts more than 700 cells per second (Fig. S1b, c). Extracted images are saved in our newly defined .h5sc file. The .h5sc format stores single-cell images based on the HDF5 containers of AnnData[37] and the scverse ecosystem[38], ensuring compatibility with existing and future tools[35,39-45]. By enabling fast reading and writing, the .h5sc format accelerates modern machine learning applications (Fig. S1d).

As we demonstrate below, scPortrait powers analyses built on generative modeling and morphology assessment and maps inferred cellular attributes back into the spatial domain, for example into tissue contexts. By making single-cell image storage findable, accessible, interoperable, and reusable (FAIR)[46], scPortrait provides the foundation for cross-dataset and cross-modality cellular representation learning, powering analyses from generative modeling to morphology-based tissue mapping[47].

To demonstrate that scPortrait supports diverse single-cell image embedding strategies, we generated a Golgi morphology benchmark using reporter HeLa cells expressing a fluorescent trans-Golgi marker. We then stimulated these Golgi reporter cells with compounds known to affect Golgi organization and imaged their Golgi morphology (Fig. S2a). Following processing with scPortrait, we featurized the resulting single-cell images in three ways: with ConvNeXt[48], a convolutional neural network (CNN) model pre-trained on natural images, with SubCell[49], a transformer-based model pre-trained on cellular images of the Human Protein Atlas[9] and with CellProfiler[25,26], a software that extracts pre-defined cellular features (Fig. S2b, c). Despite their different architectures, all embedding strategies clustered cells according to treatment, while separating them from untreated controls (Fig. S2c). In line with previous efforts to extract phenotypic information from image embedders[50], this shows that all perturbations induce reproducible and separable Golgi morphologies (Fig. S2a, c). We release this Golgi morphology dataset with scPortrait as a ready-to-use benchmark for comparing image-embedding methods.

Tissue modeling across modalities with scPortrait

scPortrait has already proven invaluable in the analysis of 120 million single-cell images from 40 million cells in a genome-scale image-based genetic screen for autophagosome formation[14] and in classifying the distinct intracellular distributions of the protein α1-antitrypsin in patient samples of the liver disease α1-antitrypsin deficiency (AATD)[51]. To highlight how scPortrait enables inference across modalities in a disease context, we analyzed a 59-plex CODEX imaging dataset of human tonsils affected by tonsillitis[21], which contained 1.1 million images of almost 20,000 cells after segmentation (Fig. S3a, b, c). To explore how individual cells contribute to tissue architecture beyond what imaging alone can provide, we sought to map publicly available dissociated CITE-seq data of human tonsils[52] onto this CODEX dataset. Since these datasets originate from different samples, there is no mapping between individual cells across modalities and neither dataset can be expected to contain all cell types or -states. This makes cross-modality alignment a non-trivial challenge, ideally suited to scPortrait's integrative framework.

To address this challenge, we turned to optimal transport, a mathematical framework that determines the most efficient way to map one distribution to another. In biology, optimal transport has recently been used to construct developmental trajectories over time and in space[42,53]. Because CODEX, via antibody staining and imaging, and CITE-seq, via sequencing of antibody-conjugated nucleotides, both measure protein abundances,

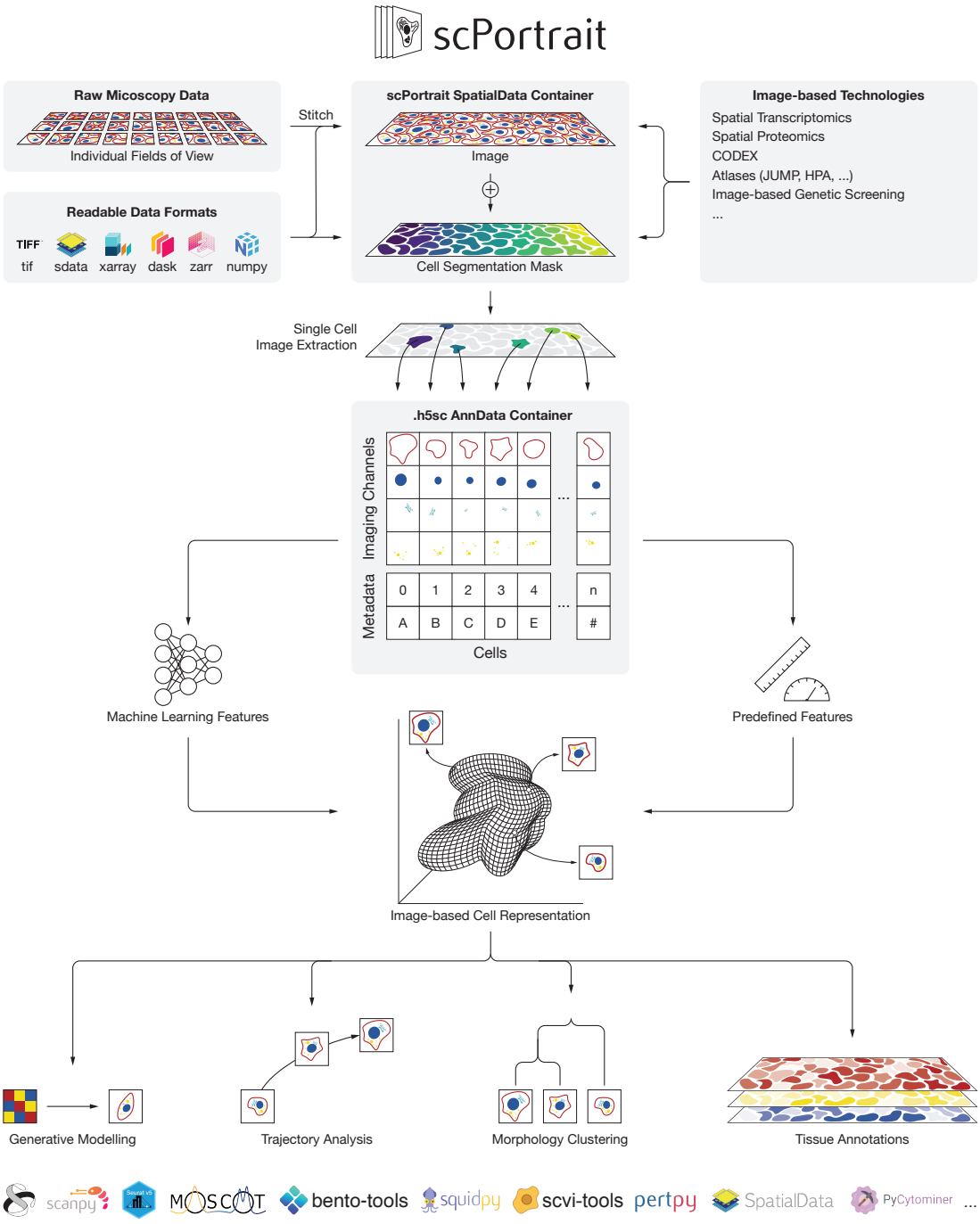


Figure 1 | The scPortrait software package and file format for single-cell image dataset generation

scPortrait generates single-cell image datasets from raw imaging inputs. It reads common image formats and stitches raw fields-of-view with high precision. Segmentation masks can then be created with built-in algorithms, external tools, or loaded directly. All intermediates are saved as SpatialData objects for compatibility with external annotations and third-party software. scPortrait then produces standardized single-cell datasets by applying segmentation masks for extraction of single-cell images. Out-of-core computation handles larger-than-memory datasets and all steps are parallelized to enable rapid processing. The resulting single-cell image datasets are stored in an HDF5-based .h5sc file format built on AnnData. This format is directly compatible with downstream analysis steps such as representation learning and integrates with existing software tools for single-cell analysis including scanpy and the scverse ecosystem.

we reasoned that these data should be well suited to generating a probability matrix linking cells across modalities under optimal transport constraints. Using this mapping, we then predicted gene expression based on CODEX features. However, this mapping is discrete, potentially suffering from sampling bias, and it does not necessarily contain all cell states of interest. To generalize beyond observed cell pairs and continuously infer gene expression based on CODEX features, we turned to flow matching[29,30]. This framework constructs probability paths that transform noise into data based on optimal transport maps, making it an ideal fit for our image-to-gene expression modeling problem. Flow matching has previously been used to predict cell behavior in response to diverse stimuli and perturbations[54-56]. To generate gene expression features conditioned on CODEX features, we trained our flow matching model by sampling pairs of source (CODEX) and target (gene expression) cells according to the probabilities in our optimal transport map (Fig. S3d).

To evaluate whether the gene expression profiles inferred by flow matching retained biological information, we tested for recovery of canonical cell type markers. We assigned cell types by k-nearest neighbors prediction to generated gene expression profiles in CITE-seq space. Indeed, we found enrichment of classic markers such as LYZ in monocytes and dendritic cells, CD3 in T cells and NKG7 in NK cells (Fig. S3e). As an aggregate measure of flow matching accuracy, UMAP representations of the measured and flow matching-inferred gene expression spaces overlap substantially, despite expected differences in cell type proportions between CODEX and CITE-seq data (Fig. S3f). These data show that flow matching predicts plausible gene expression features conditioned on CODEX data.

We then used our flow matching model to infer the expression of *TCL1A*, a marker gene of germinal center B cells (GCBC) that was not measured in the CODEX dataset. The inferred profiles revealed clusters of *TCL1A*-expressing cells precisely localized to germinal centers (Fig. S3g, h). Similarly, when inferring the expression of T cell marker *CD2*, we find strong colocalization with CD3 in the tonsil tissue (Fig. S3i, j). Beyond single marker genes, optimal transport enables transfer of higher-level annotations such as cell types from the CITE-seq reference onto the tissue, revealing structures including germinal centers (Fig. S3k). These data demonstrate that image-based cross-modality modeling can recover missing molecular features and reconstruct tissue organization, even when samples across modalities are not matched.

Modeling morphology of cells in tissues with scPortrait

Our CODEX data analysis relied on a simple featurization that collapsed single-cell images to mean intensities per channel, corresponding to protein abundances. To investigate the information contained in more complex, spatially resolved features, we turned to a joint spatial transcriptomics and fluorescence microscopy dataset of human ovarian cancer acquired using the 10x Genomics Xenium platform[57] (Fig. 2a, b). Using the provided segmentation masks we extracted 1.6 million single-cell images across four channels with scPortrait (Fig. 2c). We aimed to embed these cells into a continuous representation space based on their morphologies alone. Rather than biasing this representation towards predefined labels, we adopted a self-supervised learning approach to retain as much cell morphology information as possible. This paradigm has been shown to generate meaningful image representations irrespective of external labels by training on auxiliary tasks such as reconstructing masked image patches[58].

We fine-tuned a vision transformer-based autoencoder (ViT-MAE)[59] on the ATP1A/CD45/E-Cadherin, 18S and αSMA/Vimentin channels of the ovarian cancer dataset by mask reconstruction (Fig. 2d). The representation learned by this model's encoder captured meaningful biology, as demonstrated by its ability to separate single-cell images by their morphological phenotypes in line with a previous report[60] (Fig. 2e, S4a). Remarkably, it had also implicitly learned to group cells by type despite never being trained on cell type labels (Fig. 2e, S4a). To test whether this image-based representation contained information absent from the transcriptome, we inspected the difference in local neighborhood structure for each cell in both image- and transcriptome space. If the two modalities contained similar information, the same cells should be neighbors of one another in each embedding. Average neighborhood overlap was less than 5% across cell types, demonstrating that our image-based embedding encodes largely non-redundant information (Fig. S4b, c). Tumor cells generally exhibited the highest overlap while T and NK cells showed the lowest (Fig. S4c).

To probe the information captured by our fine-tuned ViT-MAE model in more detail, we focused on macrophages, a heterogeneous and functionally diverse cell type. Clustering all macrophages by image features revealed

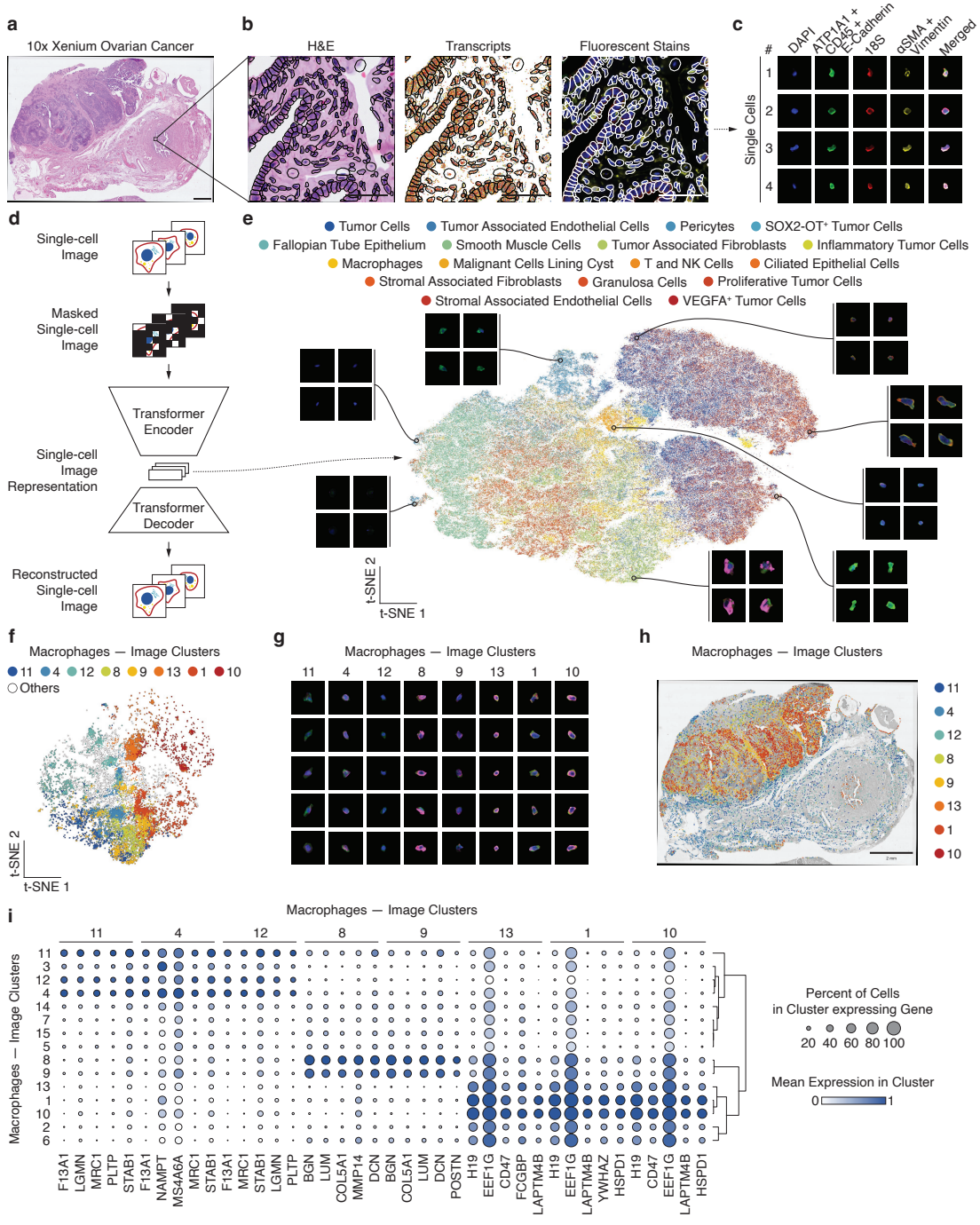


Figure 2 | scPortrait identifies macrophage subpopulations in ovarian cancer

a, H&E overview image of tissue region contained in 10x Genomics Xenium dataset of human ovarian cancer. Scalebar represents 1 mm. **b**, Magnified region of a, black outlines show cell cytosol borders. Different panels show modalities contained in the dataset. Dots in center panel correspond to probes binding individual transcripts. Scalebars represent 50 µm. **c**, Single-cell images extracted with scPortrait. ATP1A + CD45 + E-Cadherin and αSMA + Vimentin were stained together in single imaging channels. We extracted a total of 1,627,204 single-cell images across four channels. The images were min-max scaled for visualization. **d**, Overview of single-cell image embedding strategy. We trained a transformer-based encoder-decoder model on the ATP1A + CD45 + E-Cadherin, 18S and αSMA + Vimentin channels via mask reconstruction (ViT-MAE) independently of biological labels. We used the internal representation learned by this model as an image-based featurization of cells in the dataset. **e**, t-SNE visualization of ViT-MAE embeddings of single-cell images colored by dataset cell type annotation. Each dot represents one cell. Images from indicated regions were not rescaled. **f**, e filtered for macrophages. Colors indicate selected Leiden clusters. **g**, Single-cell images of macrophages from clusters in f. **h**, Distribution of macrophage clusters from f across the tissue region. **i**, Genes differentially expressed in macrophage clusters from f.

several morphologically distinct subclusters (Fig. 2f, g, S4d). Their spatial distributions differed strikingly, particularly in their intra- versus extratumoral localization (Fig. 2h, S4e, f). Thus, morphology alone was sufficient to distinguish macrophage states associated with distinct tissue niches, demonstrating that image-based embeddings can resolve biologically meaningful heterogeneity.

To explore the functional characteristics of these morphologically distinct macrophages, we turned to their spatial transcriptomics profiles. Differential gene expression analysis revealed cluster-specific signatures (Fig 2i, S4g): Clusters 4, 11 and 12, predominantly extratumoral, express *MRC1*, encoding CD206, and *STAB1*, markers of anti-inflammatory macrophage subpopulations[61,62]. In contrast, clusters 8 and 9 were almost exclusively found intratumoral and expressed *LUM* and *COL5A1*, consistent with a population of cancer-associated fibroblasts mislabeled as macrophages[63,64]. These results show that morphology-based clustering by representation learning with scPortrait can resolve functionally defined cell states from images.

scPortrait embeds cells into a transcriptome atlas based on their images

Large, labeled single-cell collections are increasingly becoming available[65-67], but most of these atlases are centered around single-cell transcriptomics. After using single-cell images to identify an anti-inflammatory macrophage subset, we asked whether single-cell images could be annotated directly from transcriptome atlases. We reasoned that a small amount of overlap in image- and transcriptome-based information might be sufficient to embed cells into transcriptomic atlases directly from their images.

We first projected all cells of the ovarian cancer data set into the SCimilarity atlas[65] based on their transcriptome. Then, using the ViT-MAE image features we generated as inputs, we trained a multilayer perceptron (MLP) as a cross-modality model to predict SCimilarity features from the ViT-MAE image embeddings (Fig. 3a). For all analyses of these data we excluded tumor cells, since SCimilarity is not trained to properly embed them[65]. On a held-out test set this cross-modality model achieved an R^2 value of 0.65 (Fig. 3b). Inspecting assigned ovary cell type labels showed that a variety of cell types was predicted, and similar cell types clustered together. To understand the biases of our model we then investigated the prediction error by cell type. Error analysis revealed similar performance across most cell types, with epithelial cells and T / NK cells showing the highest errors, suggesting that variance was not dominated by a single lineage (Fig. 3c).

An advantage of projecting cells of new experiments into existing atlases is that atlas labels can be used to infer biological information about previously unlabeled samples. Given that the ovarian cancer dataset contained cell type labelling already, we compared the provided cell types with those predicted by transcriptome- or image-based SCimilarity embeddings. Most cells showed strong agreement across all three labels (Fig. 3d, S5a, b) and mismatches typically involved related types such as smooth muscle cells and myofibroblasts. In line with the increased prediction error, ciliated epithelium and leukocytes were notable exceptions (Fig. 3c, d). To further validate the approach, we mapped predicted cell types onto a tissue region not used for training. Both transcriptome- and image-derived SCimilarity cell type labels annotated the tissue correctly, identifying epithelial structures, fibroblasts and smooth muscle cells (Fig. 3e, f, g). Together, these results demonstrate that cells can be embedded into transcriptome atlases based on images alone, recovering meaningful biological information and underscoring the role of imaging in multimodal integration and modeling.

Discussion

Microscopy is one of the most information-rich and scalable ways to study cells, yet image data have lagged behind other single-cell modalities in standardization and integration. While transcriptomics and proteomics already rely on widely adopted formats that enable large-scale analysis, image-based datasets remain fragmented, often tied to specific pipelines or instruments. scPortrait addresses this gap by introducing .h5sc, a standardized and accessible format for single-cell images, together with a scalable software framework for dataset generation and sharing.

Open formats are crucial for modern life science research[47]. By building .h5sc on HDF5 and the AnnData

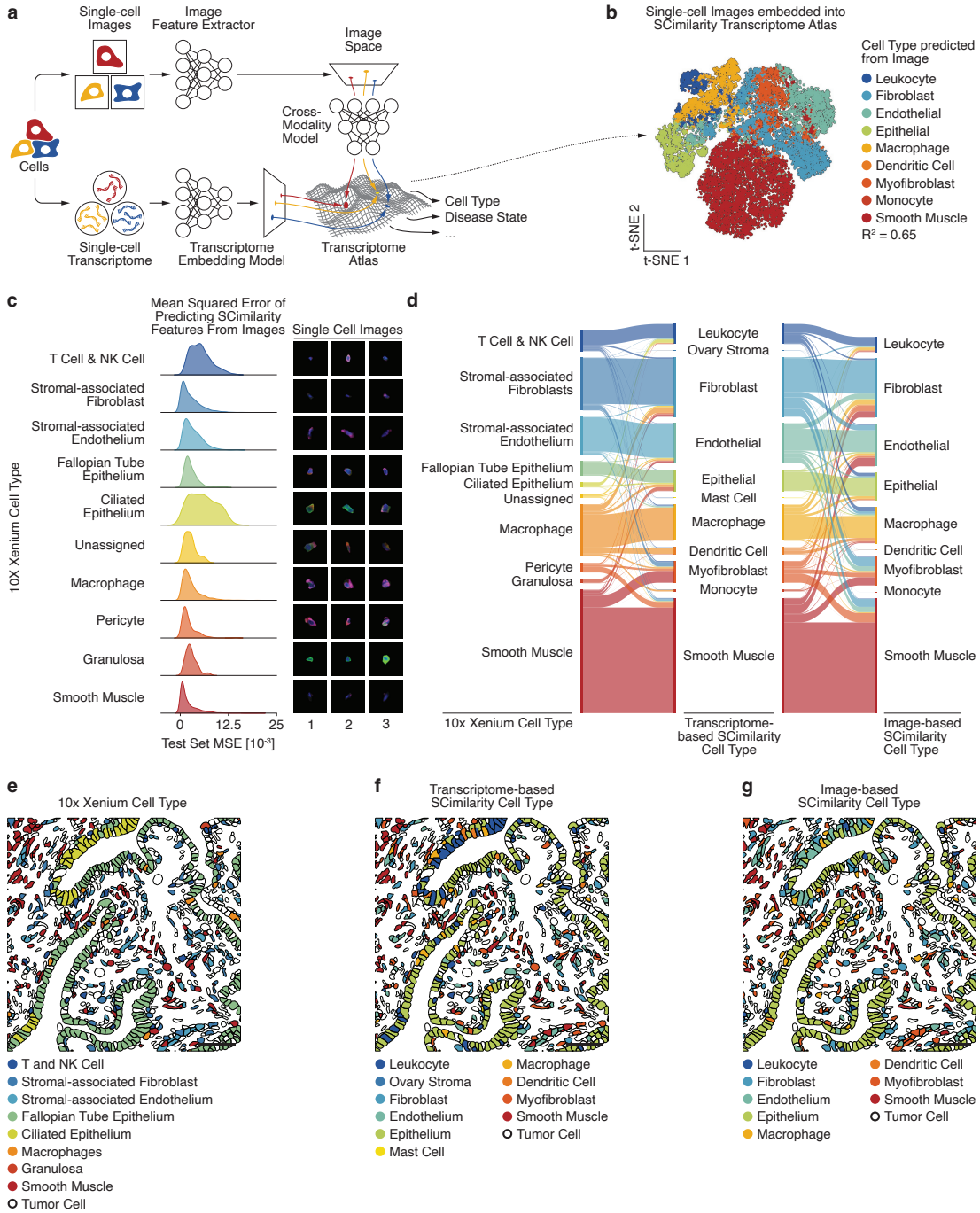


Figure 3 | scPortrait enables cross-modality embedding of single-cell images into a transcriptome atlas

a, Overview of our strategy to embed single cells into a transcriptome atlas by their images. First, we embed all cells from the ovarian cancer dataset into the single-cell transcriptome atlas SCimilarity by their transcriptome. We then train a multilayer perceptron (MLP) model to predict a cell's SCimilarity embedding from its ViT-MAE image embedding. This model embeds cells into SCimilarity even if only low-quality transcriptome information is available, based solely on their images. **b**, t-SNE of test set cells embedded into SCimilarity atlas, colored by SCimilarity cell type label. **c**, Left: Mean squared error (MSE) of SCimilarity embedding prediction from single-cell images by cell type. Only test set data are shown. Right: Corresponding single-cell images. **d**, Comparison of cell type labels in the ovarian cancer dataset (left) to transcriptome-derived SCimilarity embeddings (center) and image-derived SCimilarity embeddings (right). Only test set data are shown. **e, f, g**, Spatial distribution of cell type labels from the ovarian cancer dataset (e), transcriptome-derived SCimilarity embeddings (f) and image-derived SCimilarity embeddings (g). The depicted tissue region was excluded from the training set. Scalebars represent 50 μm.

specification[37], scPortrait ensures compatibility with the scverse ecosystem and downstream tools such as squidpy[44] or bento[40]. This design promotes interoperability, reproducibility, and reuse, aligning single-cell imaging with the FAIR principles[46]. In addition, .h5sc delivers fast random access to individual cells — essential for training large machine learning models — and avoids repeated, compute-intensive preprocessing. Thereby, scPortrait elevates single-cell imaging to the same level of accessibility and reuse that transcriptomics has already achieved, filling a critical community need.

A key advantage of images is the short turnaround time of their acquisition, enabling iterative evaluations with machine learning models in a loop. Paradigms such as active learning[68], where new observations are specifically sampled to address a model's current uncertainties, benefit from faster iteration times with scPortrait. Such approaches can be used in real time to regulate image acquisition parameters, enabling microscopes to dynamically respond to sample characteristics[69]. Across experiments, reinforcement learning can guide experimental design: A model would use the results from a given round of imaging experiments combined with the knowledge from already available representations learned by existing models[70] to recommend a new set of experiments for the next round, iteratively exploring the underlying biology. Reinforcement learning benefits directly from consistent representation of data at a defined level, achieved through the standardized .h5sc single-cell image datasets created by scPortrait.

In this study, we showed how scPortrait enables analyses that go beyond current imaging workflows. Using self-supervised representations of ovarian cancer tissue, we identified macrophage subpopulations with distinct spatial distributions and transcriptomic signatures, illustrating how morphology can resolve biologically meaningful cell states. In tonsil tissue, cross-modality mapping and flow matching allowed us to infer gene expression directly from CODEX images, recovering missing markers such as TCL1A and revealing tissue organization without matched samples. Finally, embedding cells into the SCimilarity transcriptome atlas demonstrated that morphology-derived features can generalize across datasets, opening the door to image-driven atlas annotation.

Beyond its immediate applications, image-based modeling raises unique challenges that will shape the next stage of single-cell analysis. Imaging experiments vary widely in hardware, staining protocols and preprocessing, leading to strong batch effects that hinder integration across datasets[71]. Similar challenges have been addressed in transcriptomics and proteomics[71,72], and a standardized format such as .h5sc can catalyze comparable solutions for imaging by making large, diverse datasets broadly accessible.

Another limitation is interpretability: image embeddings often lack clear biological meaning compared with transcript or protein abundances. Here, integration across modalities provides a powerful remedy. In our ovarian cancer analysis, scPortrait-derived morphology separated macrophage subsets whose molecular identities were clarified by transcriptomic profiles. Such cross-modality approaches not only validate image-based findings but also enable the discovery of new cell states that might otherwise remain hidden.

In conclusion, scPortrait provides the infrastructure needed to place imaging alongside transcriptomics and proteomics as a core modality for single-cell biology. By enabling large-scale sharing and integration, it supports the development of multimodal foundation models that learn from images as well as molecular profiles. Such models hold the promise of capturing complementary aspects of cell identity and behavior, ultimately enabling a more complete and predictive understanding of human biology.

Methods

The scPortrait software package

The Python-based scPortrait software package is available on GitHub (https://github.com/MannLabs/scPortrait) with documentation and tutorials (https://mannlabs.github.io/scPortrait/). scPortrait reads raw microscopy data and then follows a processing pipeline that generates standardized single-cell image datasets via three main steps as outlined here (https://mannlabs.github.io/scPortrait/pages/workflow.html):

- Stitching
- Segmentation
- Extraction

Two additional steps can follow:

- Featurization
- Selection of cells for downstream computational or biological analysis

All processing steps are carried out on an scPortrait project that defines a data structure on disk for saving intermediate and final results. One config file per project specifies options for each of the above steps. A variety of workflows are available for each step and can be used directly or adapted to suit specific dataset characteristics.

Processing of human tonsil CODEX data

Raw TIFF images were percentile-normalized to the 0.1 % - 99.9 % range per channel and then separated by disease status into healthy and tonsillitis tissue cores. These images were then segmented with scPortrait using the CellPose[32] nucleus model followed by mask expansion to generate cell borders. Single-cell images were then extracted into 36 × 36 px images. To preserve relative signal strengths across cells for each channel we did not rescale single-cell image intensities. Single cells were then featurized using the scPortrait CellFeaturizer, and mean intensity per channel was used for downstream analyses.

Mapping CODEX features to single-cell RNA-seq with generative Optimal Transport

Problem statement

Given a dataset of N CODEX cell images represented by D protein marker features measured by scPortrait, our goal is to derive an approximation of the gene expression profiles of the imaged cells, leveraging a multi-modal single-cell reference dataset containing gene expression counts and protein markers.

Formally, let $Y \in \mathbb{R}^{N \times D}$ be the matrix of N imaged cells across D marker-specific features. Moreover, denote $X^p \in \mathbb{R}^{M \times D}$ and $X^g \in \mathbb{R}^{M \times G}$ the *cell* × *protein* and *cell* × *gene* matrices in a reference single-cell CITE-seq atlas from the same tissue as the CODEX samples. Note that X^p has the same number of features as Y, as we subset the two feature spaces to reflect the same measured protein markers. Our goal is to derive a predicted gene expression matrix $\hat{X}^G \in \mathbb{R}^{N \times G}$ for each cell imaged by CODEX.

Our approach consists of two steps:

1. Match CODEX samples to their putative single-cell counterpart based on the similarity between image-based and CITE-seq-derived protein marker abundances. We pair the two modalities with Optimal Transport (OT).
2. Learn a mapping that transports CODEX features to the single-cell gene expression measured in the atlas.

To account for the noise in the single-cell dataset, we train a generative model based on flow matching[29,30] that generates novel gene expression profiles using the image-based marker features as input.

Learning OT with flow matching

Flow matching learns a parameterized, time-resolved vector field ${v^\theta}_t(x)$ that maps samples from a prior distribution $N(0, I)$ (by convention, at $t = 0$) to a target data distribution p at $t = 1$, thereby acting as a generative model by turning noise into data samples via the following equation:

$$\Phi(x_0) = x_0 + \int_0^1 v_t{}^\theta(x)\, dt, \text{ with } x_0 \sim N(0,I) \text{ and } \Phi(x_0) \sim p. \quad (1)$$

The integral function Φ mapping noise to data is called *flow*.

Klein et al.[28] showed that one can train flow matching to approximate a generative OT map from a source distribution q to the target data distribution p described above by conditioning the generative process with samples $y \sim q$from the source:

$$\Phi(x_0 \mid y) = x_0 + \int_0^1 v_t{}^\theta(x \mid y)\, dt, \text{ with } x_0 \sim N(0,I),\ y \sim q \text{ and } \Phi(x_0 \mid y) \sim p. \quad (2)$$

In other words, flow matching learns to transport a sample y from the source to the target stochastically according to an OT criterion, based on a pre-defined cost function.

Model training

In our setting, where the aim is to map cells from single-cell images in CODEX to single-cell CITE-seq, we indicate the image-based feature distribution as q and the single-cell CITE-seq distribution as p. We parameterize the velocity function ${v_t}^\theta(x)$ using a neural network trained with stochastic gradient descent over minibatches. One training iteration consists of the following steps:

1. Draw a random batch of B samples $Y_b = \{y_i\}^B{}_{i=1}$ from the CODEX dataset.
2. Draw a random batch of B single cells from the multi-modal reference atlas in both their protein and gene expression views as a target. We denote the target batch with $X_b = ({X_b}^G, {X_b}^P) = \{({x^G}_i, {x^P}_i)\}^B{}_{i=1}$.
3. Compute an OT coupling matrix Π between all observations in Y_b and ${X_b}^P$ minimizing the following *Euclidean cost:* $c(y_i, {x^P}_i) = ||\, y_i - {x_i}^P||_2$. The coupling approximates a joint distribution between source and target sample indices, where CODEX cells are mapped with a higher probability to atlas cells with similar marker abundance.

4. Resample couples of source and target indices from the joint distribution Π, yielding the resampled batches $\tilde{Y}_b$ and $\tilde{X}_b = \{(\tilde{x}^G{}_i, \tilde{x}^P{}_i)\}^B{}_{i=1}$. The i^{th} sample in $\tilde{Y}_b$ is transported to the i^{th} sample in $\tilde{X}_b$ with high probability Π.

5. Perform a flow matching iteration[30], training the vector field to transport a noise sample $x_0 \sim N(0, I)$ to the gene expression vector $\tilde{x}^G{}_i$ conditioned on its matched CODEX feature vector $\tilde{y}^G{}_i$.

Inference

During inference, we transport a CODEX sample y to its gene expression counterpart by sampling a noise point $x_0 \sim N(0,I)$ and computing $\hat{x}^G = \Phi(x_0 \mid y)$.

Data preprocessing

We subset the CODEX features extracted by scPortrait to the mean intensity of the marker channels and standardize individual features to mitigate skewness towards zero. To ensure correspondence between image and single-cell protein abundance data, we subset the CODEX features extracted by scPortrait and the single-cell protein abundance measurements to their 31 shared markers.

Since flow matching is a model working in continuous space, we use batch-corrected and 50-dimensional expression features extracted by the HARMONY algorithm[73] as a target representation for generation. To predict gene expression from translated CODEX features, we pre-train a decoder function that maps HARMONY embeddings to gene expression.

Model architecture and training details

1. The flow matching velocity model $v^\theta{}_t(x)$ is an MLP with 3 layers of 1024 hidden dimensions each and an ELU activation function. We train it with the AdamW optimizer with a learning rate of 1e-4, a batch size of 256 and across 2000 epochs. The velocity function $v^\theta{}_t(x)$ inputs a concatenation of the current state with a generation time embedding. To embed the generation time, we use a sinusoidal encoding[1] with 128 dimensions.

2. The decoder function, which maps HARMONY features to normalized gene expression, is an MLP with two layers of 64 hidden dimensions. We train the model for 200 epochs with a batch size of 256 and a learning rate of 1e-3. We choose a mean-squared error (MSE) loss function to reconstruct gene expression.

Processing of human ovarian cancer dataset

We downloaded the Xenium ovarian cancer dataset from 10x Genomics using this link: https://www.10xgenomics.com/datasets/xenium-prime-ffpe-human-ovarian-cancer. After reading the dataset into SpatialData[35], we generated single-cell images with scPortrait by extracting into 224 × 224 px images using the segmentation masks provided with the dataset. To preserve relative signal strengths across cells for each channel we did not rescale image single-cell intensities.

Golgi morphology experiments

On day 0, 1 million HeLa cells expressing hsTGOLN2-mCherry were plated per well of a 4-well plate containing a UV-sterilized metal frame slide with a polyphenylene sulfate (PPS) membrane. Cells were cultured in DMEM supplemented with 10 % FCS, 1 mM pyruvate, 100 U/ml penicillin and 100 µg/mL streptomycin at 37 °C and 5 % CO_2. On day 2, cells were treated with 10 µM Golgicide A, 10 µM Nigericin or 20 µM Monensin for 2 hrs or with 5 mg/mL Nocodazole for 30 minutes. Cells were then stained with 10 µg/mL WGA-Alexa-647 in PBS for 10 minutes at 37 °C before being washed 3× in PBS and then fixed in 4 % paraformaldehyde (PFA) diluted in PBS for 10 minutes at room temperature. After fixation, cells were washed 3× in PBS again before being stained with 10 µg/mL Hoechst 33342 for 15 minutes at room temperature. Afterwards, cells were washed 3× in PBS again and imaged on a Nikon Eclipse Ti2 spinning disc confocal microscope.

Training ViT-MAE

We fine-tuned a vision transformer-based masked autoencoder (ViT-MAE[59]) on the ovarian cancer dataset to learn a representation of cell morphology in an unsupervised manner, starting from a model pre-trained on natural images (https://huggingface.co/facebook/vit-mae-base)[74]. The raw images contained 4 stains: DAPI, ATP1A/CD45/E-Cadherin, 18S and αSMA/Vimentin. To match the input dimensions of the pretrained model we subsetted to three channels, discarding DAPI to focus on the functional structures of cells. The dataset contains 406,875 images of single cells, which were split into 90 %, 5 % and 5 % for training, validation and test sets respectively, including a spatially defined region in the test set. Prior to training, the images were cropped around the center at a fixed size of 128 × 128 and resized to 224 × 224 to match the expected ViT-MAE input size. The resulting single-cell image dimension was 3 × 224 × 224. No other augmentations were applied. The model was trained for 119 epochs using the scPortrait PyTorch dataloader with a batch size of 128[75]. After splitting the input image into 16 × 16 px patches, the encoder of the model passes the input through 12 transformer layer blocks, each of which uses multihead self-attention with 768 embedding dimensions per head and 12 attention heads. The patch size is 16 × 16. GELU is used as an activation function with a layernorm of eps=1e-12. The decoder of the network is lighter than the encoder but contains the same structure with 8 attention blocks of 512 embedding dimensions and 16 attention heads. The intermediate sizes of the feedforward layers are 3072 and 2048 for the encoder and the decoder. We use a masking ratio of 75 % of patches, which is set to 0 during inference after training. In total, the model has 111 million trainable parameters. We construct the latent space by average pooling across all tokens. After scaling to zero mean and unit variance we

used the Leiden algorithm with a resolution of 0.25 for clustering the latent space on 15 neighbors.

Embedding cells into SCimilarity atlas

To embed cells from the ovarian cancer dataset into SCimilarity, transcriptomics data were preprocessed as described in the SCimilarity documentation: Each cell was normalized to 10,000 counts, and the expression matrix was aligned with the SCimilarity atlas. Of note, we did not log1p transform our data, since the distribution of probe-based spatial transcriptomics counts is different from stochastically sampled dissociated transcriptomics assays. We then calculated SCimilarity features for all cells in the dataset. The SCimilarity checkpoint used was downloaded from https://zenodo.org/records/10685499.

To embed cells into SCimilarity based on their images, we trained a multilayer perceptron (MLP) to predict transcriptome-derived SCimilarity features from per-vector min-max scaled ViT-MAE embeddings. The MLP consisted of five linear layers gradually decreasing in size from the 768 dimensions of the ViT-MAE features to the 128 dimensions of the SCimilarity features. ReLU was used as an activation function. The model was trained to minimize the mean squared error (MSE) of predicted to transcriptome-derived SCimilarity features with a learning rate of 1e-5. Model training was done in PyTorch Lightning[75,76], with minimal validation loss as a selection criterion for the final set of model parameters used during inference. The same test set was held out from both ViT-MAE training and cross-modality MLP training, including a spatially defined region.

When predicting cell types for the ovarian cancer dataset we limited the list of possible cell types to those typically found in the ovary. All tumor cells were excluded from SCimilarity cell type predictions, because tumor cells were not part of the training data.

Differential marker expression testing

Subgroups for testing of marker enrichment were assigned by Leiden clustering in scanpy or via external labels such as cell types. Differentially enriched markers or expressed genes were determined using scanpy's *rank_genes_groups* function. The results were visualized using *rank_genes_groups_dotplot*.

Data Availability

All data generated as part of this work are available at https://zenodo.org/records/17162225 or can be regenerated from publicly available sources using the code in the repository at the following URL: https://github.com/MannLabs/scPortrait_manuscript.

Code Availability

The scPortrait software package is available at https://github.com/MannLabs/scPortrait. All figures in this manuscript can be recreated using the code in the repository at the following URL: https://github.com/MannLabs/scPortrait_manuscript.

Acknowledgements

We would like to thank the Center for Advanced Light Microscopy (CALM) for support with light microscopy; Marvin Thielert and Marc Oeller for testing new features of scPortrait; Magnus Schwörer for helping automate scPortrait's GitHub workflows; Piero Coronica from the Max Planck Computing and Data Facility (MPCDF) for support with improving the memory footprint of scPortrait's single-cell image dataset class during model training; and Ilan Gold for discussions about strategies for developing scPortrait as an scverse package from the ground up. We thank all scPortrait contributors and users who provided feedback. We are grateful to the communities behind the multiple open-source software packages on which we depend.

Author Contributions

S.C.M., N.A.S., F.J.T. and M.M. conceived the study. S.C.M., N.A.S., V.H., F.J.T. and M.M. designed experiments. S.C.M., N.A.S., A.N., L.H., G.W., F.J.T. and M.M. conceived the scPortrait software. S.C.M., N.A.S., A.P., A.N., A.O.C., V.V., M.A. and G.W. performed experiments. S.C.M., N.A.S., A.P., V.H., F.J.T. and M.M. wrote the manuscript with input from all authors.

Funding

S.C.M. was supported by a PhD fellowship of the Boehringer-Ingelheim Fonds. N.A.S. is supported by the Add-on Fellowship of the Joachim Herz Foundation. A.N. is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. This work was supported by the Max-Planck Society for the Advancement of Science and the ERC (ERC-2020-ADG–101018672 ENGINES). This work was also funded by the European Union (ERC, DeepCell - 101054957). This work was supported by the Helmholtz Association's Initiative and Networking Fund through CausalCellDynamics (grant # Interlabs-0029).

Declaration of competing interests

S.C.M. consulted for Lamin Labs GmbH and is a future employee of Ensocell Therapeutics. N.A.S. consulted for Lamin Labs GmbH. A.N. and L.H. are employees of Lamin Labs GmbH. G.W. is a founder of Aplusia GmbH. F.J.T. consults for Immunai, Singularity Bio, CytoReason, Cellarity, and Omniscope and has ownership interest in Dermagnostix and Cellarity. M.M. is an indirect investor in Evosep Biosystems and OmicVision Biosciences.

References

1. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
3. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *arXiv*, arXiv:1404.7828 (2014).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* **25** (2012).
5. Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84-90 (2025).
6. Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045-7063 (2024).
7. Yirmiya, E. *et al.* Structure-guided discovery of viral proteins that inhibit host immunity. *Cell* **188**, 1681-1692.e1617 (2025).
8. Seal, S. *et al.* Cell Painting: a decade of discovery and innovation in cellular imaging. *Nature Methods* **22**, 254-268 (2025).
9. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
10. Chung, K. *et al.* Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332-337 (2013).
11. Bae, J. A. *et al.* Functional connectomics spanning multiple areas of mouse visual cortex. *Nature* **640**, 435-447 (2025).
12. Chandrasekaran, S. N. *et al.* Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nat Methods* **21**, 1114-1121 (2024).
13. Bock, C. *et al.* High-content CRISPR screening. *Nature Reviews Methods Primers* **2**, 8 (2022).
14. Schmacke, N. A. *et al.* SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes. *bioRxiv*, 2023.2006.2001.542416 (2023).
15. Pitino, E. *et al.* STAMP: Single-cell transcriptomics analysis and multimodal profiling through imaging. *Cell* **188**, 5100-5117.e5126 (2025).
16. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nature Biotechnology* **40**, 1231-1240 (2022).
17. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41**, 773-782 (2023).
18. Cui, H. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623-633 (2025).
19. Ji, Y. *et al.* Scalable and universal prediction of cellular phenotypes. *bioRxiv*, 2024.2008.2012.607533 (2024).
20. Bussi, Y. & Keren, L. Multiplexed image analysis: what have we achieved and where are we headed? *Nature Methods* **21**, 2212-2215 (2024).
21. Tan, Y. *et al.* SPACEc: A Streamlined, Interactive Python Workflow for Multiplexed Image Processing and Analysis. *bioRxiv*, 2024.2006.2029.601349 (2024).
22. Schapiro, D. *et al.* MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nature Methods* **19**, 311-315 (2022).
23. Kuehl, M. *et al.* Pathology-oriented multiplexing enables integrative disease mapping. *Nature* **644**, 516-526 (2025).
24. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **7**, 16878 (2017).
25. Stirling, D. R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).
26. Muñoz, A. F. *et al.* cp_measure: API-first feature extraction for image-based profiling workflows. *arXiv*, arXiv:2507.01163 (2025).
27. Moore, J. *et al.* OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. *Nature Methods* **18**, 1496-1498 (2021).
28. Klein, D., Uscidda, T., Theis, F. & Cuturi, M. GENOT: Entropic (Gromov) Wasserstein Flow Matching with Applications to Single-Cell Genomics. *arXiv*, arXiv:2310.09254 (2023).
29. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow Matching for Generative Modeling. *arXiv*, arXiv:2210.02747 (2022).
30. Tong, A. *et al.* Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv*, arXiv:2302.00482 (2023).
31. Muhlich, J. L. *et al.* Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613-4621 (2022).
32. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* **18**, 100-106 (2021).
33. Mölder, F. *et al.* Sustainable data analysis with Snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research* **10** (2021).
34. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316-319 (2017).
35. Marconato, L. *et al.* SpatialData: an open and universal data framework for spatial omics. *Nature Methods* **22**, 58-62 (2025).
36. Sofroniew, N. *et al.* napari: a multi-dimensional image viewer for Python.
37. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *JOSS* (2024).
38. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology* **41**, 604-606 (2023).
39. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**, 293-304 (2024).
40. Mah, C. K. *et al.* Bento: a toolkit for subcellular analysis of spatial transcriptomics data. *Genome Biology* **25**, 82 (2024).
41. Serrano, E. *et al.* Reproducible image-based profiling with Pycytominer. *Nature Methods* **22**, 677-680 (2025).
42. Klein, D. *et al.* Mapping cells through time and space with moscot. *Nature* **638**, 1065-1075 (2025).
43. Heumos, L. *et al.* Pertpy: an end-to-end framework for perturbation analysis. *bioRxiv*, 2024.2008.2004.606516 (2024).
44. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* **19**, 171-178 (2022).
45. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology* **40**, 163-166 (2022).
46. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
47. Bajcsy, P. *et al.* Enabling global image data sharing in the life sciences. *Nature Methods* **22**, 672-676 (2025).
48. Liu, Z. *et al.* A ConvNet for the 2020s. *arXiv*, arXiv:2201.03545 (2022).

49. Gupta, A. *et al.* SubCell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv*, 2024.2012.2006.627299 (2024).
50. Kim, V. *et al.* Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports* **15**, 4876 (2025).
51. Rosenberger, F. A. *et al.* Deep Visual Proteomics maps proteotoxicity in a genetic liver disease. *Nature* **642**, 484-491 (2025).
52. Massoni-Badosa, R. *et al.* An atlas of cells in the human tonsil. *Immunity* **57**, 379-399.e318 (2024).
53. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e922 (2019).
54. Klein, D. *et al.* CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025.2004.2011.648220 (2025).
55. Huang, T., Liu, T., Babadi, M., Jin, W. & Ying, R. Scalable Generation of Spatial Transcriptomics from Histology Images via Whole-Slide Flow Matching. *arXiv*, arXiv:2506.05361 (2025).
56. Haviv, D., Pooladian, A.-A., Pe'er, D. & Amos, B. Wasserstein Flow Matching: Generative modeling over families of distributions. *arXiv*, arXiv:2411.00698 (2024).
57. 10x Genomics. Xenium Prime 5K In Situ Gene Expression with Cell Segmentation data for human ovarian cancer (FFPE) using the Xenium Prime 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes (v1), In Situ Gene Expression dataset analyzed using Xenium Onboard Analysis 3.0.0. (2024).
58. Giakoumoglou, N., Stathaki, T. & Gkelias, A. A Review on Discriminative Self-supervised Learning Methods in Computer Vision. *arXiv*, arXiv:2405.04969 (2024).
59. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. *arXiv*, arXiv:2111.06377 (2021).
60. Kraus, O. *et al.* Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology. *arXiv*, arXiv:2404.10242 (2024).
61. Gordon, S. & Martinez, F. O. Alternative activation of macrophages: mechanism and functions. *Immunity* **32**, 593-604 (2010).
62. Hollmén, M., Figueiredo, C. R. & Jalkanen, S. New tools to prevent cancer growth and spread: a 'Clever' approach. *British Journal of Cancer* **123**, 501-509 (2020).
63. Muhl, L. *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nature Communications* **11**, 3953 (2020).
64. Kalluri, R. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer* **16**, 582-598 (2016).
65. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085-1094 (2025).
66. Mathys, H. *et al.* Single-cell multiregion dissection of Alzheimer's disease. *Nature* **632**, 858-868 (2024).
67. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nature Medicine* **29**, 1563-1577 (2023).
68. Huang, K. *et al.* Sequential Optimal Experimental Design of Perturbation Screens Guided by Multi-modal Priors. *bioRxiv*, 2023.2012.2012.571389 (2023).
69. Mahecic, D. *et al.* Event-driven acquisition for content-enriched microscopy. *Nat Methods* **19**, 1262-1267 (2022).
70. Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633-638 (2025).
71. Arevalo, J. *et al.* Evaluating batch correction methods for image-based cell profiling. *Nat Commun* **15**, 6516 (2024).
72. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19**, 41-50 (2022).
73. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).
74. Wolf, T. *et al.* HuggingFace's transformers: State-of-the-art natural language processing. *arXiv* (2019).
75. Falcon, W. & The PyTorch Lightning team. *PyTorch Lightning*, https://www.pytorchlightning.ai (2019).
76. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv*, arXiv:1912.01703 (2019).

Supplementary Figures

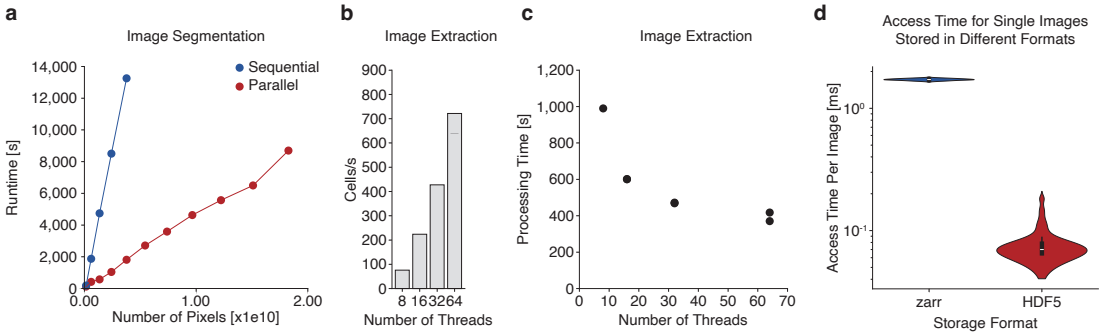


Figure S1 | The parallelization capabilities of scPortrait and its HDF5 backend enables fast image processing and access

a, Runtime comparison of parallel and sequential image segmentation. **b**, Image extraction speed in cells per second over different numbers of concurrent threads. **c**, Image extraction processing time over different numbers of concurrent threads. **d**, Access duration of image datasets locally stored as HDF5 or zarr files. 100 images were accessed per file.

s = seconds

Figure S1

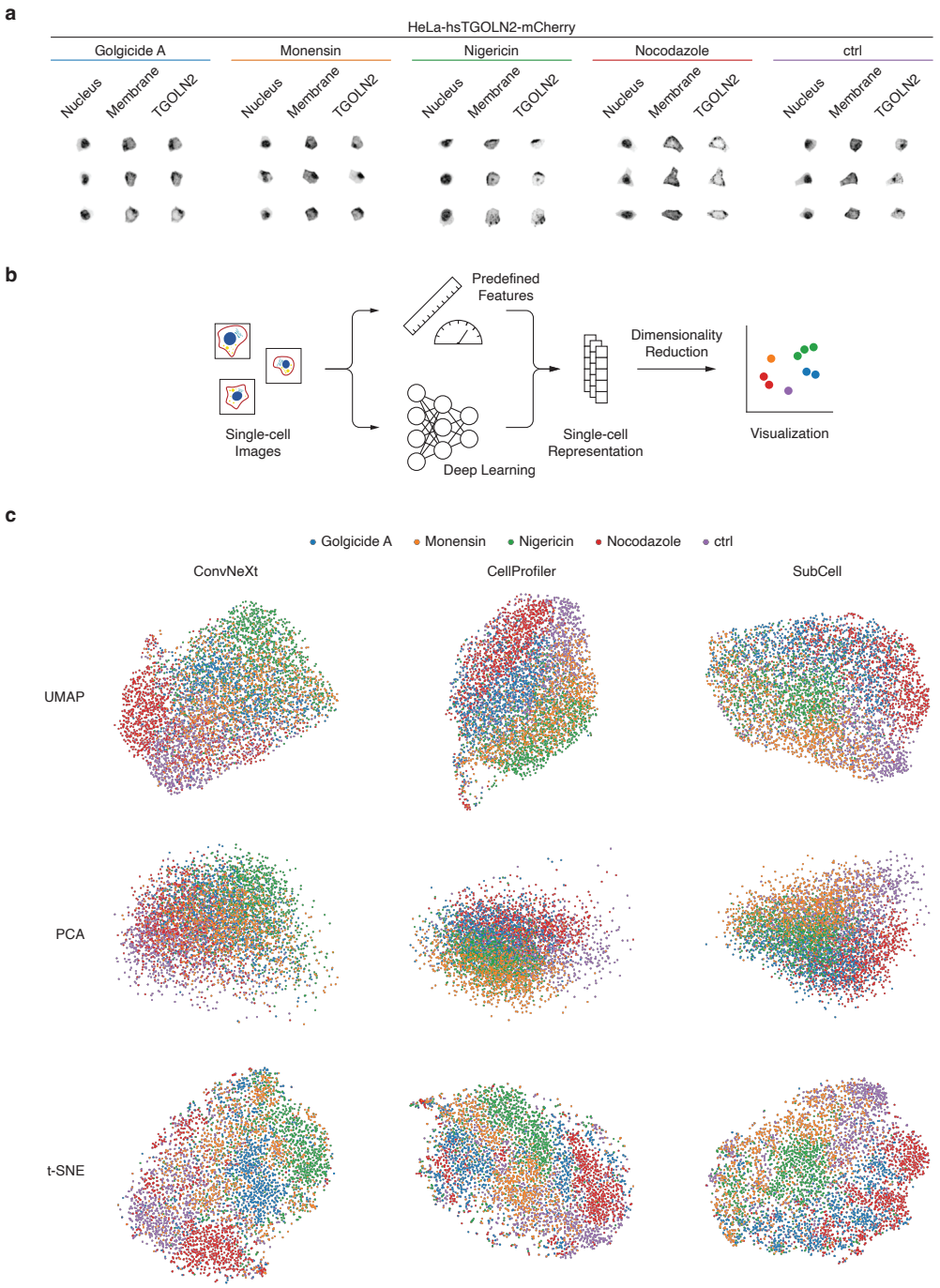


Figure S2

Figure S2 | scPortrait recognizes trans-Golgi network morphologies in combination with different single-cell image embedding strategies

a, Single-cell images of HeLa cells expressing hsTGOLN2-mCherry stimulated with the indicated compounds. **b**, Embedding strategies for images from a. **c**, Three different single-cell image embedders were used to generate representations of images shown in a: ConvNeXt, a CNN trained as a classifier on natural images, CellProfiler, a list of predefined single-cell image features and SubCell, a transformer model trained on human protein atlas images. The representations learned by these models are visualized via three different techniques: t-SNE, UMAP and PCA. Colors indicate chemical perturbations from a.

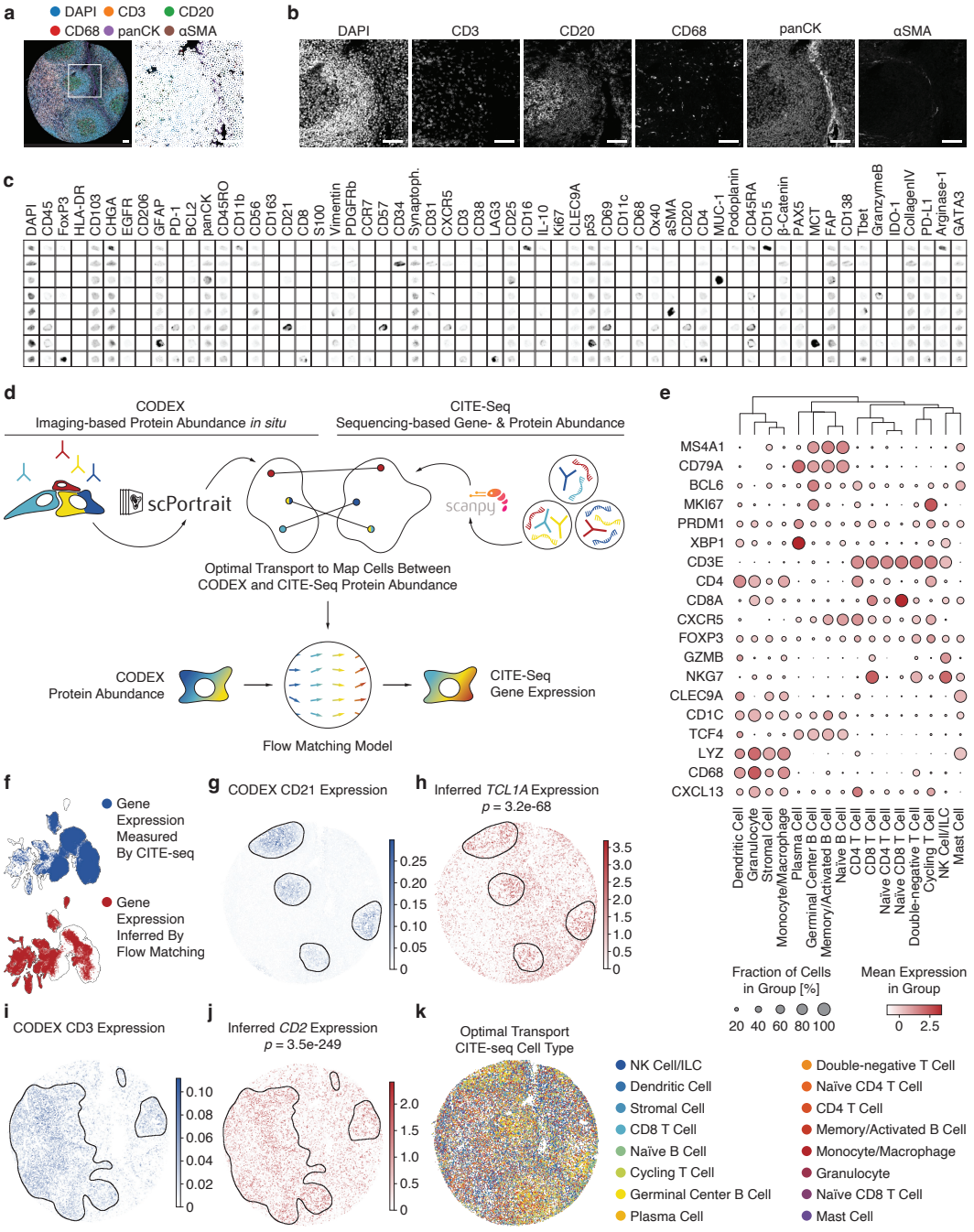


Figure S3

Figure S3 | Optimal Transport matches CODEX imaging of human tonsil with transcriptome data

a, Overview of human tonsil tissue core from a tonsillitis patient stained with 58 antibodies and DAPI using the CODEX assay. Right panels depict magnification of regions indicated on the left. Colors depict selected stains. White outlines indicate cytosol borders determined by nuclear expansion segmentation based on CellPose. Scalebars represent 75 μm. **b**, Individual imaging channels of magnified region from a. Scalebars represent 75 μm. **c**, Example single-cell images following scPortrait extraction showing all 59 channels per cell. "Synaptoph." = Synaptophysin. **d**, Overview of our strategy to match image-based CODEX data of human tonsil with unpaired single-cell transcriptomics data of human tonsil using optimal transport to train a flow matching model. We process the respective data modalities with scPortrait (images) and scanpy (transcriptomics), collapsing images to median channel intensity per cell. We then calculate a probabilistic mapping between modalities using Monge optimal transport. Sampling batches according to this optimal transport coupling between modalities, we then train a flow matching model to generate a cell's gene expressions conditioned on its CODEX profile. **e**, Flow matching-inferred expression of selected marker genes across CITE-seq derived cell types. Cell types were assigned to inferred expression profiles by k-nearest neighbor majority. **f**, UMAP representation of gene expression measured by CITE-seq or inferred by flow matching per cell. Outlines correspond to joint UMAP. **g**, Germinal center marker CD21 expression measured by CODEX in the tonsillitis sample. Black outlines show germinal centers defined by the expression of this marker. **h**, *TCL1A* expression inferred by flow matching in the tonsillitis sample. Black outlines show germinal centers defined by the expression of CD21. **i**, T cell marker CD3 expression measured by CODEX in the tonsillitis sample. Black outlines show T cell zone borders defined by the expression of this marker. **j**, *CD2* expression inferred by flow matching in the tonsillitis sample. Black outlines show T cell zone borders defined by the expression of CD3. **k**, Cell type annotation from dissociated transcriptomics data mapped onto tonsillitis tissue via optimal transport.

p-values correspond to spatially differential gene expression in- and outside of the highlighted zones and were calculated by a two-sided Mann-Whitney U test with Bonferroni correction.

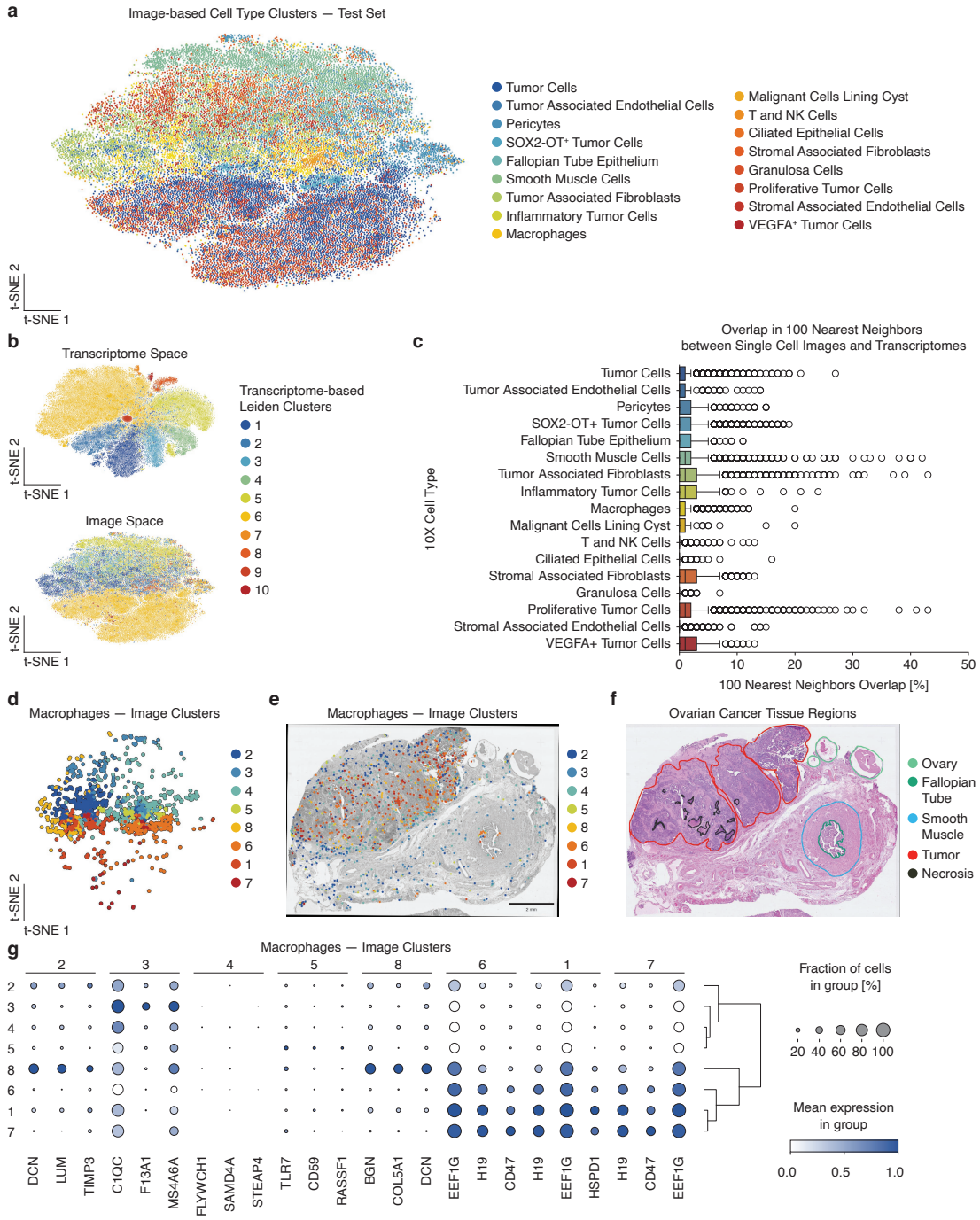


Figure S4

Figure S4 | Cross-modality modeling of spatial transcriptomics and imaging data with scPortrait

a, t-SNE visualization of ViT-MAE embeddings of single-cell images colored by 10x Xenium cell type annotation. Each dot represents one cell. Only test set data are shown. **b**, t-SNE visualization of single-cell transcriptome (top) and ViT-MAE embeddings of single-cell images (bottom) colored by transcriptome-based Leiden clusters. Each dot represents one cell. Only test set data are shown. **c**, Boxplot depicting local neighborhood overlap between transcriptome space and image-based embeddings (b). X-axis shows percent of neighbors of a given cell that are identical between transcriptome space and image-based embeddings. Y-axis shows 10x cell types. **d**, a filtered for macrophages. Colors indicate Leiden clusters. Only test set data are shown. **e**, Distribution of macrophage clusters from d across the tissue region. Only test set data are shown. **f**, Tissue region annotation of ovarian cancer dataset. **g**, Genes differentially expressed in macrophage clusters from d. Only test set data are shown.

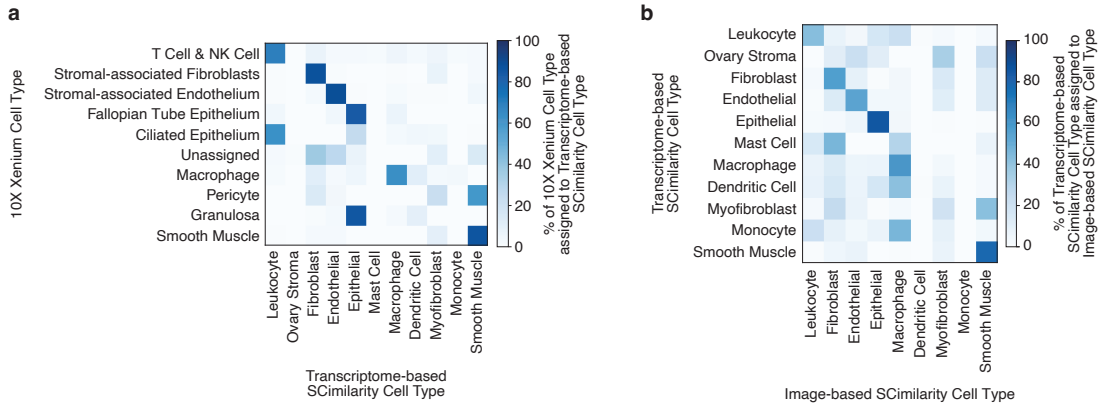


Figure S5 | Embedding of single-cell images into a transcriptome atlas with scPortrait

a, Heatmap showing which cell types cells get assigned after embedding into SCimilarity space based on their transcriptome as a percentage of 10x Xenium cell types. **b**, Heatmap showing which cell types cells get assigned after embedding into SCimilarity space based on their images as a percentage of transcriptome-based SCimilarity cell types.

Figure S5

5 Discussion

Understanding the spatial organisation of tissues and cells provides valuable insights into how life is structured. A key goal is to directly link cellular phenotypes to their underlying genetic and molecular determinants.

Through the development of SPARCS, I established a new platform that allows for the direct investigation of genetic determinants of image-based phenotypes in forward genetic screens (Publication 4.3). This technology scales to whole-genome applications and even allows for the reanalysis of archived screens for additional phenotypes when new computational models become available. As such, it presents a powerful tool for relating cellular phenotypes to their underlying biology. Since SPARCS was designed in an open and accessible manner, and works with standard light microscopy setups, I hope that it will be adopted by scientists working in different biological fields.

To facilitate the characterisation of such image-based phenotypes, I not only developed a computational framework, *scPortrait*, that provides end-to-end processing of raw microscopy images into single-cell image datasets, but also established a standardised format for the storage of such information (Publication 4.5). This framework interfaces with state-of-the-art deep learning frameworks which permits the easy integration of image-based methods to asses cellular composition into machine learning approaches.

In particular, the combination of image-based phenotyping with proteomic characterisation in scDVP (Publication 4.2), allows for the fine-grained molecular characterisation of the composition of individual cells while preserving their spatial information. This technology generates data which lays the groundwork for future models that are able to directly predict molecular composition on the basis of cellular imaging. When combined with *scPortrait*, this opens up many new approaches for the unbiased detection of cellular phenotypes. As demonstrated for AATD (Publication 4.4), this type of analysis paradigm is promising to dissect complex cellular processes in situ across a variety of different proteotoxic diseases.

Throughout this thesis, I established that deep-learning-based characterisation of cellular phenotypes based on microscopy images is a powerful approach to understanding the underlying characteristics of cellular morphology. In a supervised approach, I was able to develop and train a CNN classifier that, using intracellular LC3 distribution, was able to identify autophagy-defective cells with a high degree of confidence. This was crucial for the identification of almost all known autophagy regulators in a single experiment and also facilitated the identification of a novel previously undescribed phenotype (Publication 4.3). Even more promising are unbiased approaches where no a priori knowledge of the types of observable phenotypes is required, as demonstrated by the identification of a terminal hepatocyte state marked by globular protein aggregates, which holds the potential to understand and ultimately counteract molecular mechanisms underlying AATD disease progression (Publication 4.4).

The tools I have developed here, have the potential to advance biological discovery, contributing to a wide range of fields from tissue-based disease mechanisms to the detailed understanding of the role of individual genes in the spatial composition of cells. I have placed a strong focus on making my platforms and software tools easily accessible and open-source. I believe that this is a central cornerstone of modern science and hope that through the development of tools like SPARCS, scDVP and *scPortrait*, I can contribute to the generation of better computational models that are fully able to decode cellular function.

Index of Abbreviations

AAT	Alpha-1 Antitrypsin
AATD	Alpha-1 Antitrypsin Deficiency
Cas	CRISPR associated
CCD	Charge Coupled Device
CNN	Convolutional Neural Network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
DC	Direct Current
DDA	Data-Dependent Acquisition
DIA	Data Independent Acquisition
DNA	Deoxyribonucleic Acid
DSB	Double-Strand Break
DVP	Deep Visual Proteomics
ddNTP	Dideoxynucleotide Triphosphate
dNTP	Deoxynucleotide Triphosphate
ESI	Electrospray Ionisation
GFP	Green Fluorescent Protein
GPU	Graphics Processing Unit
HDR	Homology-Directed Repair

ILSVRC ImageNet Large Scale Visual Recognition Challenge

LC Liquid Chromatography

LC-MS Liquid Chromatography-Mass Spectrometry

lncRNA Long Non-Coding RNA

MLP Multilayer Perceptron

mRNA Messenger RNA

MS Mass Spectrometry

MS/MS Tandem Mass Spectrometry (also known as MS^2)

NHEJ Non-Homologous End Joining

NGS Next-Generation Sequencing

NLP Natural Language Processing

PALM Photo-Activated Localisation Microscopy

PAM Protospacer Adjacent Motif

PCR Polymerase Chain Reaction

PPS Polyphenylene Sulfide

PTM Post-Translational Modification

ReLU Rectified Linear Unit

RF radio frequency

RNA Ribonucleic Acid

rRNA Ribosomal RNA

SPARCS Spatially Resolved CRISPR Screening

scDVP Single-Cell Deep Visual Proteomics

sgRNA single-guide RNA

STED Stimulated Emission Depletion

TALE Transcription Activator-Like Effector

TALEN Transcription Activator-Like Effector Nuclease

TOF Time-of-Flight

ViT Vision Transformer

ZF Zinc Finger

ZFN Zinc Finger Nuclease

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Matthias. His unwavering support and trust have been instrumental through my PhD. I am deeply grateful for his mentorship, which has not only refined my scientific thinking but has also instilled in me a profound appreciation for the rigorous pursuit of knowledge.

I am also immensely thankful to all the members of our lab. Working alongside such a talented and collaborative group of individuals has been a truly enriching experience. The vibrant discussions and the camaraderie within the lab have made this journey both enjoyable and intellectually stimulating. Each member has, in their own way, contributed to the successful completion of this thesis, and for that, I am profoundly grateful.

My heartfelt appreciation extends to my family. Their unwavering faith in my abilities and their constant support have been the foundation upon which this achievement stands. Their patience, understanding, and encouragement have given me the strength to persevere through the most challenging times. I am indebted to them for their unconditional love and sacrifices, which have enabled me to pursue my dreams.

Finally, to my partner Niklas: thank you for being my rock throughout this journey. Without you none of this would have been possible. You have been both my partner, and my mentor. Your boundless patience and emotional support, and your belief in me, especially during my moments of self-doubt, has been a source of unwavering strength and motivation for me. I am profoundly thankful for your love, understanding, and encouragement and am so grateful to have you by my side.

Bibliography

Adams, M. D. et al. (2000). "The Genome Sequence of Drosophila melanogaster". In: *Science* 287.5461, pp. 2185–2195. ISSN: 0036-8075. DOI: 10.1126/science.287.5461.2185.

Aebersold, R. and M. Mann (2016). "Mass-spectrometric exploration of proteome structure and function". In: *Nature* 537.7620, pp. 347–355. ISSN: 0028-0836. DOI: 10.1038/nature19949.

Alberts, B., J. Wilson, and T. Hunt (2008). *Molecular biology of the cell*. English. 5th ed. New York: Garland Science. ISBN: 9780815341055.

Amarasinghe, S. L. et al. (2020). "Opportunities and challenges in long-read sequencing data analysis". In: *Genome Biology* 21.1, p. 30. ISSN: 1474-7596. DOI: 10.1186/s13059-020-1935-5.

Anfinsen, C. B. (1973). "Principles that Govern the Folding of Protein Chains". In: *Science* 181.4096, pp. 223–230. ISSN: 0036-8075. DOI: 10.1126/science.181.4096.223.

Arber, W. (1978). "Restriction Endonucleases". In: *Angewandte Chemie International Edition in English* 17.2, pp. 73–79. ISSN: 0570-0833. DOI: 10.1002/anie.197800733.

Arnott, D., J. Shabanowitz, and D. F. Hunt (1993). "Mass spectrometry of proteins and peptides: sensitive and accurate mass measurement and sequence analysis". In: 39.9, pp. 2005–2010. ISSN: 0009-9147. DOI: 10.1093/clinchem/39.9.2005.

Aston, F. (1920). "XLIV. The constitution of atmospheric neon". In: *Philosophical Magazine Series 6* 39.232, pp. 449–455. ISSN: 1941-5982. DOI: 10.1080/14786440408636058.

Avery, O. T., C. M. MacLeod, and M. McCarty (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types". In: *The Journal of Experimental Medicine* 79.2, pp. 137–158. ISSN: 0022-1007. DOI: 10.1084/jem.79.2.137.

Baltimore, D. (1970). "Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses". In: *Nature* 226.5252, pp. 1209–1211. ISSN: 0028-0836. DOI: 10.1038/2261209a0.

Barrangou, R. et al. (2007). "CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes". In: *Science* 315.5819, pp. 1709–1712. ISSN: 0036-8075. DOI: 10.1126/science.1138140.

Betzig, E. et al. (2006). "Imaging Intracellular Fluorescent Proteins at Nanometer Resolution". In: *Science* 313.5793, pp. 1642–1645. ISSN: 0036-8075. DOI: 10.1126/science.1127344.

Bian, Y., F. P. Bayer, et al. (2021). "Robust Microflow LC-MS/MS for Proteome Analysis: 38000 Runs and Counting". In: *Analytical Chemistry* 93.8, pp. 3686–3690. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.1c00257.

Bian, Y., R. Zheng, et al. (2020). "Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS". In: *Nature Communications* 11.1, p. 157. DOI: 10.1038/s41467-019-13973-x.

Boveri, T. (1902). "Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns". In: *Verhandlungen der Physikalisch-Medizinischen Gesellschaft zu Würzburg* 35. Signatur: 8R 82.280/30, pp. 67–90.

Boyle, W. S. and G. E. Smith (1970). "Charge coupled semiconductor devices". In: *The Bell System Technical Journal* 49.4, pp. 587–593. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1970.tb01790.x.

Brenner, S. (1974). "The Genetics of Caenorhabditis Elegans". In: *Genetics* 77.1, pp. 71–94. ISSN: 0016-6731. DOI: 10.1093/genetics/77.1.71.

Brenner, S., F. Jacob, and M. Meselson (1961). "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis". In: *Nature* 190.4776, pp. 576–581. ISSN: 0028-0836. DOI: 10.1038/190576a0.

Bridges, C. B. and T. H. Morgan (1923). "The third-chromosome group of mutant characters of Drosophila melanogaster". In: DOI: 10.5962/bhl.title.24013.

Brouns, S. J. J. et al. (2008). "Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes". In: *Science* 321.5891, pp. 960–964. ISSN: 0036-8075. DOI: 10.1126/science.1159689.

Brunner, A.-D. et al. (2022). "Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation". In: *Molecular Systems Biology* 18.3, e10798. ISSN: 1744-4292. DOI: 10.15252/msb.202110798.

Caron, M. et al. (2021). "Emerging Properties in Self-Supervised Vision Transformers". In: *arXiv*. DOI: 10.48550/arxiv.2104.14294.

Chang, H. H. Y., N. R. Pannunzio, N. Adachi, and M. R. Lieber (2017). "Non-homologous DNA end joining and alternative pathways to double-strand break repair". In: *Nature Reviews Molecular Cell Biology* 18.8, pp. 495–506. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.48.

Chargaff, E. (1950). "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation". In: *Experientia* 6.6, pp. 201–209. ISSN: 0014-4754. DOI: 10.1007/bf02173653.

– (1951). "Some recent studies on the composition and structure of nucleic acids". In: *Journal of Cellular and Comparative Physiology* 38.S1, pp. 41–59. ISSN: 0095-9898. DOI: 10.1002/jcp.1030380406.

Chargaff, E., R. Lipshitz, C. Green, and M. Hodes (1951). "The Composition Of The Desoxyribonucleic Acid Of Salmon Sperm". In: *Journal of Biological Chemistry* 192.1, pp. 223–230. ISSN: 0021-9258. DOI: 10.1016/s0021-9258(18)55924-x.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *arXiv*. DOI: 10.48550/arxiv.2002.05709.

Chen, X. et al. (2016). "Multi-View 3D Object Detection Network for Autonomous Driving". In: *arXiv*. DOI: 10.48550/arxiv.1611.07759.

Chinwalla, A. T. et al. (2002). "Initial sequencing and comparative analysis of the mouse genome". In: *Nature* 420.6915, pp. 520–562. ISSN: 0028-0836. DOI: 10.1038/nature01262.

Consortium, I. H. G. S. et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062.

Crick, F. H. (1958). "On protein synthesis." In: *Symposia of the Society for Experimental Biology* 12, pp. 138–63. ISSN: 0081-1386.

– (1970). "Central Dogma of Molecular Biology". In: *Nature* 227.5258, pp. 561–563. ISSN: 0028-0836. DOI: 10.1038/227561a0.

Crick, F. H., J. S. Griffith, and L. E. Orgel (1957). "Codes without Commas". In: *Proceedings of the National Academy of Sciences* 43.5, pp. 416–421. ISSN: 0027-8424. DOI: 10.1073/pnas.43.5.416.

Cui, H. et al. (2024). "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". In: *Nature Methods* 21.8, pp. 1470–1480. ISSN: 1548-7091. DOI: 10.1038/s41592-024-02201-0.

Dahm, R. (2008). "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research". In: *Human Genetics* 122.6, pp. 565–581. ISSN: 0340-6717. DOI: 10.1007/s00439-007-0433-0.

Darwin, C. (1951). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. English. F390b. New York: D. Appleton.

Dikic, I. (2016). "Proteasomal and Autophagy Degradation Systems". In: *Annual Review of Biochemistry* 86.1, pp. 1–32. ISSN: 0066-4154. DOI: 10.1146/annurev-biochem-061516-044908.

Dobzhansky, T. G. (1941). *Genetics and the Origin of Species, By Theodosius Dobzhansky*. Columbia University Press.

Doll, S. and A. L. Burlingame (2015). "Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications". In: *ACS Chemical Biology* 10.1, pp. 63–71. ISSN: 1554-8929. DOI: 10.1021/cb500904b.

Dosovitskiy, A., L. Beyer, et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv*. DOI: 10.48550/arxiv.2010.11929.

Dosovitskiy, A., P. Fischer, et al. (2014). "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks". In: *arXiv*. DOI: 10.48550/arxiv.1406.6909.

Dunham, I. et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. ISSN: 0028-0836. DOI: 10.1038/nature11247.

Ellis, H. M. and H. Horvitz (1986). "Genetic control of programmed cell death in the nematode C. elegans". In: *Cell* 44.6, pp. 817–829. ISSN: 0092-8674. DOI: 10.1016/0092-8674(86)90004-8.

Eng, J. K., A. L. McCormack, and J. R. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". In: *Journal of the American Society for Mass Spectrometry* 5.11, pp. 976–989. ISSN: 1044-0305. DOI: 10.1016/1044-0305(94)80016-2.

Esteva, A. et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639, pp. 115–118. ISSN: 0028-0836. DOI: 10.1038/nature21056.

Feng, X. et al. (2019). "Computer vision algorithms and hardware implementations: A survey". In: *Integration* 69, pp. 309–320. ISSN: 0167-9260. DOI: 10.1016/j.vlsi.2019.07.005.

Fenn, J. B. et al. (1989). "Electrospray Ionization for Mass Spectrometry of Large Biomolecules". In: *Science* 246.4926, pp. 64–71. ISSN: 0036-8075. DOI: 10.1126/science.2675315.

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.

Fraenkel-Conrat, H. and R. C. Williams (1955). "Reconstruction of Active Tobacco Mosaic Virus From its Inactive Protein and Nucleic Acid Components*". In: *Proceedings of the National Academy of Sciences* 41.10, pp. 690–698. ISSN: 0027-8424. DOI: 10.1073/pnas.41.10.690.

Franklin, R. E. and R. G. Gosling (1953). "Molecular Configuration in Sodium Thymonucleate". In: *Nature* 171.4356, pp. 740–741. ISSN: 0028-0836. DOI: 10.1038/171740a0.

Gamow, G. (1954). "Possible Relation between Deoxyribonucleic Acid and Protein Structures". In: *Nature* 173.4398, pp. 318–318. ISSN: 0028-0836. DOI: 10.1038/173318a0.

Gamow, G. and M. Yčas (1955). "STATISTICAL CORRELATION OF PROTEIN AND RIBONUCLEIC ACID COMPOSITION". In: *Proceedings of the National Academy of Sciences* 41.12, pp. 1011–1019. ISSN: 0027-8424. DOI: 10.1073/pnas.41.12.1011.

Gasiunas, G., R. Barrangou, P. Horvath, and V. Siksnys (2012). "Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria". In: *Proceedings of the National Academy of Sciences* 109.39, E2579–E2586. ISSN: 0027-8424. DOI: 10.1073/pnas.1208507109.

Gierer, A. and G. Schramm (1956). "Infectivity of Ribonucleic Acid from Tobacco Mosaic Virus". In: *Nature* 177.4511, pp. 702–703. ISSN: 0028-0836. DOI: 10.1038/177702a0.

Gillet, L. C. et al. (2012). "Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis*". In: *Molecular & Cellular Proteomics* 11.6, O111.016717. ISSN: 1535-9476. DOI: 10.1074/mcp.o111.016717.

Goffeau, A. et al. (1996). "Life with 6000 Genes". In: *Science* 274.5287, pp. 546–567. ISSN: 0036-8075. DOI: 10.1126/science.274.5287.546.

Griffith, F. (1928). "The Significance of Pneumococcal Types". In: *Journal of Hygiene* 27.2, pp. 113–159. ISSN: 0022-1724. DOI: 10.1017/s0022172400031879.

Guzman, U. H. et al. (2024). "Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition". In: *Nature Biotechnology*, pp. 1–12. ISSN: 1087-0156. DOI: 10.1038/s41587-023-02099-7.

Haldane, J. (1932). *The causes of evolution*. English. London: Longmans, Green and Co.

Hartwell, L. H., J. Culotti, and B. Reid (1970). "Genetic Control of the Cell-Division Cycle in Yeast, I. Detection of Mutants". In: *Proceedings of the National Academy of Sciences* 66.2, pp. 352–359. ISSN: 0027-8424. DOI: 10.1073/pnas.66.2.352.

Hayden, E. C. (2014). "Technology: The $1,000 genome". In: *Nature* 507.7492, pp. 294–295. ISSN: 0028-0836. DOI: 10.1038/507294a.

He, K., X. Zhang, S. Ren, and J. Sun (2015). "Deep Residual Learning for Image Recognition". In: *arXiv*. DOI: 10.48550/arxiv.1512.03385.

Hell, S. W. and J. Wichmann (1994). "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy". In: *Optics Letters* 19.11, pp. 780–782. DOI: 10.1364/ol.19.000780.

Hendrickson, E. A. (1997). "Cell-Cycle Regulation of Mammalian DNA Double-Strand-Break Repair". In: *The American Journal of Human Genetics* 61.4, pp. 795–800. ISSN: 0002-9297. DOI: 10.1086/514895.

Hershey, A. D. and M. Chase (1952). "Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage". In: *The Journal of General Physiology* 36.1, pp. 39–56. ISSN: 0022-1295. DOI: 10.1085/jgp.36.1.39.

Huang, G., Z. Liu, L. v. d. Maaten, and K. Q. Weinberger (2016). "Densely Connected Convolutional Networks". In: *arXiv*. DOI: 10.48550/arxiv.1608.06993.

Hunkapiller, T., R. J. Kaiser, B. F. Koop, and L. Hood (1991). "Large-Scale and Automated DNA Sequence Determination". In: *Science* 254.5028. doi: 10.1126/science.1925562, pp. 59–67. DOI: 10.1126/science.1925562.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3, pp. 318–356. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(61)80072-7.

Jacob, T. M. and H. G. Khorana (1965). "Studies on Polynucleotides. XLIV. 1 The Synthesis of Dodecanucleotides Containing the Repeating Trinucleotide Sequence Thymidylyl- (3'→5') -thymidylyl- (3'→5') -deoxycytidine 2,3". In: *Journal of the American Chemical Society* 87.13, pp. 2971–2981. ISSN: 0002-7863. DOI: 10.1021/ja01091a029.

Jain, M., R. Abu-Shumays, H. E. Olsen, and M. Akeson (2022). "Advances in nanopore direct RNA sequencing". In: *Nature Methods* 19.10, pp. 1160–1164. ISSN: 1548-7091. DOI: 10.1038/s41592-022-01633-w.

Jain, M., S. Koren, et al. (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nature Biotechnology* 36.4, pp. 338–345. ISSN: 1087-0156. DOI: 10.1038/nbt.4060.

Jain, M., H. E. Olsen, B. Paten, and M. Akeson (2016). "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biology* 17.1, p. 239. ISSN: 1474-7596. DOI: 10.1186/s13059-016-1103-0.

Jinek, M. et al. (2012). "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity". In: *Science* 337.6096, pp. 816–821. ISSN: 0036-8075. DOI: 10.1126/science.1225829.

Johannsen, W. (1909). *Elemente der Exakten Erblichkeitslehre*. Deutsch. Jena: Verlag von Gustav Fischer.

Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. ISSN: 0028-0836. DOI: 10.1038/s41586-021-03819-2.

Kasianowicz, J. J., E. Brandin, D. Branton, and D. W. Deamer (1996). "Characterization of individual polynucleotide molecules using a membrane channel". In: *Proceedings of the National Academy of Sciences* 93.24, pp. 13770–13773. ISSN: 0027-8424. DOI: 10.1073/pnas.93.24.13770.

Kayagaki, N. et al. (2015). "Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling". In: *Nature* 526.7575, pp. 666–671. ISSN: 0028-0836. DOI: 10.1038/nature15541.

Khorana, H. G. et al. (1966). "Polynucleotide Synthesis and the Genetic Code". In: *Cold Spring Harbor Symposia on Quantitative Biology* 31.0, pp. 39–49. ISSN: 0091-7451. DOI: 10.1101/sqb.1966.031.01.010.

Kim, V. et al. (2024). "Self-supervision advances morphological profiling by unlocking powerful image representations". In: *bioRxiv*, p. 2023.04.28.538691. DOI: 10.1101/2023.04.28.538691.

Kim, Y. G., J. Cha, and S. Chandrasegaran (1996). "Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain." In: *Proceedings of the National Academy of Sciences* 93.3, pp. 1156–1160. ISSN: 0027-8424. DOI: 10.1073/pnas.93.3.1156.

Kossel, A. (1911). *Ueber die chemische Beschaffenheit des Zellkerns (Nobel-Vortrag).* Vol. 58. München Med. Wochenschrift.

Krishna, R. et al. (2024). "Generalized biomolecular modeling and design with RoseTTAFold All-Atom". In: *Science* 384.6693, eadl2528. ISSN: 0036-8075. DOI: 10.1126/science.adl2528.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". English. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. ISSN: 0018-9219. DOI: 10.1109/5.726791.

Leeuwenhoek, A. V. (1997). "Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused". In: *Philosophical Transactions of the Royal Society of London* 12.133. doi: 10.1098/rstl.1677.0003, pp. 821–831. DOI: 10.1098/rstl.1677.0003.

Lehman, I. R. (1974). "DNA Ligase: Structure, Mechanism, and Function". In: *Science* 186.4166, pp. 790–797. ISSN: 0036-8075. DOI: 10.1126/science.186.4166.790.

Levene, P. A. and W. A. Jacobs (1909). "Über die Hefe-Nucleinsäure". In: *Berichte der deutschen chemischen Gesellschaft* 42.2, pp. 2474–2478. ISSN: 0365-9496. DOI: 10.1002/cber.190904202148.

Levene, P. and E. London (1929). "The Structure Of Thymonucleic Acid". In: *Journal of Biological Chemistry* 83.3, pp. 793–802. ISSN: 0021-9258. DOI: 10.1016/s0021-9258(18)77108-1.

Liang, F., M. Han, P. J. Romanienko, and M. Jasin (1998). "Homology-directed repair is a major double-strand break repair pathway in mammalian cells". In: *Proceedings of the National Academy of Sciences* 95.9, pp. 5172–5177. ISSN: 0027-8424. DOI: 10.1073/pnas.95.9.5172.

Ligon, W. V. (1979). "Molecular Analysis by Mass Spectrometry". In: *Science* 205.4402, pp. 151–159. ISSN: 0036-8075. DOI: 10.1126/science.205.4402.151.

Lin, T., Y. Wang, X. Liu, and X. Qiu (2021). "A Survey of Transformers". In: *arXiv*. DOI: 10.48550/arxiv.2106.04554.

Liu, Z. et al. (2022). "A ConvNet for the 2020s". In: *arXiv*. DOI: 10.48550/arxiv.2201.03545.

Luckey, J. A. et al. (1990). "High speed DNA sequencing by capillary electrophoresis". In: *Nucleic Acids Research* 18.15, pp. 4417–4421. ISSN: 0305-1048. DOI: 10.1093/nar/18.15.4417.

Ludwig, C. et al. (2018). "Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial". In: *Molecular Systems Biology* 14.8, e8126. ISSN: 1744-4292. DOI: 10.15252/msb.20178126.

Lykke-Andersen, S. and T. H. Jensen (2015). "Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes". In: *Nature Reviews Molecular Cell Biology* 16.11, pp. 665–677. ISSN: 1471-0072. DOI: 10.1038/nrm4063.

Mädler, S. C. et al. (2025). "scPortrait integrates single-cell images into multimodal modeling". In: *bioRxiv*. DOI: 10.1101/2025.09.22.677590.

Makarov, A. (2000). "Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis". In: *Analytical Chemistry* 72.6, pp. 1156–1162. ISSN: 0003-2700. DOI: 10.1021/ac991131p.

Mann, M., C. K. Meng, and J. B. Fenn (1989). "Interpreting mass spectra of multiply charged ions". In: *Analytical Chemistry* 61.15, pp. 1702–1708. ISSN: 0003-2700. DOI: 10.1021/ac00190a023.

Mardis, E. R. (2011). "A decade's perspective on DNA sequencing technology". In: *Nature* 470.7333, pp. 198–203. ISSN: 0028-0836. DOI: 10.1038/nature09796.

Marvin, M. (1961). "Microscopy apparatus".

Mattick, J. S. et al. (2023). "Long non-coding RNAs: definitions, functions, challenges and recommendations". In: *Nature Reviews Molecular Cell Biology* 24.6, pp. 430–447. ISSN: 1471-0072. DOI: 10.1038/s41580-022-00566-8.

McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133. ISSN: 0007-4985. DOI: 10.1007/bf02478259.

Meier, F. et al. (2018). "BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes". In: *Nature Methods* 15.6, pp. 440–448. ISSN: 1548-7091. DOI: 10.1038/s41592-018-0003-5.

Mendel, G. (1866). *Versuche über Pflanzen-Hybriden.*

Michaelis, A. C. et al. (2023). "The social and structural architecture of the yeast protein interactome". In: *Nature* 624.7990, pp. 192–200. ISSN: 0028-0836. DOI: 10.1038/s41586-023-06739-5.

Miescher, F. (1871). "Über die chemische Zusammensetzung der Eiterzellen". In: *Medizinisch-Chemische Untersuchungen.*

Miller, J. C., M. C. Holmes, et al. (2007). "An improved zinc-finger nuclease architecture for highly specific genome editing". In: *Nature Biotechnology* 25.7, pp. 778–785. ISSN: 1087-0156. DOI: 10.1038/nbt1319.

Miller, J. C., S. Tan, et al. (2011). "A TALE nuclease architecture for efficient genome editing". In: *Nature Biotechnology* 29.2, pp. 143–148. ISSN: 1087-0156. DOI: 10.1038/nbt.1755.

Minsky, M. (1988). "Memoir on inventing the confocal scanning microscope". In: *Scanning* 10.4, pp. 128–138. ISSN: 0161-0457. DOI: 10.1002/sca.4950100403.

Miyawaki, A. (2011). "Proteins on the move: insights gained from fluorescent protein technologies". In: *Nature Reviews Molecular Cell Biology* 12.10, pp. 656–668. ISSN: 1471-0072. DOI: 10.1038/nrm3199.

Mondello, L. et al. (2023). "Comprehensive two-dimensional liquid chromatography". In: *Nature Reviews Methods Primers* 3.1, p. 86. DOI: 10.1038/s43586-023-00269-0.

Moore, J. K. and J. E. Haber (1996). "Cell Cycle and Genetic Requirements of Two Pathways of Nonhomologous End-Joining Repair of Double-Strand Breaks in Saccharomyces cerevisiae". In: *Molecular and Cellular Biology* 16.5, pp. 2164–2173. ISSN: 0270-7306. DOI: 10.1128/mcb.16.5.2164.

Moore, R., A. Chandrahas, and L. Bleris (2014). "Transcription Activator-like Effectors: A Toolkit for Synthetic Biology". In: *ACS Synthetic Biology* 3.10, pp. 708–716. ISSN: 2161-5063. DOI: 10.1021/sb400137b.

Morgan, T. H. (1910). "Sex Limited Inheritance in Drosophila". In: *Science* 32.812, pp. 120–122. ISSN: 0036-8075. DOI: 10.1126/science.32.812.120.

Morgan, T. H. (1911). "The Origin of Nine Wing Mutations in Drosophila". In: *Science* 33.848, pp. 496–499. ISSN: 0036-8075. DOI: 10.1126/science.33.848.496.

Morgan, T. H. (1919). *Contributions to the genetics of Drosophila melanogaster ...* Washington: Carnegie Institution of Washington.

Mullis, K. B. and F. A. Faloona (1987). "[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction". In: *Methods in Enzymology* 155, pp. 335–350. ISSN: 0076-6879. DOI: 10.1016/0076-6879(87)55023-6.

Mund, A. et al. (2021). "AI-driven Deep Visual Proteomics defines cell identity and heterogeneity". In: *bioRxiv*, p. 2021.01.25.427969. DOI: 10.1101/2021.01.25.427969.

Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of Shotgun Proteomic Data". In: *Molecular & Cellular Proteomics* 4.10, pp. 1419–1440. ISSN: 1535-9476. DOI: 10.1074/mcp.r500012-mcp200.

Nirenberg, M., T. Caskey, et al. (1966). "The RNA Code and Protein Synthesis". In: *Cold Spring Harbor Symposia on Quantitative Biology* 31.0, pp. 11–24. ISSN: 0091-7451. DOI: 10.1101/sqb.1966.031.01.008.

Nirenberg, M. and P. Leder (1964). "RNA Codewords and Protein Synthesis". In: *Science* 145.3639, pp. 1399–1407. ISSN: 0036-8075. DOI: 10.1126/science.145.3639.1399.

Nirenberg, M. W. and J. H. Matthaei (1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". In: *Proceedings of the National Academy of Sciences* 47.10, pp. 1588–1602. ISSN: 0027-8424. DOI: 10.1073/pnas.47.10.1588.

Nishimura, S. et al. (1965). "Studies on polynucleotides XLVII. The in vitro synthesis of homopeptides as directed by a ribopolynucleotide containing a repeating trinucleotide sequence. New codon sequences for lysine, glutamic acid and arginine". In: *Journal of Molecular Biology* 13.1, pp. 283–301. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(65)80097-3.

Nurk, S. et al. (2022). "The complete sequence of a human genome". In: *Science* 376.6588, pp. 44–53. ISSN: 0036-8075. DOI: 10.1126/science.abj6987.

Nüsslein-Volhard, C. and E. Wieschaus (1980). "Mutations affecting segment number and polarity in Drosophila". In: *Nature* 287.5785, pp. 795–801. ISSN: 0028-0836. DOI: 10.1038/287795a0.

Parnas, O. et al. (2015). "A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks". In: *Cell* 162.3, pp. 675–686. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.06.059.

Paul, W. and H. Steinwedel (1953). "Notizen: Ein neues Massenspektrometer ohne Magnetfeld". In: *Zeitschrift für Naturforschung A* 8.7, pp. 448–450. ISSN: 0932-0784. DOI: 10.1515/zna-1953-0710.

Pfaendler, R., J. Hanimann, S. Lee, and B. Snijder (2023). "Self-supervised vision transformers accurately decode cellular state heterogeneity". In: *bioRxiv*, p. 2023.01.16.524226. DOI: 10.1101/2023.01.16.524226.

Poplin, R. et al. (2018). "A universal SNP and small-indel variant caller using deep neural networks". In: *Nature Biotechnology* 36.10, pp. 983–987. ISSN: 1087-0156. DOI: 10.1038/nbt.4235.

Preissler, S. and E. Deuerling (2012). "Ribosome-associated chaperones as key players in proteostasis". In: *Trends in Biochemical Sciences* 37.7, pp. 274–283. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2012.03.002.

Press release: The Nobel Prize in Chemistry 2002 - NobelPrize.org (2024).

Press release: The Nobel Prize in Chemistry 2020 - NobelPrize.org (2024).

Prober, J. M. et al. (1987). "A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides". In: *Science* 238.4825. doi: 10.1126/science.2443975, pp. 336–341. DOI: 10.1126/science.2443975.

Ran, F. A. et al. (2013). "Genome engineering using the CRISPR-Cas9 system". In: *Nature Protocols* 8.11, pp. 2281–2308. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.143.

Rhie, A. et al. (2021). "Towards complete and error-free genome assemblies of all vertebrate species". In: *Nature* 592.7856, pp. 737–746. ISSN: 0028-0836. DOI: 10.1038/s41586-021-03451-0.

Robertson, L. A. (2023). "Antoni van Leeuwenhoek 1723–2023: a review to commemorate Van Leeuwenhoek's death, 300 years ago". In: *Antonie van Leeuwenhoek* 116.10, pp. 919–935. ISSN: 0003-6072. DOI: 10.1007/s10482-023-01859-4.

Ronneberger, O., P. Fischer, and T. Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv*. DOI: 10.48550/arxiv.1505.04597.

Rosenberger, F. A., S. C. Mädler, et al. (2025). "Deep Visual Proteomics maps proteotoxicity in a genetic liver disease". In: *Nature* 642.8067, pp. 484–491. ISSN: 0028-0836. DOI: 10.1038/s41586-025-08885-4.

Rosenberger, F. A., M. Thielert, et al. (2023). "Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome". In: *Nature Methods* 20.10, pp. 1530–1536. ISSN: 1548-7091. DOI: 10.1038/s41592-023-02007-6.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6, pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519.

Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton (Project Para)*. English. Buffalo: Cornell Aeronautical Laboratory.

– (1960). "Perceptron Simulation Experiments". In: *Proceedings of the IRE* 48.3, pp. 301–309. ISSN: 0096-8390. DOI: 10.1109/jrproc.1960.287598.

– (1961). "Principles Of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms". In: DOI: 10.21236/ad0256582.

Rowley, M. J. and V. G. Corces (2018). "Organizational principles of 3D genome architecture". In: *Nature Reviews Genetics* 19.12, pp. 789–800. ISSN: 1471-0056. DOI: 10.1038/s41576-018-0060-8.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536. ISSN: 0028-0836. DOI: 10.1038/323533a0.

Saleh-Gohari, N. and T. Helleday (2004). "Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells". In: *Nucleic Acids Research* 32.12, pp. 3683–3688. ISSN: 0305-1048. DOI: 10.1093/nar/gkh703.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467. ISSN: 0027-8424. DOI: 10.1073/pnas.74.12.5463.

Schäfer, E. A. (1897). *The Cell in Development and Inheritance*. 2d ed., rev. & enl. Vol. 55. Nature. New York: Macmillan. DOI: 10.1038/055530a0.

Schmacke, N. A., S. C. Mädler, et al. (2023). "SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes". In: *bioRxiv*, p. 2023.06.01.542416. DOI: 10.1101/2023.06.01.542416.

Schmacke, N. A., F. O'Duill, et al. (2022). "IKKβ primes inflammasome formation by recruiting NLRP3 to the trans-Golgi network". In: *Immunity* 55.12, 2271–2284.e7. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2022.10.021.

Shalem, O. et al. (2014). "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells". In: *Science* 343.6166, pp. 84–87. ISSN: 0036-8075. DOI: 10.1126/science.1247005.

Shi, J. et al. (2015). "Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death". In: *Nature* 526.7575, pp. 660–665. ISSN: 0028-0836. DOI: 10.1038/nature15514.

Shimomura, O., F. H. Johnson, and Y. Saiga (1962). "Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea". In: *Journal of Cellular and Comparative Physiology* 59.3, pp. 223–239. ISSN: 0095-9898. DOI: 10.1002/jcp.1030590302.

Simonyan, K. and A. Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv*. DOI: 10.48550/arxiv.1409.1556.

Sinha, A. and M. Mann (2020). "A beginner's guide to mass spectrometry–based proteomics". In: *The Biochemist* 42.5, pp. 64–69. ISSN: 0954-982X. DOI: 10.1042/bio20200057.

Söll, D. et al. (1965). "Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids." In: *Proceedings of the National Academy of Sciences* 54.5, pp. 1378–1385. ISSN: 0027-8424. DOI: 10.1073/pnas.54.5.1378.

Song, C., C. Ye, Y. Sim, and D. S. Jeong (2024). "Hardware for Deep Learning Acceleration". In: *Advanced Intelligent Systems*. ISSN: 2640-4567. DOI: 10.1002/aisy.202300762.

Song, J. G. and J. W. Lee (2023). "CNN-Based Object Detection and Distance Prediction for Autonomous Driving Using Stereo Images". In: *International Journal of Automotive Technology* 24.3, pp. 773–786. ISSN: 1229-9138. DOI: 10.1007/s12239-023-0064-z.

Spitz, F. and E. E. M. Furlong (2012). "Transcription factors: from enhancer binding to developmental control". In: *Nature Reviews Genetics* 13.9, pp. 613–626. ISSN: 1471-0056. DOI: 10.1038/nrg3207.

Steen, H. and M. Mann (2004). "The abc's (and xyz's) of peptide sequencing". In: *Nature Reviews Molecular Cell Biology* 5.9, pp. 699–711. ISSN: 1471-0072. DOI: 10.1038/nrm1468.

Stevens, N. M. (1905). *Studies in spermatogenesis with espeical reference to the acccessory chromosome*. English. Washington: Carnegie Institution of Washington. DOI: 10.5962/bhl.title.23606.

– (1906). *Studies in Spermatogenesis. Part II: A comparative study of the heterochromosomes in certain species of coleoptera, hemiptera and lepidoptera, with especial reference to sex determination*. Washington: Carnegie Institution of Washington.

Stewart, H. I. et al. (2023). "Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis". In: *Analytical Chemistry* 95.42, pp. 15656–15664. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.3c02856.

Stringer, C., T. Wang, M. Michaelos, and M. Pachitariu (2021). "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature Methods* 18.1, pp. 100–106. ISSN: 1548-7091. DOI: 10.1038/s41592-020-01018-x.

Sun, Y., N. B. Agostini, S. Dong, and D. Kaeli (2019). "Summarizing CPU and GPU Design Trends with Product Data". In: *arXiv*. DOI: 10.48550/arxiv.1911.11313.

Sutton, W. S. (1902). "On the morphology of the chromosome group in Brachystola magna". In: *Biological Bulletin* 14.4, pp. 24–39.

– (1903). "The Chromosomes In Heredity". In: *The Biological Bulletin* 4.5. doi: 10.2307/1535741, pp. 231–250. ISSN: 0006-3185. DOI: 10.2307/1535741.

Szegedy, C. et al. (2014). "Going Deeper with Convolutions". In: *arXiv*. DOI: 10.48550/arxiv.1409.4842.

Tan, M. and Q. V. Le (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *arXiv*. DOI: 10.48550/arxiv.1905.11946.

Temin, H. M. and S. Mizutani (1970). "Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus". In: *Nature* 226.5252, pp. 1211–1213. ISSN: 0028-0836. DOI: 10.1038/2261211a0.

The Nobel Prize in Physiology or Medicine 1995 - NobelPrize.org (2024).

The Nobel Prize in Physiology or Medicine 2001 - NobelPrize.org (2024).

The Nobel Prize in Physiology or Medicine 2002 - NobelPrize.org (2024).

Theodoris, C. V. et al. (2023). "Transfer learning enables predictions in network biology". In: *Nature* 618.7965, pp. 616–624. ISSN: 0028-0836. DOI: 10.1038/s41586-023-06139-9.

Thielert, M. et al. (2023). "Robust dimethyl-based multiplex-DIA doubles single-cell proteome depth via a reference channel". In: *Molecular Systems Biology* 19.9, e11503. ISSN: 1744-4292. DOI: 10.15252/msb.202211503.

Thomson, J. J. (1897). "XL. Cathode Rays ". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.269. doi: 10.1080/14786449708621070, pp. 293–316. ISSN: 1941-5982. DOI: 10.1080/14786449708621070.

– (1921). *Rays of positive electricity and their application to chemical analyses*. London: Longmans, Green, and co.,

Tran, J. C. et al. (2011). "Mapping intact protein isoforms in discovery mode using top-down proteomics". In: *Nature* 480.7376, pp. 254–258. ISSN: 0028-0836. DOI: 10.1038/nature10575.

Valen, D. A. V. et al. (2016). "Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments". In: *PLoS Computational Biology* 12.11, e1005177. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1005177.

Vaswani, A. et al. (2017). "Attention Is All You Need". In: *arXiv*. DOI: 10.48550/arxiv.1706.03762.

Venable, J. D. et al. (2004). "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra". In: *Nature Methods* 1.1, pp. 39–45. ISSN: 1548-7091. DOI: 10.1038/nmeth705.

Venter, J. C. et al. (2001). "The Sequence of the Human Genome". In: *Science* 291.5507, pp. 1304–1351. ISSN: 0036-8075. DOI: 10.1126/science.1058040.

Wallmann, G. et al. (2024). "AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics". In: *bioRxiv*, p. 2024.05.28.596182. DOI: 10.1101/2024.05.28.596182.

Wang, J. Y. and J. A. Doudna (2023). "CRISPR technology: A decade of genome editing is only the beginning". In: *Science* 379.6629, eadd8643. ISSN: 0036-8075. DOI: 10.1126/science.add8643.

Wang, T., J. J. Wei, D. M. Sabatini, and E. S. Lander (2014). "Genetic Screens in Human Cells Using the CRISPR-Cas9 System". In: *Science* 343.6166, pp. 80–84. ISSN: 0036-8075. DOI: 10.1126/science.1246981.

Waterston, R. and J. Sulston (1995). "The genome of Caenorhabditis elegans." In: *Proceedings of the National Academy of Sciences* 92.24, pp. 10836–10840. ISSN: 0027-8424. DOI: 10.1073/pnas.92.24.10836.

Watson, J. D. and F. H. Crick (1953a). "Genetical Implications of the Structure of Deoxyribonucleic Acid". In: *Nature* 171.4361, pp. 964–967. ISSN: 0028-0836. DOI: 10.1038/171964b0.

– (1953b). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356, pp. 737–738. ISSN: 0028-0836. DOI: 10.1038/171737a0.

Wheeler, D. A. et al. (2008). "The complete genome of an individual by massively parallel DNA sequencing". In: *Nature* 452.7189, pp. 872–876. ISSN: 0028-0836. DOI: 10.1038/nature06884.

Whitehouse, C. M., R. N. Dreyer, M. Yamashita, and J. B. Fenn (1985). "Electrospray interface for liquid chromatographs and mass spectrometers". In: *Analytical Chemistry* 57.3, pp. 675–679. ISSN: 0003-2700. DOI: 10.1021/ac00280a023.

Wiedenheft, B., S. H. Sternberg, and J. A. Doudna (2012). "RNA-guided genetic silencing systems in bacteria and archaea". In: *Nature* 482.7385, pp. 331–338. ISSN: 0028-0836. DOI: 10.1038/nature10886.

Wolff, M. M. and W. E. Stephens (1953). "A Pulsed Mass Spectrometer with Time Dispersion". In: *Review of Scientific Instruments* 24.8, pp. 616–617. ISSN: 0034-6748. DOI: 10.1063/1.1770801.

Workman, R. E. et al. (2019). "Nanopore native RNA sequencing of a human poly(A) transcriptome". In: *Nature Methods* 16.12, pp. 1297–1305. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0617-2.

Wright, S. (1931). "Evolution In Mendelian Populations". In: 16.2, pp. 97–159. ISSN: 1943-2631. DOI: 10.1093/genetics/16.2.97.

Zeng, W.-F. et al. (2022). "AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics". In: *Nature Communications* 13.1, p. 7238. DOI: 10.1038/s41467-022-34904-3.

Zhang, J., Y. Fei, L. Sun, and Q. C. Zhang (2022). "Advances and opportunities in RNA structure experimental determination and computational modeling". In: *Nature Methods* 19.10, pp. 1193–1207. ISSN: 1548-7091. DOI: 10.1038/s41592-022-01623-y.

Zuylen, J. v. (1981). "The microscopes of Antoni van Leeuwenhoek". In: *Journal of Microscopy* 121.3, pp. 309–328. ISSN: 0022-2720. DOI: 10.1111/j.1365-2818.1981.tb01227.x.