Establishing blood-based classifiers for decentralized and privacy-preserving clinical disease prediction

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

in der Fachrichtung Molekulare Biomedizin

vorgelegt von

Stefanie Warnat-Herresthal

aus

Trier

Bonn, April 2023

Angefertigt mit Genehmigung und nach den Richtlinien der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Joachim L. Schultze
2. Gutachter: Prof. Dr. Jan Hasenauer

Tag der Promotion: 12.01.2024

Erscheinungsjahr: 2024

I Table of Contents

II Index of Figures

III List of abbreviations

AI	Artificial Intelligence
AML	Acute Myeloid Leukemia
ALL	Acute Lymphocytic Leukemia
BMP	Basic Metabolic Panel
CBC	Complete Blood Count
CDM	Common Data Model
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resource
GP	General Practitioner
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine Learning
ODHSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OOD	Out of distribution
PBMC	Peripheral Blood Mononuclear Cells
PPV	Positive Predictive Value
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
SNP	Single-Nucleotide Polymorphism
TB	Tuberculosis
WHO	World Health Organization

IV Abstract

Big data is increasingly being generated across all sectors of medicine and high-dimensional omics data have been shown to be a powerful data space for AI applications that can assist in primary and differential diagnostics, therapeutic decision making, disease outcome prediction and clinical planning, as well as for an improved understanding of disease pathophysiology. Particularly blood transcriptomics was shown to be informative for a variety of machine learning tasks while at the same time being a technology that is clinically feasible.

This thesis first exemplifies the enormous potential of leveraging AI for medical diagnosis by demonstrating that accurate prediction of Acute Myeloid Leukemia is possible on a large reference dataset across more than 120 independent studies covering two microarray platforms as well as bulk RNA-sequencing data. A large range of clinically relevant scenarios are evaluated, addressing several bottlenecks on the path towards clinical deployment of this technology. I provide evidence that accurate prediction of AML by near-automated machine learning algorithms based on high-dimensional omics data is possible. At the same time, translation of these possibilities into clinical practice is not happening yet due to limitations which are inherent to the health care sector. Mainly, aggregation of datasets which are sufficiently large to train robust classifiers that are generalizable across sites is not feasible due to strict regulations on data privacy. To overcome this obstacle, I introduce Swarm Learning in my thesis as a new framework for collaborative and decentral machine learning based on blockchain technology. I am showing that SL enables training of accurate classification algorithms which outperform local models across a wide range of scenarios predicting AML, tuberculosis, COVID-19 and lung diseases utilizing even different medical data spaces, while at the same time preserving data privacy.

Consequently, I propose a new approach to be included into the development of routine diagnostics in the health care sector, which utilizes the power of machine learning for physicians to extract information from high-dimensional molecular data for data-driven prediction of disease or therapy outcome, thereby augmenting clinical decision making and ultimately leading to a more precise approach to medicine, while at the same time acknowledging data privacy and the need for democratic structures in collaborative machine learning across the health care sector.

1. Introduction & Background

The capabilities of artificial intelligence (AI) to accurately detect patterns in almost any type of medical data ranging from image data to electronics health records (EHR), pulmonary function tests, data from ECG records or high-dimensional data as produced by genome and transcriptome technologies are immense (Rajkomar et al. 2018; Topalovic et al. 2019; Attia et al. 2019; Libbrecht and Noble 2015; Eraslan et al. 2019). While the possibilities of AI to potentially support various aspects of primary and differential diagnostics, patient morbidity or mortality risk prediction, prognosis, formulation of treatment recommendations and clinical decision making are widely appreciated (Rajpurkar et al. 2022; Acosta et al. 2022; Obermeyer and Emanuel 2016), the utilization of AI systems for clinical routine diagnostic methodologies is still in its infancy and the long list of publications on the subject in form of academic articles, health policy reports, statements from professional societies, and popular media coverage stands in stark contrast to the low number of applications that are already used in routine medical care (Meskó and Görög 2020; Benjamens, Dhunnoo, and Meskó 2020). At the same time however, high-dimensional omics technologies have been established and have shown to be very informative in human health and disease conditions (Rood et al. 2022) and particularly blood transcriptomics is known to provide a powerful data space to characterize a multitude of conditions (Zak et al. 2016; Altman et al. 2021; Ulas et al. 2020). The aim of this thesis is (1) to provide evidence on whether transcriptome datasets can be utilized for ML-based diagnostic purposes and (2) to propose a new concept of collaborative machine learning in combination with high-dimensional medical data that could make it possible to utilize the power of AI in medicine in a powerful, effective, and clinically feasible way.

In this introduction, I put the topics of my theses into a broader perspective about how algorithm-based applications shape medicine in general and medical diagnostics in particular. Furthermore, the role of omics technologies, particularly blood transcriptomics, in future precision medicine approaches is highlighted (1.1). Next, the molecular and technical basis of transcriptomic technologies are briefly described (1.2). Third, developments and basic terminology of artificial intelligence and machine learning are introduced with emphasis on algorithms that are suitable for the transcriptomic data space (1.3). Lastly, medical data characteristics such as decentralized structures and strict privacy regulations are described, which need to be considered in any AI framework suitable for medical use cases, and the respective terminology is clarified (1.4).

1.1. The route towards artificial intelligence in medical diagnostics

Artificial intelligence (AI)-guided technology is fundamentally shaping modern society in many regards and is supporting critical data interpretation and decision-making processes in politics, research, media and economics. Consequently, this general shift towards computer-aided data interpretation is also altering biomedical research and healthcare in an unprecedented fashion and AI-based applications have the potential to influence every area and aspect of medicine from clinical routines to laboratory diagnostics. The route towards AI-guided medicine is considered to be the “most significant transformation for healthcare in generations" (E. J. Topol 2022) and roles and functions of clinical staff are expected to fundamentally change in all professions (E. Topol 2019).

This implies that biomedical research as well as applied medicine are becoming increasingly data-intensive and technology-driven. Large volumes of biomedical data are produced at high speed of access and analysis, with substantial data heterogeneity across producing sites and data types (Price and Cohen 2019). At the same time, the whole field is exploring the armamentarium of computer science, mathematics, and computational modeling to turn this wealth of data into knowledge and utilize it for medical applications. Consequently, statistics and machine learning have become essential tools to find descriptive patterns and to utilize such data to answer biomedical questions. Examples of applications include the usage of machine learning tools to identify previously unknown subtypes of diseases in genomic, transcriptomic and epigenomic data (Dai et al. 2022; Nakauma-González et al. 2022; Cancer Genome Atlas Research Network 2015), to identify new therapeutic targets in proteome data (Piazza et al. 2020) or to link biochemical pathways to a disease in a multi-omics setting (Frost and Amos 2018).

1.1.1. AI for primary diagnostics

This work focuses on the usage of AI for diagnostic purposes and various predictive algorithms have been proposed to potentially assist practitioners in clinical decision making (Rajpurkar et al. 2022). There are, however, different diagnostic requirements depending on the concrete medical sector that need to be considered. In the context of primary diagnostics at the general practitioner (GP), disease prevalence is usually rather low compared to the situation at medical specialists. Patients at the GP often present with milder forms of disease and rather nonspecific symptoms, which leads to a diagnostic uncertainty that is considered inherent to general medical practice (Wübken, Oswald, and Schneider 2013). At the same time, early diagnosis

and referral of the patient to a specialized clinician are crucial for a better prognosis and lower mortality in severe diseases, as it could be modeled by the consequences of systemic delay in cancer diagnostics during the COVID-19 pandemic (Sud et al. 2020; Maringe et al. 2020). Therefore, GPs are expected to identify individuals with severe diseases early on while they also need to avoid excessive testing or inappropriate onward referrals (Summerton and Cansdale 2019).

Consequently, standard diagnostic tests that are performed at the GP are rather broad and include tests to evaluate the overall health of patients and to check for a variety of conditions, such as anemia, infection, and inflammation. One common test is the complete blood count (CBC) (George-Gay and Parker 2003), which measures the levels of red blood cells, white blood cells and platelets in the blood, as well as hemoglobin and hematocrit levels. Another example is the basic metabolic panel (BMP) (Kildow et al. 2018), which measures several substances in the blood, such as sodium, potassium, and calcium, as well as glucose and kidney function markers like creatinine and blood urea nitrogen, which are generally considered parameters for body function and provide a broad and unspecific picture of a patient's overall health condition. These standard blood tests are very easy-to-use, quick and inexpensive and pose relatively little risk of harm to the patient. However, despite their usability for the large "problem space" that general practitioners face (Knottnerus 1991), the efficacy of such non-specific markers for early detection of severe diseases could not be demonstrated (Boland, Wollan, and Silverstein 1996; S Rüttimann and Clémençon 1994; Sigmund Rüttimann 1992; Alpert, Greiner, and Hall 2004). Even when all tested values turn out to be in the normal range, none of these tests should generally be used as a "rule out" test for malignant diseases (J. Watson et al. 2019). Furthermore, individual blood tests that test for single disease-specific biomarkers only have a low positive predictive value (PPV) and high false-positive rates in low-risk populations such as those seen by GPs (J. Watson et al. 2019).

Interestingly, with the advent of machine learning technologies, it became apparent that routine blood test results contain much more information than commonly recognized in primary care. In contrast to human evaluation of those tests, where physicians tend to pay particular attention mostly to values outside of a particular reference range (Luo et al. 2016), machine learning algorithms are able to detect non-obvious and latent relationships in the data and subtle correlated deviations in measured parameters. It has been shown that parameters collected from standard blood tests can be used to detect e.g. hematologic diseases (Gunčar et al. 2018) and liver fibrosis (Blanes-Vidal et al. 2022), but also infectious diseases like COVID-19 (Kukar et

al. 2021) as well as brain tumors with a reported diagnostic accuracy that is comparable to that of neuroimaging studies (Podnar et al. 2019). First products are on the market that make use of this, e.g., the SABS software, a clinical decision support system that interprets blood test results and expands differential diagnostics (Smart Blood Analytics 2023b) and mobile applications such as mySmartBlood (Smart Blood Analytics 2023a), which makes it possible for laypersons to enter their blood test results and with that determine their most likely groups of diseases.

Similarly, other directions have been proposed on how AI can enhance primary care decision making and to develop screening tests. For example, risk prediction by modeling on Electronic Health Record (EHR) data has been shown to be helpful for scheduling primary care interventions in the ambulatory setting (Lin, Mahoney, and Sinsky 2019). Also, several ideas are being proposed on how for instance smartphone technology can help to make health metrics accessible outside of specialized settings, like recording and translating heart sounds with smartphones (Bae et al. 2022) or using a smartphone camera to detect eye diseases (Babenko et al. 2022). This is particularly relevant for developing nations and remote areas, where AI-driven health interventions outside of specialized settings might accelerate the achievement of health-related sustainable development goals (Schwalbe and Wahl 2020). Taken together, the use of AI to support primary care is seen to have great potential in improving diagnostic accuracy and efficiency, to reduce risks to patients safety associated with human frailties such as cognitive bias in primary diagnostics (Summerton and Cansdale 2019) and to ultimately lead to a more precise treatment (E. Topol 2019), which takes into account a much more holistic space of disease parameters than current approaches. However, besides first landmark demonstrations, the translation of such data-driven concepts to routine primary care is still challenging and considered a largely unfulfilled opportunity today (Rajpurkar et al. 2022).

1.1.2 AI in medical specialties

In contrast to the demand for accurate screening tests suitable for primary health care described above, medical specialists usually require differential diagnostics. Patients have been referred to medical specialists by general practitioners and usually show more severe and specific symptoms. Consequently, diagnostics performed at medical specialists are usually aimed at clearly defined hypotheses about the type of disease and the predictive value of diagnostics tests is generally higher in such a specialized setting due to the higher prevalence (Knottnerus 1991). Additionally, specific companion diagnostics may be performed that aim to identify patients who are most likely to benefit from a particular therapeutic intervention.

Medical disciplines that heavily rely on image interpretation have particularly profited from AI-based applications already with most FDA approved technologies having been developed for Radiology, Cardiology and Internal Medicine (Benjamens, Dhunnoo, and Meskó 2020). Data spaces which have been studied extensively to be used for AI-guided diagnosis in specialized medical settings are for example computed tomography (CT) images (P. Huang et al. 2019), optical coherence tomography (OCT) images (Kermany et al. 2018), X-ray images (X. Wang et al. 2017), and images of histopathological slides (Echle et al. 2021), which have also been proposed to be used for companion diagnostics in prostate cancer (Leo et al. 2021). Neural networks have been shown to perform equally well or better than oncologists in the detection of skin cancer (Esteva et al. 2017), breast cancer (McKinney et al. 2020; Wu et al. 2020) or lung cancer (Ardila et al. 2019) on such images and they can furthermore be used for risk prediction (P. Huang et al. 2019; Elshafeey et al. 2019), assisting in treatment decisions (G. Kaissis et al. 2019) and for assessing tumor composition and prognosis (Fu et al. 2020) and it's genomic characterization (Lu et al. 2019). Also for gastroenterology, it has been demonstrated how AI-guided systems can assist in detection of adenoma (Schauer et al. 2022) and randomized controlled studies have been performed (Gong et al. 2020; P. Wang et al. 2020), which show how these technologies can have a quantifiable effect in real-world settings. Additionally, advances have been made in the field of ophthalmology, where retinal fungus imaging can be used to detect diabetic retinopathy (Abràmoff et al. 2013), which was further demonstrated in a pivotal trial (Abràmoff et al. 2018). In fact, a diabetic retinopathy diagnostic system became the first fully autonomous AI-based system approved for marketing in the USA (Keane and Topol 2018; FDA 2018).

Besides image interpretation, there are many different types of health data and corresponding technological platforms which were shown to be usable in AI-based applications with the goal to improve and accelerate medical diagnostics. Medical texts can be screened by natural language processing (J. Lee et al. 2020), and medical signal data such as electroencephalograms (EEG) (Claassen et al. 2019) or electrocardiograms (ECG) (Porumb et al. 2020; Attia et al. 2019) can be used to predict medical outcomes (Rajpurkar et al. 2022). Furthermore, wearables can be used to record medical health data and predict disease (Sabry et al. 2022) and microphone and speakers of smartphones can be used e.g. to detect the presence of middle ear fluid, a key diagnostic marker for common pediatric ear diseases (Chan et al. 2019).

1.1.3. The role of omics technologies and systems medicine

Many of the above-mentioned applications have the great benefit that they mostly rely on technologies which are already being present at clinical sites, like e.g. CT imaging, which makes an add-on application of trained AI models for evaluation of this data relatively easy to implement, provided that the metadata for these data is available in machine-readable formats according to international standards such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) by the Observational Health Data Sciences and Informatics (ODHSI) program (Sivesind et al. 2022) or the Fast Healthcare Interoperability Resource (FHIR) (Ayaz et al. 2021). However, given the huge technological advances that have been made in molecular biomedicine (Rajewsky et al. 2020; Rood et al. 2022) and the potential of utilizing such big data to transform clinical medicine (Shilo, Rossman, and Segal 2020), current applications of AI in diagnostics particularly based on omics data are still in their infancy. While present diagnostic practice usually relies on single or few markers in diagnostic tests which are evaluated manually, high-throughput technologies for DNA- and RNA-profiling, epigenomics, metabolomics and proteomics have been introduced which make it possible to access almost all biological layers of a biological sample with high resolution. Most importantly, these data have been shown to be highly informative for a multitude of disease conditions (Apweiler et al. 2018; Aronson and Rehm 2015; Davis, Tato, and Furman 2017; Rajewsky et al. 2020). Systems Medicine approaches combine these technological advances with expertise from biology, biostatistics, informatics, mathematics, and computational modeling to extract previously inaccessible knowledge from high-dimensional omics data in order to enhance clinical decision making (Apweiler et al. 2018). Such approaches do not only have the potential to improve diagnostic sensitivity, but could also enable a more holistic understanding of disease cause, precise therapeutic targeting and outcome risk prediction (Rood et al. 2022), which goes beyond “simple” screenings for individual diseases by taking into account a multitude of parameters (Ashley 2016).

Particularly whole blood as well as peripheral blood mononuclear cells (PBMC) have been shown to be especially well-suited for data-driven assessments of molecular phenotypes in human health and disease (Chaussabel 2015) and compared to other sources of biological material that can be used for system medicine approaches, blood comes with several advantages. First, blood is easily accessible and taking blood is considered a procedure that poses relatively little risk of harm to patients (Wisser et al. 2003), which enables longitudinal sampling before, during and after any medical intervention (Brodin, Duffy, and

Quintana-Murci 2019). Second, robust sampling protocols are in place for molecular profiling of blood, which can be implemented on a large scale both within and outside of clinical settings (Mahajan et al. 2016; Speake et al. 2017), Third, the blood circulatory system connects every organ and every tissue throughout the body and thus signals of health and disease are transferred via the blood and can often be measured directly, such as cell-free DNA (cfDNA), which is known to be present in raised concentrations in patients with cancer, but also other conditions (Wan et al. 2017) as well as direct detection of pathogens in infectious diseases by DNA or RNA sequencing (Ko et al. 2019; Miller et al. 2013). Finally, blood captures dynamic systemic immunological responses in patients (Chaussabel, Pascual, and Banchereau 2010). In other words, not only can infectious diseases be identified by targeting the infectious agents, but also the biological processes underpinning infection and host response can be monitored using molecular profiles of the blood. This provides possibilities for more precise treatment decisions as well as outcome prediction and disease subtyping in a wide range of disease conditions and a large body of literature has been published on blood transcriptome profiling in health and disease. Examples include studies on viral (Ulas et al. 2020; Hou et al. 2014), bacterial and fungal infections (Mahajan et al. 2016; Piasecka et al. 2018) and transcriptomic profiling of various non-infectious conditions such as autoimmune diseases (Acquaviva et al. 2020) or cancer (Dumeaux et al. 2015; Nichita et al. 2014). The information content of the blood compartment in respect to biological heterogeneity can also be dissected in higher granularity by single-cell RNA sequencing in health and disease, and large studies have been performed on infections (Oelen et al. 2022; Schulte-Schrepping et al. 2020a) and chronic inflammatory and autoimmune diseases (Perez et al. 2022) for example.

This vision of a blood-based molecular diagnosis can be formulated as to have a future “Complete Blood Count 2.0”, meaning an omics-based test which would measure many more parameters than in current approaches and by that providing a high-resolution portrait of the molecular profiles of nucleated blood cells (Rood et al. 2022). Results of such a test could then be compared against known blood profiles of known diseases and conditions and would provide a comprehensive view of the status of the immune network (Bonaguro, Schulte-Schrepping, Ulas, et al. 2022). Such a concept would potentially be applicable to all mentioned diagnostic tasks, such as screening tests in primary care (Montgomery, Bernstein, and Wheeler 2022), as well as disease subtyping (Chen et al. 2020) and companion diagnostics (Mellors et al. 2020) and therapy response prediction (Sammut et al. 2022) in specialized hospital settings.

1.2 Biological and technological background of omics technologies

1.2.1 Molecular biomedicine: From DNA-sequencing to systems biology

As introduced above, high-throughput omics technologies such as whole genome DNA- and RNA-profiling are integral part of any systems-level approach to medicine. While there is no binding definition of "omics" technologies, the word affix goes back to the Latin word "omnis", everything. Accordingly, all of these technologies aim to expand the number of measured parameters from single genes, proteins or metabolites, to the entirety of all genes, proteins or metabolites in a given sample (Vogeser and Bendt 2023).

To put this into context, modern molecular biomedicine had its foundation in the middle of the last century (Schrödinger 1944) and starting from the discovery of the double-helical structure of DNA (J. D. Watson and Crick 1953), major technological milestones were achieved in the following decades, which were prerequisite for today's standard high-throughput technologies, such as the development of a technique to systematically determine a sequence of DNA (Sanger, Nicklen, and Coulson 1977) and the invention of PCR (Mullis et al. 1986). Following that, the Human Genome Project was initiated and the first draft of the human genome was published in 2001 (International Human Genome Sequencing Consortium et al. 2001), a landmark achievement which can be seen as a starting point for any further "omics" technologies. Since then, DNA sequencing has become one of the most influential technologies in biomedical research and large studies on population genetics have been conducted with the aim to improve our understanding of health and disease by investigating the combined effects of genetic and environmental influences (Allen et al. 2014; All of Us Research Program Investigators et al. 2019). However, most diseases that manifest in clinical phenotypes are not attributable to single genetic variants alone, and relationships between genetic variations and phenotypic traits, large genetic association studies (GWAS) are conducted to identify susceptibility loci for different diseases (Severe Covid-19 GWAS Group et al. 2020). However, this is often not straightforward and association signals tend to be spread across most of the genome, which include many genes without an obvious connection to disease (Boyle, Li, and Pritchard 2017). In contrast to Mendelian or monogenic diseases which are mostly caused by changes in protein-coding regions, complex diseases are often associated with non-coding variants that affect gene regulation (Y. I. Li et al. 2016). For example, variants that are causally linked to autoimmune diseases mostly map to immune enhancers (Farh et al. 2015), which are important regulatory elements for gene expression. Furthermore, common single-nucleotide polymorphisms (SNPs) with small effect sizes account for genetic variance in many

traits (Boyle, Li, and Pritchard 2017; H. Shi, Kichaev, and Pasaniuc 2016) and a large meta-analysis on height has shown that over 12,000 independent SNPs are significantly associated with that trait, which accounts for almost all of the SNP-based heredity (Yengo et al. 2022). This implies that, to understand how phenotypic alterations in health and disease can be causally linked to biological mechanisms, it is important to consider not only genomic alterations, but also which genomic information is actively transcribed into RNA and finally translated to protein, which is in the end serving a biological function and constitutes a particular phenotype. This process of gene expression cannot be attributed to genetic information alone, it is rather the "interpretation" of genomic information by the cell, which is highly context-dependent and involves heavy regulation at all levels from RNA transcription to RNA-splicing, translation of mRNA to protein and finally post-translational modifications of proteins (Figure 1).

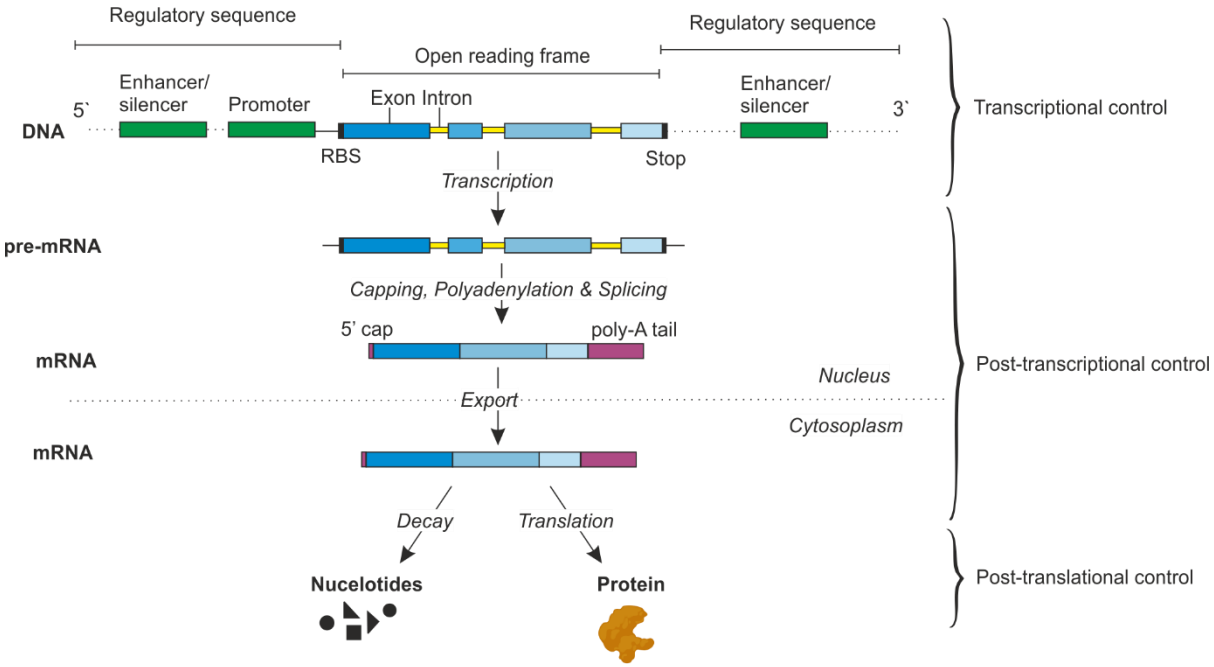


***Figure 1**: Schematic representation of gene expression from DNA to mRNA to protein. The upper part shows the general structure of the DNA of a protein-coding gene, including its 5' and 3' regulatory elements, transcription start and stop sites as well as exons and introns. Nucleotides of the open reading frame are transcribed to pre-mRNA, which is further processed to mRNA. mRNA is exported to the cytosol, where it is either translated to protein or subjected to decay mechanisms. Gene expression is regulated at transcription from DNA to mRNA, at post-transcriptional and at post-translational level. Figure adapted from* (Halbeisen et al. 2008).

One particularly informative snapshot of this complex process is the transcriptome, which describes the complete set of ribonucleic acid (RNA) molecules that are present at a given time within an organism, tissue, or a single cell (Zhong Wang, Gerstein, and Snyder 2009). In contrast to the genome, which is incredibly stable over replications, the transcriptome changes dynamically in response to various factors like environmental influences, developmental stages, or disease conditions (Magnuson, Bedi, and Ljungman 2016). RNA types present in the cell are protein-coding messenger RNA (mRNA) and non-coding RNAs, which participate in transcription, RNA processing (e.g., small nuclear RNAs) and translation (e.g., ribosomal RNAs, transfer RNAs, microRNAs) as well as other RNAs that are involved in other processes and whose functions are less well characterized or unknown (Cooper, Wan, and Dreyfuss 2009).

Especially mRNA molecules are of great interest since they are transient intermediary molecules in the process of translating genetic information to biological function: Protein coding genes are first transcribed into pre-mRNA, which is then spliced to remove introns, resulting in mature mRNA, which will finally be translated into protein products (Figure 1). The information which mRNA molecules are present in a biological sample can therefore be seen as a direct surrogate of active gene expression. Depending on the transcriptome technology and data processing method being used, mRNA quantification can even be performed on splice variant level (Zhong Wang, Gerstein, and Snyder 2009). In addition, long non-coding RNAs (lncRNA) have been studied with increased interest in the last years and are known to be crucial for gene regulation. They affect gene expression in different biological and pathophysiological contexts, for example by modulation of chromatin function or by altering the stability and translation of cytoplasmic mRNA (Statello et al. 2021). Capturing the full set of coding and non-coding RNA molecules in a biological sample can therefore serve as a snapshot of the cell's biological state, be it in steady state or in response to environmental or pathophysiological conditions, which can also be used for diagnostic purposes. Particularly circulating micro RNAs (miRNAs), a group of lncRNAs that suppress gene expression both by inhibiting protein translation and promoting mRNA cleavage, have been indicated to be attractive biomarkers for disease diagnosis and prognosis in cancer and parkinsons disease, for example (Ravanidis et al. 2020; Asakura et al. 2020; Sharifi, Talkhabi, and Taleahmad 2022).

Different technologies exist that can measure gene expression levels in a given sample. One fundamental difference between technologies is whether these approaches target a predefined set of RNA molecules, such as in DNA microarrays, or whether in principle all RNA molecules

can be detected and therefore also unknown RNA molecules can be found, such as in next generation RNA sequencing. Today, most transcriptome data is generated using RNA sequencing. However much of the data that has been generated during the last decades is based on DNA microarray technology, which can be repurposed and re-analyzed also for the applications discussed in this work. Therefore, DNA microarrays are shortly introduced here.

1.2.2 DNA Microarrays for targeted detection of mRNA

DNA microarrays are a group of technologies that assess gene expression levels by measuring the amount of fluorescently labeled target nucleotides which hybridize to a set of DNA-probes on a surface. To measure a sample with a DNA microarray, the RNA has to be isolated and to be enriched for mRNA. This mRNA is either labeled directly or converted to cDNA and then labeled (Bumgarner 2013). Different methods exist for labeling cDNA (Richter et al. 2002), but usually fluorochrome dyes are used. The labeled cDNA is then hybridized to the probes on the microarray. These are short DNA sequences usually between 25 and 60 bps and each probe targets a short region of a specific transcript ("Microarray Probe Mapping" 2023). Molecules that successfully bind to the plate emit light, which can be measured. Since the location of each probe on the chip is known, the fluorescence intensity can be used as a surrogate for the transcript abundance of the target transcript. This can be done for an arbitrary number of probes. For example, the Affymetrix U133 2.0 chip consists of 604,258 probes covering 17,271 gene symbols (Robinson and Speed 2007). To analyze microarray data, the scanned microarray image needs to be converted into quantifiable values, which can be used for downstream computing. These are usually stored in a binary format (e.g. a CEL file) or as a text file. After a quality control and a correction for background fluorescence, the data is normalized to control for technical variation between arrays, e.g. by Robust Multichip Average (RMA) (Irizarry et al. 2003). The resulting data is usually a table with genes in rows and samples in columns, and each value represents an intensity value.

After the technology was introduced in the 1990s (Lockhart et al. 1996), DNA microarrays have been important to generate first transcriptomic datasets on a larger scale (Su et al. 2002, 2004). However, reliability and consistency issues were raised regarding microarray data and their potential application in clinical and regulatory settings (Marshall 2004; Miklos and Maleszka 2004). This led to the creation of microarray standards, quality measures (MAQC Consortium et al. 2006) and a consensus on data analysis for the development and validation of predictive models (L. Shi et al. 2010). Accordingly, a large body of clinical research has

been proposed using transcriptome data generated by microarrays, including prediction of diseases such as multiple myeloma (Kuiper et al. 2012; Zhan et al. 2007) or breast cancer (van ’t Veer et al. 2002; Zemmour et al. 2015) and the discovery (Bullinger et al. 2004; Alizadeh et al. 2000) and prediction (Andersson et al. 2007) of disease subtypes.

1.2.3 RNA sequencing

Today, the use of microarray-based transcriptome technologies has largely been replaced by next-generation RNA sequencing (RNA-seq). This method not only quantifies known transcripts but also identifies alternatively spliced genes, detects allele-specific expression and unannotated transcripts (Zhong Wang, Gerstein, and Snyder 2009). Compared to microarrays, RNA-seq also has a wider dynamic range of detecting transcripts, which makes it the method of choice for recent clinical transcriptomic studies (Van den Berge et al. 2019). In addition, single-cell sequencing techniques (Potter 2018) have made it possible to study the gene expression of individual cells, which is of great advantage when studying disease biology at the cellular level (Schulte-Schrepping et al. 2020b; Rood et al. 2022; Oelen et al. 2022) and developing disease- and cell type-specific predictive signatures (Bernardes et al. 2020). Furthermore, it is possible to couple RNA-seq with methods to measure genomic, epigenomic and proteomic profiles thereby generating multi-omics data, also at single-cell resolution (Vandereyken et al. 2023).

RNA-seq protocols are manifold and can differ in the RNA species they detect. While most protocols aim at detecting polyadenylated mRNA (Van den Berge et al. 2019), it is also possible to investigate non-polyadenylated RNA species, such as certain lncRNAs (Kukurba and Montgomery 2015), which are known to be important regulatory elements in a variety of diseases (DiStefano 2018). During the procedure, RNA is extracted and specific RNA species are selected either by enriching polyA-containing transcripts or by depleting highly abundant ribosomal RNA (Zhao et al. 2018). Then, the selected RNA is fragmented, converted into cDNA, sequencing adapters are added and the cDNA is amplified to construct a library for sequencing. For Illumina sequencing, libraries are loaded onto a flow cell, where complementary oligonucleotides bind to cDNA molecules. Then, sequencing by synthesis (Ju et al. 2006) is performed with the primary sequencing output being binary per-cycle BCL basecall files which store base calls and quality for each sequencing cycle. These files are then translated into FASTQ files containing the base sequences of the library fragments, which correspond to the transcribed RNA molecules of the original specimen, along with

corresponding quality scores. In order to retain gene expression information from those fragments of sequences, they need to be mapped to a reference genome or transcriptome, depending on the application (Van den Berge et al. 2019). Following this, the relative expression levels of each feature (transcript or gene) are quantified, which results in a count matrix that summarizes expression values per sample.

The major focus of this work is the usage of transcriptomic data for disease prediction in a clinical setting, which could eventually become routine diagnostics pipelines for infective as well as non-transmittable diseases. A large body of work has been using transcriptomic profiling to develop RNA-based diagnostic and prognostic biomarker panels, for infectious (de Araujo et al. 2016; Hou et al. 2014; Thompson et al. 2017) and non-infectious diseases like cancer (Bhalla et al. 2017; Sanchez and Mackenzie 2020; Biswas et al. 2019; Abdul Aziz et al. 2016) or diabetes (Y. Wang, Wang, and Zhang 2018). Transcriptomic data was shown to be applicable for predictions of disease outcome (Xuan Liu et al. 2017; Abdul Aziz et al. 2016), for detection and characterization of disease subtypes based on molecular features (Figgett et al. 2019; Ulas et al. 2020; Ben Azzouz et al. 2021; Y.-R. Liu et al. 2016), for predicting drug responses (Carraro et al. 2022; Preuer et al. 2018; Cherlin et al. 2020), and also to model quasi loss- or gain-of-function experiments based solely on transcriptomic data (Bonaguro, Schulte-Schrepping, Carraro, et al. 2022). To clarify how such applications relate to artificial intelligence and machine learning, I will next introduce the respective terminology and provide background on the basic concepts of AI.

1.3 Artificial Intelligence methods for medical diagnostics

1.3.1 Defining Artificial Intelligence and Machine Learning

Research on artificial intelligence (AI) has been influenced by various disciplines with contributions from mathematics, philosophy, neuroscience, economics, psychology and computer engineering. Consequently, AI is not a monolithic term, but describes a broad, interdisciplinary and dynamic field of research as well as practical applications that build on it. Today, artificial intelligence is generally considered to be both theory and development of computer systems which are able to perform tasks that normally require human intelligence (Figure 2) (Lidströmer, Aresu, and Ashrafian 2022; Toosi et al. 2021).

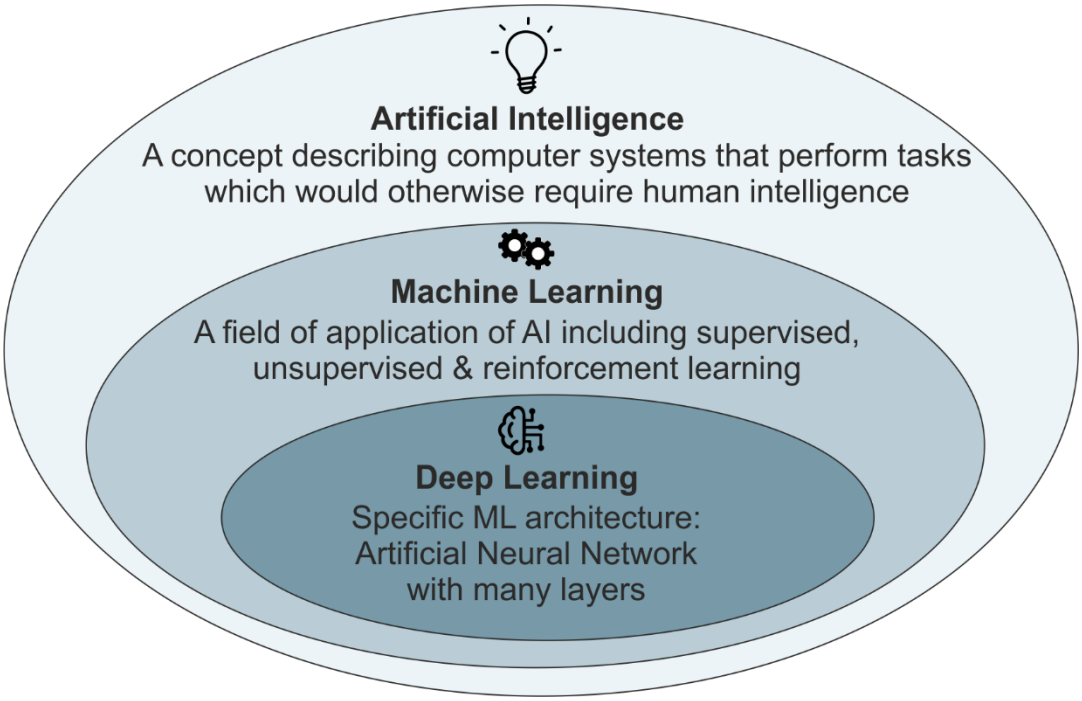


Figure 2: *Terminology as used within this thesis. Artificial Intelligence is a broad concept describing both theory and application of computer systems which can perform tasks that normally require human intelligence. Machine Learning is an application of AI, which allows machines to extract specific patterns from data autonomously. Deep learning is a specific architecture of ML that uses artificial neural networks with more than two layers.* (Lidströmer, Aresu, and Ashrafian 2022)

While the idea of non-human artificial or mechanistic intelligence builds upon concepts and myths that reach way back to ancient Greece (Nilsson 2009), one of the first influential modern theoretical concepts of AI was initiated by Alan Turing. He proposed that a machine should be considered intelligent if a human interrogator, when having a prolonged conversation with the machine via written questions and responses, could not tell the machine apart from a human being (Turing 1950). This idea of a machine with human-like behavior has also been extended to other forms of human-machine interactions and has inspired parts of the research on robotics and computer vision, leading to another definition of machine intelligence as "acting humanely" (S. J. Russell 2010). Taking this understanding of AI one step further, "general" or "strong" AI would ultimately be a human-like system which can reason independently of external human input, can apply knowledge and skills in different contexts, plan for the future and have self-consciousness. This idea of AI benchmarked by human-like performance has been inspirational for the philosophical debate about the nature of self-consciousness and human intelligence (Bringsjord and Govindarajulu 2022; Müller 2021) and is a commonly covered theme in public discussions about AI ("Can We Stop AI Outsmarting Humanity? " 2019), especially after the recent introduction of multimodal large language models like GPT-4, in which some already see sparks of strong AI (Bubeck et al. 2023) and call on a temporary

moratorium before developing even more powerful systems (Future of Life Institute 2023). However, whether a truly human-like or "general" AI is at all reachable is subject to debate (Butz 2021) and goes beyond the scope of this introduction.

In contrast to that, the great successes of AI systems in ubiquitous applications today rely on algorithms that have been designed for well-defined and structured tasks. Here, computers can easily outperform human cognition, especially when it comes to large and high-dimensional data. Even though various AI-related research such as robotics and natural language processing has been inspired by the goal of mimicking human behavior, this is not the core function of AI applications. In that sense, all currently used AI is "narrow" or "weak", since it works rationally within a set of predefined algorithms and it cannot just veer off that programmed path (S. Russell and Norvig 2016). Nevertheless, it does involve sophisticated algorithms that learn patterns from high-dimensional data, can solve complex problems and perform certain well-defined tasks more efficiently than humans.

There has been a significant development in algorithmic and computational techniques since the early applications of artificial intelligence (Schneider 2022). One of the first approaches for clinical usability of AI was to hard-code scientific knowledge into formal language. With this, a computer could reason automatically within statements and use formal inference rules to answer questions or solve problems posed to it. Such "expert systems" were designed for very specific tasks, for example to identify unknown chemical compounds (B. Buchanan and Sutherland 1969). In the 1970s, Stanford University developed a system to identify bacteria which cause an infection and to recommend antibiotics with dosages according to each individual's body weight (B. G. Buchanan and Shortliffe 1984). In principle, such logic-based approaches were sought to be well-suited for medical decision making, since clinical knowledge and decision-trees are usually very structured to prevent malpractice. For example, the clinical decision tree for the diagnosis of diabetes can be displayed in a rule-based decision tree with solely if-then conditions, which takes as input the result of a fasting blood sugar test (Buchard and Richens 2022). However, solely knowledge-based approaches were not largely successful, since they turned out to be useful only in very narrowly defined tasks and rely on explicit, structured knowledge, which requires human experts to encode their knowledge in a set of rules.

The difficulties of such systems suggested that AI systems needed not to rely on hard-coded information but to acquire their own knowledge by extracting patterns from data. This can be

accomplished by algorithms of machine learning (ML), which automatically identify such patterns from data and are able to make predictions based on past data. The terminology is often not clear-cut and AI and ML are used as synonyms in many cases or in combination. However, ML is usually considered a subset of AI techniques as well as a general term to describe applications in which computers learn from data, which is how the terminology is used throughout this work (see Figure 2) (Lidströmer, Aresu, and Ashrafian 2022). ML is different from classical statistical approaches in terms of the question that is being asked in either of the disciplines. While traditional statistical methods aim to find a model that explicitly formalizes the relationship between a set of variables and can be used to test a hypothesis about how the system behaves, machine learning focuses on generating a model that allows for accurate outcome prediction of previously unseen data and does not necessarily provide detailed information on the underlying model architecture itself (Bzdok, Altman, and Krzywinski 2018). Thus, machine learning methods are particularly helpful when dealing with "long data", in which the number of input variables exceeds the number of measured samples, which is an important aspect of any omics type of data (Bzdok, Altman, and Krzywinski 2018).

One important distinction of ML tasks is whether the algorithm is trained with a set of labeled or unlabeled data. The first type of task is called supervised machine learning. Here, labeled data is used to train or "supervise" the algorithm, which can then predict the label of previously unseen data. When the output labels are categorical quantities, the algorithm is called classification and if the output is a quantity, it is called regression. AI-based diagnosis of a disease condition is usually performed as classification tasks and the output variables are either two or more disease categories. The second family of ML algorithms is unsupervised machine learning, which can detect patterns directly from unlabeled data and is used for ubiquitous analysis tasks in medical research. To give an example, the exploratory data analysis of high-dimensional omics-data requires dimensionality reduction (Becht et al. 2018) and clustering for quality control as well as to identify biological heterogeneity in the data (Duò, Robinson, and Soneson 2018). Another type of machine learning is reinforcement learning, which involves training an agent to make decisions in an environment in order to maximize a reward (Esteva et al. 2019).

1.3.2 Classification algorithms for medical diagnostics

Many classification algorithms exist that can perform diagnostic modeling tasks; however, they differ in their specifications and the type of data they can be used for. Linear regression is a

simple algorithm which models a linear relationship of one or more independent variables, or features, to a numeric dependent variable, like for example a drug dosage. Generalized linear models can also take other variables as output, for example logistic regression, which can predict a categorical variable, or multinomial models, if the output variable has more than two categories (McCullagh and Nelder 1989). In such a setting, the association between input and prediction output is transparent and explainable and the algorithm learns how each of the features correlated with various outcomes (Goodfellow, Bengio, and Courville 2016). This makes linear modeling particularly interesting for clinical applications, for example to estimate the risk of breast cancer from a set of clinical parameters (Chhatwal et al. 2009) or to predict long-COVID conditions from electronic health records (Kulenovic and Lagumdzija-Kulenovic 2022). However, these algorithms are not particularly well-suited for high-dimensional data such as genomics data, where the feature space is very large and exceeds the sample number by far. There, models are likely to overfit and generalize very poorly to other data sets. In such cases, algorithms are preferred which can perform feature selection and remove irrelevant predictors from the model (Saeys, Inza, and Larrañaga 2007; Bühlmann and Van De Geer 2011). One example of a feature selection method is the Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani 1996), which can be applied to linear modeling. There, predictors which have no discriminatory power are shrunken to zero, while features with nonzero coefficients represent features that can separate classes (Ghosh and Chinnaiyan 2005). This makes the algorithm well-suited for e.g. genomics and epigenomics data and it has been e.g. applied for biomarker selection and disease prediction in chronic kidney disease (Xiao et al. 2019). Interestingly, it has been shown that Lasso performs equally well as neural networks on prediction tasks such as readmission prediction from administrative clinical data (Allam et al. 2019).

Another algorithm which is widely used in the context of disease prediction is the support vector machine (SVM) (James et al. 2021a). In this method, two classes are divided by finding a hyperplane with a maximal margin between them and in a linear SVM, a hyperplane is a line that separates samples in two dimensions. The data points that lie closest to the separating hyperplane are called “support vectors” since they “support” the hyperplane: If they would be moved, then the hyperplane would move as well (see Figure 3A).

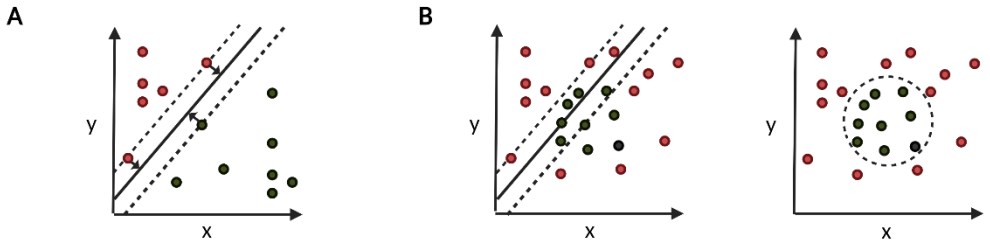


Figure 3*: Concept of a Support Vector Machine. A two-dimensional dataset is shown that consists of two classes (grey and red). A) Separating hyperplanes are depicted as solid lines and margins as dotted lines. Three support vectors are shown and their distance to the hyperplane is indicated with arrows. New observations will be classified according to their position in relation to the separating hyperplane. B) The example shows data points that cannot be separated by a linear hyperplane (left). After transformation using the radial kernel trick, a linear separation becomes possible (right). Figures adapted from* (James et al. 2021b)*.*

This indicates that more distant data points do not influence the hyperplane, which is robust to the behavior of those observations. However, in most cases a perfectly separating hyperplane is either not possible or not desirable, as accounting for outliers would lead to overfitting. To get a greater robustness to individual observations and to achieve a better classification of most of the testing observations, it is reasonable to allow for some training observations to be on the "wrong" side of the hyperplane. The amount of allowed hyperplane violation can be controlled by using a tuning parameter, which can be thought of as a "budget" for margin violations. It controls the bias-variance trade-off (James et al. 2021a): If the margin is rather large, the classifier will have low variance, but high bias, whereas a small margin will highly depend on individual observations, which increases its variance and lowers the according bias. But even when a penalty parameter C is included, it is not possible for a linear function to separate two given sets of data points in many cases. Then, the data can be transformed using the kernel trick: Upon transformation in higher dimensional spaces, the data can become linearly separable (see Figure 3B). Commonly used kernel transformations include polynomial, radial and sigmoid functions. SVMs are widely used for clinical use cases as they account for around 40% of ML papers in healthcare (Richens and Buchard 2022) and have been shown to be applicable to various clinical predictions, such as the prediction of diabetes from clinical parameters (Yu et al. 2010; Barakat, Bradley, and Barakat 2010) or for automated recognition of obstructive sleep apnea from ECG recordings (Almazaydeh, Elleithy, and Faezipour 2012). SVMs have also been used for image classification such as for breast cancer classification from magnetic resonance imaging (Vidić et al. 2018), however, they need a pre-processing step for

feature extraction. Furthermore, SVMs have been influential for the analysis of transcriptomic data, where they are used for a variety of disease prediction studies in early microarray studies (Statnikov et al. 2005) as well as in bulk (S. Huang et al. 2018) and single cell sequencing data (Hu et al. 2016; Alquicira-Hernandez et al. 2019).

Another widely used machine learning algorithm that can be used for medical diagnostics is the random forest classifier (James et al. 2021b; Breiman 2001; Díaz-Uriarte and Alvarez de Andrés 2006). It relies on an ensemble of decision trees, which are rather easy to interpret and can be displayed graphically (see Figure 4A).

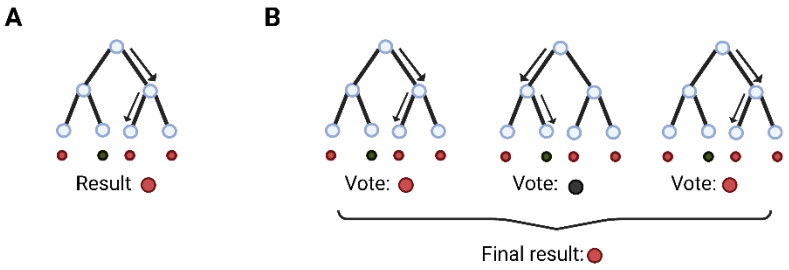


Figure 4*: Concept of classification based on decision trees. A) A decision tree is created based on a top-down binary splitting of the training data. Each terminal node represents a class label and each top-down path represents a classification rule. B) Ensemble of three decision trees that combines the votes of individual models. In this case, the final output is the class selected by most trees.*

Starting at the top of the tree, each branch represents a decision. This brings the predictors in an intuitive order of importance since the influence of a predictor directly corresponds to its position in the tree and there is one prediction result for each tree in the random forest. However, single decision trees usually don't show a very good performance, because they are unstable and small changes in the learning sample impair the prediction accuracy in the test samples. Instead, by aggregating many decision trees as done in random forest classification (see Figure 4B) gives a greater prediction performance. In random forest, each split on each tree is performed using a random subset of the features. In order to increase interpretability of the random forest, additional measures for variable selection have been introduced, such as the Gini importance (Nembrini, König, and Wright 2018). Random forest has been used on transcriptome data, examples include prediction of survival in patients with Ebola virus infection based on blood transcriptome data (Xuan Liu et al. 2017), inferring stages of multiple

sclerosis from blood transcriptomes (Acquaviva et al. 2020) and subclassification of patients with systemic lupus erythematosus (Figgett et al. 2019).

In addition to the algorithms mentioned above, neural networks have been successfully used in medical diagnostics and are applicable to various kinds of medical data (Keane and Topol 2018; Piccialli et al. 2021). Especially such neural networks that contain many layers, also termed deep neural network (DNN) (LeCun, Bengio, and Hinton 2015) (see Figure 2), outperform other methods in machine learning in tasks like visual object (Krizhevsky, Sutskever, and Hinton 2012) or speech recognition (Hinton et al. 2012), but also in predicting the effects of mutations on non-coding DNA on gene expression (Xiong et al. 2015) for example. One prominent example is AlphaFold, a deep learning model which has been shown to predict the three-dimensional structure of a protein solely based on the amino acid sequence with high accuracy, a task that has previously been considered one of biology's grandest unresolved challenges (Jumper et al. 2021; Callaway 2020). One important aspect, particularly for image processing, is that DNNs can learn directly from raw data, meaning that it is not necessary to run feature extractors that transform raw data into suitable representations from which the algorithm can detect patterns (Esteva et al. 2019).

Neural network algorithms are inspired by biological neural networks. In principle, any such network consists of nodes and connections between the nodes. Nodes are considered to perform basic features of a biological neuron. In biological neurons, dendrites receive signals from other neuronal cells. The cell body summarizes this signal, and the axon transmits a signal to other cells, only if a certain threshold is reached. In analogy to that, artificial neurons receive signals from other nodes, process them and can signal to other nodes, only if a certain threshold is reached. They can be mathematically described with an activation function, which "fires", when a linear combination of its inputs exceeds a certain threshold (Rosenblatt 1958). Having only one such artificial neuron, without connections to other nodes, would be equivalent to a logistic regression model (Dybowski 2022). A neural network is a combination of many of such artificial neurons and these connections, also called edges, are inspired by biological synapses. Each connection has a numeric weight associated with it, which determines the strength and the sign of the connection. This link between two nodes serves to propagate the activation from one node to the next. To define the numerical values of the weights and the biases, the network has to be fitted to the data. That is, it starts out with unknown parameter values and those are estimated to fit to the dataset using backpropagation, a process which can become enormously complex and time-consuming depending on the size of the training data

and the network architecture (LeCun, Bengio, and Hinton 2015). Once the parameters are fixed, the trained model can be used to calculate outputs based on a given input. The properties of a neural network are determined by its topology and the characteristics of the included neurons and there are different ways on how such a network can be constructed. When all connections are pointing in the same direction, then the network is called a feed-forward network. Every node receives input from upstream nodes and gives output to downstream nodes and there are no loops. One can think of an easy example of a feed-forward neural network, where there is one input node, which takes as an input a dosage of a medicine, one output node, which would give out effectiveness of the drug and two intermediate nodes (see Figure 5A).

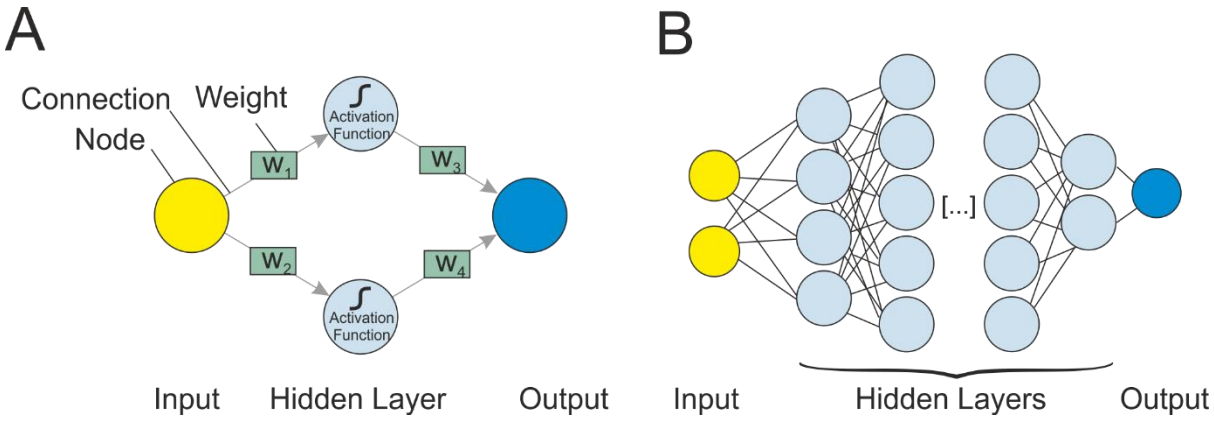


Figure 5: *Schematic diagrams of neural network. A) A simple neural network consists of a single layer of neurons, with input connections that receive information from external sources, such as sensors or other networks. The neurons in the layer process the input signals and generate output signals that are transmitted to other networks or to the outside world. B) The deep neural network consists of multiple layers of neurons, each layer processing the output signals of the previous layer. The input connections receive input signals as in the simple neural network, but the information is processed through multiple layers of nonlinear transformations, enabling the network to learn complex patterns and relationships in the input data.*

Such a network structure can already provide a non-linear decision boundary for classification (Dybowski 2022). However, in practice this can get much more complicated. Usually, neural networks have more than one input node, more than one output node (e.g. for different classes in a multilabel prediction), different layers of nodes between the input and output nodes and a spider web of connections between each layer of nodes (Figure 5B). The layers of nodes between the input and output nodes are called hidden layers. When a network consists of

several hidden layers, it is considered a deep neural network (Dybowski 2022). Other network architectures also exist, for example recurrent networks, where outputs can be fed back to the input of the network (Choi et al. 2017), or convolutional neural networks, which are used in image processing (Yadav and Jadhav 2019) .

Given their advantages compared to classical machine learning algorithms, neural networks have also become increasingly important in biomedical research. Applications of these algorithms are numerous and they have been widely applied to genomics (Kelley et al. 2018), transcriptomics (Eraslan et al. 2019) and epigenomics data (Angermueller et al. 2017), single cell RNA sequencing data analysis (Zheng and Wang 2019) or in multi omics datasets (Leng et al. 2022), where they are used for tasks such as dimensionality reduction, batch removal, classification and clustering tasks.

Convolutional neural networks have been used extensively for computational image analysis for various clinical use cases. Retinal images for example can be used to detect chronic kidney disease (Sabanayagam et al. 2020), to assess risk of cardiovascular diseases (Cheung et al. 2021) and to detect early signs of Alzheimer's (Cheung et al. 2022). Deep neural networks have been developed to detect tuberculosis from chest X-Rays (Lakhani and Sundaram 2017) and outperformed dermatologists in detection of skin cancer (Codella et al. 2017; Haenssle et al. 2018; Esteva et al. 2017). Another field of application is natural language processing, for example recurrent neural networks can be used to process electronic health records (Rajkomar et al. 2018). A disadvantage is however that neural networks are considered "black boxes", and approaches for increasing interpretability of neural networks are being developed (Novakovsky et al. 2023; Lauritsen et al. 2020).

1.4 Considerations for machine learning frameworks in the clinical setting

As outlined above, machine learning applications can successfully be showcased for various diagnostic purposes and have been proposed to support medical practice in many regards. However, in order to establish clinical classifiers that can predict well across real-world settings, it is important to thoroughly consider how predictive algorithms perform in out-of-distribution (OOD) data characterized by e.g. technical imbalances and sample heterogeneity, which are inherent to any clinical setting. A model trained on data from one healthcare system or hospital, for example, may not be generalizable to another site due to differences in local demographics, laboratory equipment and assays, data measurement frequency, and clinical and administrative practices, such as coding and definition of diagnoses. Furthermore, individual

rare diseases may be too infrequent for a per-condition classification in individual training datasets, but can collectively be common and clinically significant when considering a larger population. There is a large body of research on improving OOD robustness of predictive algorithms by e.g. data augmentation methods (Hendrycks et al. 2021; Shorten and Khoshgoftaar 2019), by using synthetic data for training on under-represented labels (Röglin et al. 2022) or by using transfer learning for generalization of algorithms across clinical sites (Wardi et al. 2021). Also, it is important to quantify prediction uncertainty under OOD conditions, so that models can on the one hand accurately predict classes that have been seen during training, but also provide reliable informative estimates about uncertainty and flag such abnormalities for further analyses (Guha Roy et al. 2022; Zimmerer et al. 2022; Linmans et al. 2023). However, despite these important efforts in dealing with distribution shifts in medical data, it is widely acknowledged that model generalizability across OOD conditions can best be achieved when large, heterogeneous, and high-quality clinical data are available for training, which represent clinical distributions as close as possible. (Mårtensson et al. 2020; Zech et al. 2018; Guha Roy et al. 2022).

Such heterogeneous medical data is not readily available for machine learning algorithms to be trained for a variety of reasons. First, medical infrastructure is inherently decentralized. Each patient is an individual data source and healthcare infrastructure is distributed across various providers and systems, which makes it difficult to transfer data between computing centers. Second, the medical sector is highly valuing patient's privacy and confidentiality, which results in a reluctance to share data beyond what is necessary for patient care. Third, data privacy is strictly protected (Price and Cohen 2019) and regulated by legal frameworks like the General Data Protection Regulation (GDPR) ("General Data Protection Regulation" 2023) in the EU and the Health Insurance Portability and Accountability Act (HIPAA) in the USA ("HIPAA and Administrative Simplification" 2022). In case of HIPAA, one of the main strategies to protect medical patient data is to de-identify the data by removing identifying elements such as email addresses or social security numbers (El Emam 2011). The problem with this, however, is that it often results in the removal of useful information and that it does not ensure adequate privacy protection of the individual, since de-identified data is not necessarily anonymous. This concerns all medical data types that are commonly used in machine learning. Not only do genomics data have the potential to uniquely identify individuals (Oestreich et al. 2021), but the same is true for electronic health records for example, where only a few annotations are enough for patient re-identification (L Sweeney 2000). Furthermore, patient

identification from medical imaging data is possible (Esmeral and Uhl 2022; Rieke et al. 2020). On the other hand, the GDPR is an attempt to regulate this by putting strict regulations on sharing of *any* personal medical data, which is relevant to any international research collaboration as well as to nearly all large global research companies which build machine learning models using personal data to train them. As a result, concepts based on data sharing are difficult to realize in the medical sector overall. (McCall 2018; Oestreich et al. 2021)

Nevertheless, huge endeavors have been undertaken to collect curated high-dimensional atlas-like datasets that can serve as reference datasets for the scientific community, and which also have been used to exemplify how disease classifiers could be trained on such data. For example The Cancer Genome Atlas Network (Cancer Genome Atlas Network 2015; Cancer Genome Atlas Research Network et al. 2013) alone has generated 2.5 petabytes of genomic, epigenomic, transcriptomic and proteomic data on 33 cancer types (Hoadley et al. 2018). Other datasets cover population-based data such as the UK biobank, a population-based prospective cohort of 500,000 men and women in the UK (Allen et al. 2014) and e.g. Scotland' Safe Havens, where EHR data are stored centrally on the NHS network and de-identified data is being made available for research purposes. Further large datasets are the NIH initiative "all of us" (All of Us Research Program Investigators et al. 2019) collected genomic information together with electronic health records. Prospective population-based cohorts such as the Dutch LifeLines DEEP aim to collect multi-omics data to understand environmental influences e.g. on the development of chronic diseases (Tigchelaar et al. 2015; Zijlema et al. 2016). Furthermore, large databases exist e.g. on Alzheimer-related data (Weiner et al. 2013) as well as clinical radiology image data, which has been used for training of predictive classifiers (Prior et al. 2017; X. Wang et al. 2017). Additional, huge datasets e.g. on single cell transcriptomics have been published, which are being used to generate transcriptomic signatures for diseases in specific cell-types (Perez et al. 2022; Nichita et al. 2014; Schulte-Schrepping et al. 2020a). However, when compared to the amount of data which was used e.g. to train the algorithmic agents in the game of Go (Silver et al. 2016) or in the field of research on autonomous vehicles (Fridman et al. 2019; G. A. Kaissis et al. 2020), these carefully collected datasets still come from a relatively small number of sources or from single technological platforms with strict protocol harmonization. Therefore, to establish machine learning applications which can learn from data on real-world scale datasets, other solutions have to be put forward that go beyond data that can be collected in an academic context, even within large consortia.

Cloud computing infrastructures have been proposed as one major framework to overcome limitations of local model training and have been proposed as a solution to integrate data from genomics, systems biology and biomedical data mining without having to have the full computing infrastructure as well as the data locally. However, even though this comes with several advantages compared to local computing, having data and models stored at a central instance comes with the disadvantages such as data duplication and increased data traffic as well as challenges for data privacy and security (Dove et al. 2015). Furthermore, such an architecture favors data monopoles and among others large corporations such as Meta (formerly known as Facebook), Amazon, and Alphabet (formerly known as Google) have started massive AI initiatives within health, and providing personalized health data to such large companies might be critical in terms of data protection (Lidströmer, Aresu, and Ashrafian 2022).

Federated learning (FL) has been developed as an alternative approach for confidential machine learning without data sharing (McMahan et al. 2016b). Here, a federation of participating sites (or nodes) each train a local model. When each different site with local data has finished calculating, the model weights are sent to a central server, which is merging them to create a global model and those updated parameters are sent back to the participating sites. With this, model training is decoupled from the need for direct access to the raw training data, which makes FL an attractive approach for distributed machine learning and various applications for healthcare data have been published (Xu et al. 2021; Rieke et al. 2020; X. Li et al. 2020; Pati et al. 2022). However, the star-shaped structure is kept with one central server being responsible for aggregating the parameters, which decreases fault tolerance and requires trust in the central instance. Therefore, a computational infrastructure is needed that fully matches the needs of the medical domain for collaborative, yet confidential machine learning.

2 Aim of the thesis

Advances in AI for medical diagnostics hold great promise for assisting and improving clinical decision making by translating high-dimensional biomedical data into scores and recommendations on disease diagnostics, disease severity, and treatment options, which are clinically usable and demonstrate added value in comparison to current standard diagnostic workflows. However, AI models presented in literature oftentimes fail to materialize such demonstrable value and may even be detrimental when being deployed into clinical applications (S.-C. Huang et al. 2022). Therefore, new computational frameworks are needed to finally translate innovations in technological and algorithmic research into improvements in individual medical care. It is the aim of this thesis to contribute to this translation by evaluating the use of high-dimensional transcriptomic datasets for privacy-preserving disease classification and to put this into the broader context of using omics technologies for AI-driven, systems-level medical diagnostics.

My work is presented as a cumulative thesis consisting of two publications:

In the first publication, I present my work on disease classification based on high-dimensional transcriptomics data, exemplified for the prediction of Acute Myeloid Leukemia (AML). I provide evidence that transcriptome data can be utilized for highly accurate and scalable ML-based primary disease diagnosis (Warnat-Herresthal et al. 2020). In the second publication, I introduce Swarm Learning as a new concept of privacy-preserving, collaborative machine learning, which can generate shared classification models on high-dimensional data, without sharing any personal data. This is exemplified for transcriptomic data on AML, COVID-19 and tuberculosis, as well as for X-Ray data on lung diseases and covering a variety of clinically relevant prediction scenarios (Warnat-Herresthal et al. 2021).

In summary, the research presented here provides evidence that high-dimensional medical data, in combination with a decentral machine learning framework such as Swarm Learning, could make it possible to utilize the power of AI for medicine in an effective and clinically feasible way.

3 Developing blood-based disease classifiers for prediction of AML

Accompanying text for the following publication:

Stefanie Warnat-Herresthal, Konstantinos Perrakis, Bernd Taschler, Matthias Becker, Kevin Baßler, Marc Beyer, Patrick Günther, Jonas Schulte-Schrepping, Lea Seep, Kathrin Klee, Thomas Ulas, Torsten Haferlach, Sach Mukherjee, Joachim L. Schultze. Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. iScience. 2020 Jan 24;23(1):100780. doi: 10.1016/j.isci.2019.100780.

Acute myeloid leukemia (AML) is a severe, often fatal cancer of the hematopoietic system, which is characterized by uncontrolled proliferation of malignant bone marrow stem cells. It is the most common form of acute leukemia in adults, with a median age at diagnosis of 68 years and an estimated 5-year overall survival rate of 30% (Shimony, Stahl, and Stone 2023). AML presents with very unspecific symptoms such as infection, anemia and bleeding and despite enormous improvements in diagnosis and therapy during the last years, primary diagnosis is often delayed. While subclassification of AML is already mainly based on molecular features, particularly taking into account the mutational status of the leukemic cells (Shimony, Stahl, and Stone 2023), guidelines for primary diagnosis and management of the disease recommend diagnosis by a rather conservative diagnostic pipeline consisting of a combination of cytomorphology, cytogenetics and immunophenotyping (Eckardt et al. 2020; Döhner et al. 2017). At the same time, AML has been used as a prime showcase for data-driven disease prediction and subtype discovery since the very first landmark studies that demonstrated the usability of machine learning for transcriptome data already in 1999 (Golub et al. 1999). Many research articles followed that analyzed blood transcriptome data on AML and related diseases (Goswami et al. 2009; Zhang et al. 2018), as well as computational models that were presented to detect prognostic gene-expression profiles (Valk et al. 2004; Z. Li et al. 2013; T Haferlach et al. 2007) or predict the disease and its subtypes (Bullinger et al. 2004; Wouters et al. 2009; Torsten Haferlach et al. 2010, 2005). Furthermore mutational status in 25 genes have been proposed for AML stratification and 14 separate disease subtypes with distinct diagnostic features and clinical outcomes have been proposed (Papaemmanuil et al. 2016), underlining the molecular heterogeneity of the disease. Furthermore, prediction of AML has been showcased to be suitable for other ML-based automated diagnostic frameworks, such as automated detection and classification of leukemia based on microscopic images of blood cells (Bibi et al. 2020), underlining the various possibilities in which AML diagnostics could potentially be enhanced. However, despite this rich body of data and research that has been

gathered, there has been little translation of such approaches into clinical practice, which prompted the starting point for the present study.

As the data sciences have developed and transcriptome data from peripheral blood, including samples from AML patients, have become more available, we sought to test whether primary and differential diagnosis would be possible by using transcriptome-based machine learning approaches. For that aim, we first collected all data on human blood transcriptomes of AML patients that was available for two microarray platforms as well as for Illumina RNA sequencing. This resulted in a unique transcriptomics dataset for classifier development consisting of three independent datasets with a total of 12,029 samples from 105 studies, containing 4,145 AML samples of diverse subtypes, 7,884 other samples derived from healthy controls (n = 904), patients with acute lymphocytic leukemia (ALL, n = 3,466), chronic myeloid leukemia (CML, n = 162), chronic lymphocytic leukemia (CLL, n = 770), myelodysplastic syndrome (MDS, n = 267), and other non-leukemic diseases (n = 2,312). To the best of our knowledge, the dataset as presented in this work is still the most comprehensive transcriptome dataset on AML prediction to date.

After quality control and preprocessing, the data was used for the prediction of AML in several clinically relevant scenarios, such as the prediction of AML vs. all other samples, including healthy controls and the prediction of AML in a differential diagnosis setting, where AML is tested against other leukemias. We found that classification accuracy, specificity, and sensitivity are already very good when sampling sizes are small, but the performance of prediction still increased when the training dataset size was larger. This is particularly relevant when considering low prevalence scenarios and we demonstrate that even small gains in overall accuracies can lead to massive improvements in positive predictive values (PPVs) in low-prevalence settings. Nine different prediction algorithms were applied to all scenarios, and overall, the prediction task was very robust regardless of the algorithm used. Also, prediction accuracy was not dependent on specific AML subtypes.

However, it is known that effects of site-specific factors, such as technical batch effects or differences due to sample distribution in different studies can pose problems for generalizing predictive models. Following that, we analyzed the effect of cross-study variation on predictive performance on our dataset. Due to the large number of studies included, it was possible to perform an entirely disjoint cross-study analysis, meaning that we were able to separate samples of complete studies in training and test sets and permuted this setting 100 times. As

expected, performance measures were worse than in the random sampling approach, where training and test samples came from a pool of all studies. However, performance measures got better with increasing training set size, meaning that large datasets can help to overcome such study-specific batch effects, which demonstrated the power of current machine learning algorithms even when using such a heterogeneous sample sets which were collected from many individual sites.

Another major part of the work concentrated on the question whether classifiers could be translated across technical platforms. Technologies evolve and it is important that classifiers can cope with changes of the technical platforms. To this end, we evaluated whether an algorithm which was trained on one microarray platform could be used to predict data which was generated by another microarray platform or RNAseq data and saw huge differences in prediction accuracies depending on the technological platform used for training. However it was possible to rescue classifier performance by using rank transformation (Zwiener, Frisch, and Binder 2014) applied on each dataset independently, meaning that transcriptomic signatures do translate robustly across technologies.

Lastly, we analyzed the relationship of features establishing a predictive signature compared to genes that qualify as differentially expressed (DE) and genes that are known to be important for AML biology. Genes which were selected by the lasso classifier were often not differentially expressed and the classification algorithm was shown to be robust to removal of known AML-related genes, pointing out that it can be beneficial for prediction to consider data-driven, genome-wide signatures rather than taking into account sets of single genes.

Taken together, we demonstrated that the combination of machine learning and transcriptomics can yield highly accurate and robust classifiers, which can be translated across clinical batches as well as technological platforms. This supports the vision that transcriptomic-based ML could be introduced to clinical practice to support AML primary diagnosis, particularly in settings where hematological expertise is not sufficiently available and lays the ground for the design of future prospective studies to assess diagnostic utility.

For this publication I was responsible for major areas of the work including the design of the study search strategy, the collection and preprocessing of datasets 1, 2 and 3, writing code and performing calculations, evaluation of results, preparation of figures and I was the major contributor to paper writing.

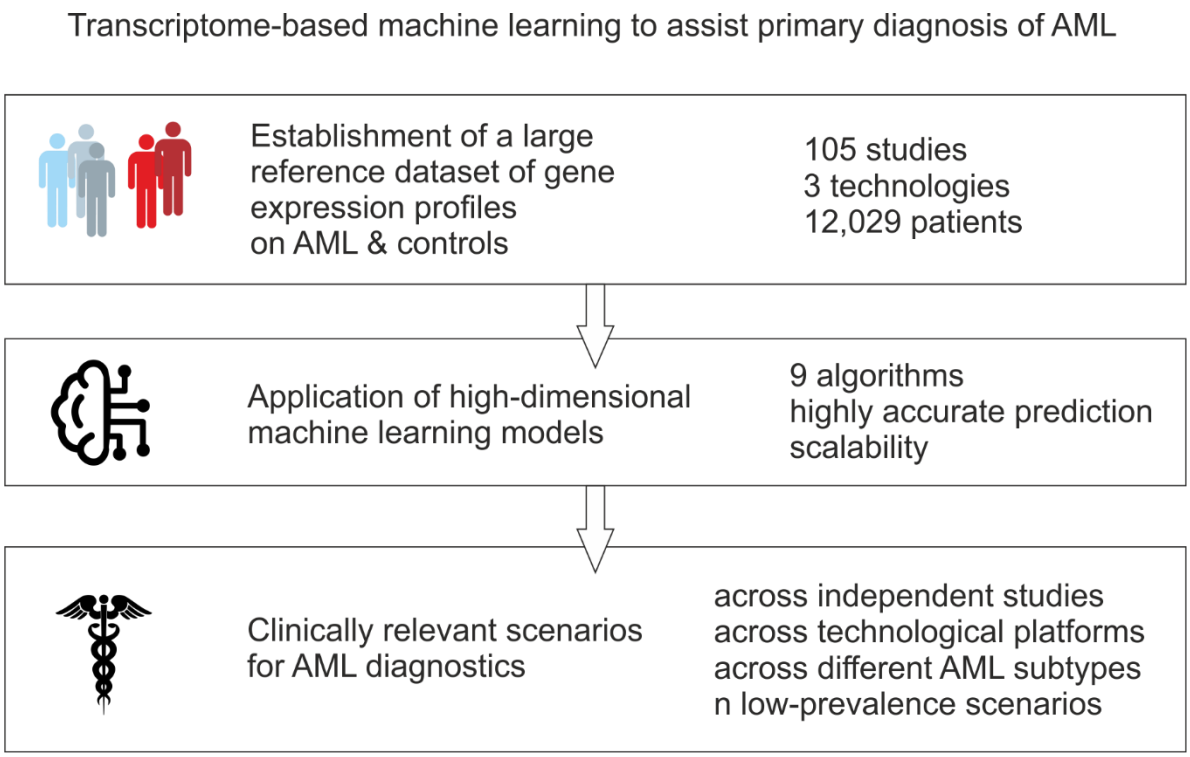


Figure 6: *Graphical abstract of Warnat-Herresthal et al. 2020. A comprehensive meta-analysis on the prediction of AML based on published gene expression profiles was performed. Nine high-dimensional machine-learning algorithms were trained in different clinically relevant scenarios. We report highly accurate disease classifiers, which are scalable and can be translated across site-specific clinical study batches as well as technical platforms.*

4 Swarm Learning as a decentral and privacy-preserving machine learning approach for disease classification

Accompanying text for the following publication:

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, Ahmad Aziz, Sofia Ktena, Florian Tran, Michael Bitzer, Stephan Ossowski, Nicolas Casadei, Christian Herr, Daniel Petersheim, Uta Behrends, Fabian Kern, Tobias Fehlmann, Philipp Schommers, Clara Lehmann, Max Augustin, Jan Rybniker, Janine Altmüller, Neha Mishra, Joana P. Bernardes, Benjamin Krämer, Lorenzo Bonaguro, Jonas Schulte-Schrepping, Elena De Domenico, Christian Siever, Michael Kraut, Milind Desai, Bruno Monnet, Maria Saridaki, Charles Martin Siegel, Anna Drews, Melanie Nuesch-Germano, Heidi Theis, Jan Heyckendorf, Stefan Schreiber, Sarah Kim-Hellmuth, COVID-19 Aachen Study (COVAS), Jacob Nattermann, Dirk Skowasch, Ingo Kurth, Andreas Keller, Robert Bals, Peter Nürnberg, Olaf Rieß, Philip Rosenstiel, Mihai G. Netea, Fabian Theis, Sach Mukherjee, Michael Backes, Anna C. Aschenbrenner, Thomas Ulas, Deutsche COVID-19 Omics Initiative DeCOI), Monique M. B. Breteler, Evangelos J. Giamarellos-Bourboulis, Matthijs Kox, Matthias Becker, Sorin Cheran, Michael S. Woodacre, Eng Lim Goh, Joachim L. Schultze (2021). Swarm Learning for decentralized and confidential clinical machine learning, Nature 594, 265–270.

Applications of machine learning, including those that are building on genomic and transcriptomic data, have the potential to make medical diagnostics more accurate, efficient, and accessible for patients worldwide, and to radically transform several aspects of medical practice. However, as the possibilities of ML/AI technologies for medicine are appreciated scientifically, major challenges remain in the deployment of such technologies in the clinical setting. Those include concerns about implementation, accountability and fairness of AI algorithms (Rajpurkar et al. 2022). One central aspect underlying these concerns is the generalizability of models, meaning the performance of disease prediction on data that has not been included in training (Mårtensson et al. 2020; Zech et al. 2018). To overcome these limitations, classifiers need to be trained on large datasets from various clinical sites, as it has been demonstrated in the previous chapter of this dissertation. Technically, if sufficient data is available at any clinical site, model training for disease prediction can be performed locally (see Figure 7A). However, such a model will likely not generalize well across other clinical instances due to sources of variation, such as different technological platforms or different data handling and processing protocols, different distribution of patient characteristics such as age, ethnicity and disease subtype (Obermeyer and Emanuel 2016). Furthermore, rare disease conditions are likely to be missed from individual local training data.

Data centralization, for example by cloud solutions, has become one alternative to overcome the limitations of local computing (Figure 5B), including access to sufficiently large datasets and adequate infrastructure for data storage, processing, and analysis. (Dove et al. 2015; Langmead and Nellore 2018; Ping et al. 2018) Compared to local approaches, cloud computing can significantly increase the amount of data for training of ML algorithms, therefore reducing prediction biases and improving testing performance. However, cloud computing has disadvantages such as data duplication from local to central data storage, increased data traffic and issues with locally differing data privacy and security regulations. Furthermore, such an architecture favors data monopolies.

As an alternative, federated learning (FL) has been developed (Konečný et al. 2016; McMahan et al. 2016a) (Figure 5C). Unlike in a conventional centralized machine learning framework, participants do not need to share their data to a central custodian. Instead, only the parameters of the model are shared, which is trained on local data (McMahan et al. 2016b). The parameters are then merged by a central server and communicated back to the participating sites (G. A. Kaissis et al. 2020; Konečný et al. 2016; Zhilin Wang and Hu 2021; Pati et al. 2022). FL is a pioneering ML architecture in the sense that it enables the establishment of a shared ML model without sharing the actual data. However, a remainder of the central architecture is kept, which requires trust, hampers implementation across different jurisdictions and therefore still requires the respective legal negotiations. Furthermore, the risk for a single point of failure at the central structure reduces fault-tolerance.

To overcome these limitations, we introduced Swarm Learning (SL) as a completely decentralized machine learning architecture, which facilitates integration of any omics data from any data owner world-wide without violating privacy laws (Figure 5D). In SL, the dedicated central parameter server is dismissed, which is facilitated by blockchain-based peer-to-peer networking and coordination in a completely decentralized network with immutable parameter sharing. This enables different organizations or consortia to efficiently collaborate in a completely decentralized and privacy-preserving manner and thereby goes beyond federated learning.

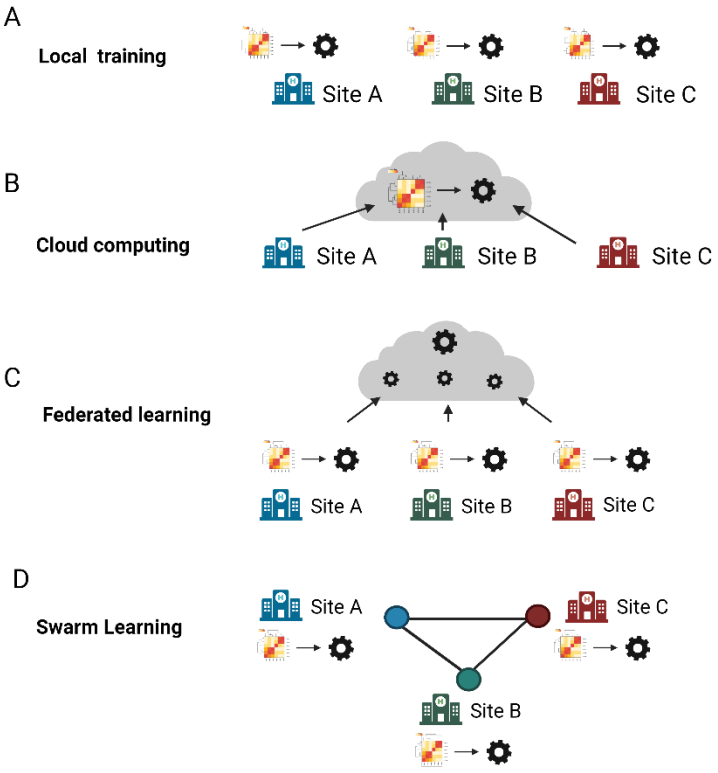


Figure 7*. Schematic overview of Machine Learning architectures with three independent clinical sites. A) Local Training. Each hospital trains its own model B) Cloud computing: Data is sent to a central instance and a shared model is trained C) Federated Learning: Data stays local, model is trained locally, and model parameters are shared to a central parameter server C) Swarm Learning: Model and parameters are both local. Model parameters are shared within the decentral Swarm Network.*

In the paper presented here, we used more than 16,400 blood transcriptomes derived from 127 individual studies with non-uniform distribution of cases and controls and significant study biases. First, we used the transcriptome dataset consisting of peripheral blood mononuclear cell (PBMC) transcriptomes established in Warnat-Herresthal et al. 2020. With this, we illustrated the feasibility of SL to develop disease classifiers based on distributed data for Acute Myeloid Leukemia by siloing various distributions of cases and controls per node and comparing model performance of individual nodes against performance of the model achieved within the Swarm Network. SL outperformed prediction results of single nodes in most settings and most importantly, we could demonstrate that Swarm Learning can overcome single-source biases of nodes as well as prediction biases based on technological differences between nodes. We extended this to the prediction of samples from patients with acute lymphoblastic leukemia (ALL) as cases, ran a multi-class prediction across four major types of leukemia, extended the number of nodes to 32, tested onboarding of nodes at later time points and replaced the deep

neural network with LASSO, and the results echoed the initial findings, illustrating that SL outperforms local learning.

In a second use case, we sought to investigate whether SL would also be feasible in conditions which are expected to be more heterogeneous, such as infectious diseases like tuberculosis (TB). Previous work in smaller studies had already suggested that active TB as well as outcome of TB treatment can be revealed by blood transcriptomics (Zak et al. 2016; Leong et al. 2018; de Araujo et al. 2016; Verma et al. 2018; Thompson et al. 2017). In the present study, we demonstrated that this prediction task is also feasible in a Swarm Learning setting and that patients with active and latent tuberculosis can indeed be identified based on their blood transcriptomes. For this, we tested several distributions of cases and controls, sizes of nodes and numbers of nodes. In each of the tested settings, models trained by the Swarm Network outperformed models trained at single nodes.

Next, we extended our analysis to X-Ray images to also include a second medical data space, which is already extensively used for developing AI-guided diagnostics. For this, we used a publicly available dataset of more than 95,000 chest X-rays which has been used before to benchmark machine learning approaches. Also here, we showed that SL outperforms the contributing nodes in a multi-class and multi-label classification problem.

Finally, we evaluated how clinical prediction would work in an outbreak scenario of a newly identified disease, as it has been encountered in the world-wide pandemic of COVID-19. Usually, COVID-19 patients are identified by PCR-based assays to detect viral RNA (Corman et al. 2020). However, we used this case as a proof-of-principle study to illustrate how SL could be used even very early on during an outbreak based on the patients' immune response captured by analysis of the circulating immune cells in the blood. Here, blood transcriptomes only present a potential feature space to illustrate the performance of SL and it has been shown that COVID-19 can be predicted based on several data modalities (Zoabi, Deri-Rozov, and Shomron 2021; Laguarta, Hueto, and Subirana 2020). However, using blood transcriptomes for assessing the specific host response, in addition to disease prediction, might be beneficial in situations for which the pathogen is unknown or specific pathogen tests are not yet possible. Furthermore, blood transcriptomics can contribute to the understanding of the host's immune response (Schulte-Schrepping et al. 2020b; Bernardes et al. 2020; Krämer et al. 2021), and also includes information about COVID-19 severity, which cannot be assessed by viral testing alone (Ulas et al. 2020).

To extend this to a real-life scenario, we evaluated how data from different individual centers with very different patient characteristics in terms of their local controls, age-, sex- and disease severity distributions would perform in a SL setting to collectively predict COVID-19. This Swarm network of six participating centers was additionally tested on two external datasets, one with only convalescent COVID-19 cases and one dataset of only granulocyte COVID-19 samples instead of whole blood samples. SL outperformed all nodes in AUC for the prediction of the global test dataset and could successfully predict the external datasets. In addition, we tested further unbalanced scenarios with training e.g., on male samples and testing on females or using test nodes with very different disease severity distributions. In all cases, SL outperformed models trained on single nodes.

Collectively, the SL approach combines blockchain technology with decentralized machine learning to create an entirely democratized approach that eliminates the need for a centralized custodian. It therefore represents a uniquely suitable strategy to utilize the possibilities of machine learning in the medical domain, which is inherently decentralized and fundamentally based on trust. Blood transcriptomes were used since they combine blood as the most widely used surrogate tissue for diagnostic purposes with an omics technology that is providing high-dimensional data. However, the here presented approach can be translated to a multitude of clinical use-cases, algorithms and data spaces and this way enable global cooperation on a variety of use-cases, ultimately leading to a more data-driven systems-level approach to medicine.

The results of this publication were obtained in close collaboration with different research groups, clinical centers and in cooperation with Hewlett Packard Enterprise (HPE), who developed the Swarm Learning Library**.** My specific contributions include the design of all predictive scenarios presented within this publication as well as the collection, preprocessing and quality control of datasets A1, A2, A3, B, D, and E. Furthermore, I was responsible for the calculation of predictive performance measures from raw prediction output files and I was the main contributor to the interpretation of the results, the preparation of figures and to paper writing.

5 Summary & Outlook

The first part of the thesis illustrates that accurate detection of AML is possible solely by machine learning algorithms based on blood transcriptome data without requiring additional expert input. In Warnat-Herresthal et al. 2020 a unique reference dataset of annotated public transcriptome data has been compiled for this purpose, combining 105 independently performed studies across three technological platforms with more than 12,000 samples in total including AML as well as other leukemic diseases, healthy and other non-leukemic control samples. This made it possible to thoroughly evaluate the performance of classifiers across a wide range of clinically relevant scenarios, including prediction across independent studies and translation of trained models between technological platforms. ML-based detection of AML was highly accurate across all tested scenarios and robust in terms of the used prediction algorithm. In this way, this publication provides evidence that combination of ML approaches with existing technologies for high-dimensional blood transcriptomics can yield highly effective, robust, and purely data-driven classifiers for near-automated primary diagnosis of AML, opening the possibility for prospective clinical trials based on these results.

The second part of this thesis covers Swarm Learning as a decentralized machine learning framework which enables collaborative machine learning across independent institutions, as introduced in Warnat-Herresthal et al. 2021. In SL, medical data from any data owner world-wide can be used to establish shared, data-driven models for disease prediction at the same time supporting data privacy by design. This approach was evaluated for the prediction of AML, active and latent tuberculosis as well as COVID-19 and other lung pathologies based on five transcriptome datasets consisting of more than 16,400 blood transcriptomes derived from 127 independent clinical studies with non-uniform distributions of cases and controls and substantial study biases, as well as more than 95,000 chest X-ray images. High performance of SL was demonstrated across a multitude of real-world prediction scenarios, such as primary diagnosis of AML and tuberculosis, as well as simulation of an COVID-19 outbreak scenario across independent clinical sites and SL outperformed single-site predictions across all evaluated scenarios. In this publication we proposed that SL represents a uniquely suitable strategy to utilize the possibilities of machine learning in the medical domain, and to ultimately enable systems-level diagnostics based on high-dimensional molecular profiles, while at the same time ensuring data confidentiality and privacy.

With these two elements, first a proof-of-concept on the robustness of ML models to predict AML based on high-dimensional transcriptomic profiles and second, the introduction of a machine learning framework for confidential and decentralized, collaborative machine learning, I have laid the foundation to formulate a concrete hypothesis on how future AI-assisted medical diagnosis can be envisioned.

AML has served as a showcase within this thesis to elaborate the power of high-dimensional omics profiles, with blood transcriptomics providing a particularly informative snapshot of the immunological state across the body as a whole. Importantly, AML is a disease with enormous diagnostic and therapeutic complexity, and any future ML-based diagnostic workflow needs to account for this. The disease heterogeneity is becoming even more pronounced with a continuously improved understanding of the molecular pathophysiology, illustrated by the fact that two competing classification systems on AML were recently published by independent expert groups (Khoury et al. 2022; Arber et al. 2022), which introduce non-overlapping subtype terminology and contain different additions or updates of disease entities based on clinical, immunophenotypic, and molecular data (Falini and Martelli 2023).

The obvious question is, however, how this phenotypic complexity of the disease can be tackled best by any diagnostic workflow. As introduced, technological tools for accessing high-dimensional molecular profiles are in place, have been tested to be robust for clinical use and could, together with state-of-the-art machine learning algorithms, be utilized for an accurate, data-driven, and diagnostic pipeline that analyzes a multitude of parameters in parallel. In fact, the notion that gene expression profiles particularly of leukemic patients contain highly informative patterns which can be detected by classification algorithms is not new (Golub et al. 1999) and have proposed early on as an unbiased and robust readout that could complement current diagnostic pipelines of AML (Torsten Haferlach et al. 2010). If and where to position transcriptome technologies within the current standard diagnostic pipeline of leukemia is however still subject to debate and the translation of the large body of academic findings on the topic into clinical practice is not happening at a large scale.

Interestingly, recent classification proposals and guidelines clearly *emphasize* the integration of molecular characterization into clinical practice and, in addition to the genetic characterization of leukemic subtypes, results of transcriptomic profiling are starting to be described (Khoury et al. 2022; Arber et al. 2022; Shimony, Stahl, and Stone 2023). For example, the international consensus classification (ICC) of myeloid Neoplasms and Acute

Leukemias lists new subcategories of ALL with driver structural lesions, which are also recognizable by their distinct gene expression signatures (Arber et al. 2022). Additionally, the revised 2022 ELN risk classification includes new response criteria and treatment recommendations based on increased understanding of the molecular pathogenesis of AML (Döhner et al. 2022).

Meanwhile, research on usage of ML for diagnostics of leukemia is accelerating and could be integrated at different levels of the diagnostic pipeline. Genome-based molecular signature profiles have been shown to define previously unrecognized subgroups of AML which reflect disease prognosis, potentially harmonizing the disease classification (Awada et al. 2021). Bulk transcriptomic profiles can be linked to drug sensitivity in AML treatment (S.-I. Lee et al. 2018) and single-cell RNA sequencing profiles reveal distinct disease hierarchies that are relevant to disease progression (van Galen et al. 2019). Beyond omics data, it has been shown that AML diagnostics based on blood smear image data can be automated (Fatma and Sharma 2014; Das and Dutta 2019; Shafique and Tehsin 2018; Jha, Das, and Dutta 2020), that cell morphology can serve as a predictor or therapy response (Duchmann et al. 2022) and that clinical data includes valuable information for algorithmic classifiers, e.g. to predict relapse in childhood ALL (Pan et al. 2017), underlining the strength and flexibility of ML-based approaches overall.

This raises the general question of why such innovative AI concepts have not been translated to clinical practice much more quickly. Clearly, this question goes beyond the presented showcase of AML diagnostics, but touches upon the relationship of AI and medicine in general (Rajpurkar et al. 2022). The recent COVID-19 pandemic has been an extraordinary example, where the interaction of biomedical science, AI developers, medical stakeholders, media, politics and society could be monitored “fast forward” and made it possible to identify structural deficits in the process of developing and deploying medical AI models (S.-C. Huang et al. 2022). While large global efforts have been undertaken to leverage digital health data and machine learning techniques to tackle the challenges posed by the pandemic, and to provide e.g. models for early detection and prognostication (Mei et al. 2020), severity scoring (Frid-Adar et al. 2021), long-term outcome and mortality predictions (Ramtohul et al. 2020), most of these models failed to provide demonstrable practical value, and some may even have been detrimental if rolled out to the clinics (Roberts et al. 2021), partly generating disillusionment and distrust in the potential of AI to impact medicine (Wilkinson et al. 2020).

The answers on which lessons can be learned from the 'pandemic experience' are manifold and concern the whole process of building models for clinical applications, from having a more clear definition of actual clinical needs that need to be tackled upfront, to better curation of data for model training and more thorough model testing as well as careful evaluations of models also post-deployment (S.-C. Huang et al. 2022). Importantly, there is broad consensus that AI models, to finally deliver improvements in individual care, need to be trained on data which is resembling the clinical situation as close as possible, and any retrospective collection of public datasets is necessarily lacking this. While retrospective data collection often is the natural starting point for model development, prospective trials must follow that touch the clinical "ground truth", which is mirroring actual deployment scenarios. Such data is characterized not only by a more specific set of samples than covered in most retrospectively collected datasets, e.g. particularly including early disease samples for screening applications, but also factors of technical and biological variation will need to be considered that by definition cannot be ruled out *a priori*, especially for high-dimensional omics data, as described in this thesis. Depending on the concrete diagnostic setting in each and every medical site, there might be different sample handling and data processing protocols in place, different technological platforms being used for data generation, and different institutions, countries and continents will include different patient populations in terms of ethnicity, age, sex and social background. Sites may even apply varying disease classification scores and terminology to describe clinical phenotypes. Ultimately, estimating how well any machine learning model is generalizing across this range of unknown variation is only possible *a posteriori*, meaning after thoroughly testing them within these described settings. In addition, randomized clinical trials would be necessary to proof clinical benefit, which is to show that clinical decisions made on the bases of new AI-procedures have a beneficial influence on the disease course in relation to well-defined outcome measures, such as e.g. increase in the five-year survival rate (Vogeser and Bendt 2023). Furthermore, possibilities for continuous evaluation and model improvement need to be considered also after deployment, to detect distribution shifts that were not anticipated beforehand. Clearly, collection of such data goes beyond what can be performed in an academic setting and it requires medical institutions to allow data management to become an integral part of any clinical daily routine, so that continuous integration of new knowledge and true data distributions are included. While further digitalization of the medical sector will likely be pushed by political authorities, the adoption of machine-learning tools as part of clinical routines is, to date, rather alien for most clinicians and requires individual engagement, motivation, and trust. To account for this and to enable medicine and the health care sector to

profit from knowledge contained in big data, which is not accessible via current laboratory diagnostics, IT frameworks must account for the characteristics of the medical and health care sector setting, not *vice versa*.

Swarm Learning accounts for this by translating the core values of the medical traditions such as confidentiality and knowledge sharing to a state-of-the-art machine learning framework and could therefore become the basis for large prospective trials on ML-based applications as well as integral part of future diagnostic pipelines. Interestingly, the concept of Swarm Learning has been taken up by the scientific community in biomedicine as well as computer science and is currently being explored further. An independent research group has shown that SL can be used to predict the mutational status of colorectal cancer based on histopathology images (Saldanha et al. 2022) and for establishing molecular biomarkers in gastric cancer (Saldanha et al. 2023), and the European Union recently funded a large consortium to establish a Swarm Learning Network in the context of breast cancer screening ("ODELIA " 2023). Technical extensions of the SL concept have been proposed in the field of computer science, e.g. adding a human-in-the loop for integrating user feedback in the loop of learning (Dong, Sarker, and Qian 2022) and a swarm deep reinforcement learning framework in the field of robotics (Zhu, Zhang, and Li 2022).

In essence, SL is a computational framework which is agnostic to the concrete machine learning algorithm that is used for training, and the concept is in principle applicable to any model architecture which is defined by sharable parameters. Therefore, when arguing for SL in diagnostic routines, a multitude of additional perspectives relate to this. One central aspect is model interpretability, which is not only necessary to generate trust among clinicians and patients, but it has also recently been formulated as a requirement for GDPR-compliance (Sovrano, Vitali, and Palmirani 2020). The lack of explainable models is considered as one of the key reasons for the restricted uptake of AI algorithms in healthcare and when models or systems cannot be well interpreted, it can be difficult to accept and communicate their conclusions (Reddy 2022). This is especially important when high-performing models are found to use wrong or confounding variables, as it has been described for the image-based prediction of hip fractures (Badgeley et al. 2019) or the detection of COVID-19 (DeGrave, Janizek, and Lee 2021). A whole field of research has emerged to address these issues and to illuminate and communicate decision processes of deep-learning architectures (Mukhtorov et al. 2023; Javed et al. 2023; Kandul et al. 2023). Of note, many of these improvements that are being made on explainable AI can also be integrated in a clinical pipeline based on the SL

framework. The same is true for additional measures of data safety and security such as homomorphic encryption (Froelicher et al. 2021) and differential privacy (Abadi et al. 2016), which are in principle orthogonal to the SL framework and can be added to it.

Another important aspect for a potential deployment of a decentralized machine learning system in the healthcare sector is the option to deploy a model sequentially at different sites since data might not be available at all contributing centers simultaneously. For this, continuous learning algorithms are needed, where a single model continues to learn from new cohorts of patients and maintains generalizability. Clinical examples of this have been reported (Amrollahi et al. 2022; Kiyasseh, Zhu, and Clifton 2021) and implementing such an approach in a Swarm Learning-based application would be of high importance.

To exemplify how SL can be utilized at different stages of a future diagnostic workflow, consider a patient with undiagnosed AML that presents at the GP with unspecific symptoms such as weakness and fever (Figure 8). Currently, the most likely diagnostic path is that this patient faces several referrals before the correct primary diagnosis is being made, especially in early stages of the disease. Also, standard blood tests that are being performed at the GP are providing only single parameters, which need to be evaluated manually. When being referred to a hematologist, AML subtyping is performed and the therapy is started. In contrast, an alternative workflow which considers also high-dimensional omics profiles and world-wide insights based on Swarm Learning would look differently. The general practitioner would still examine the overall health of the patient by routine procedures, but additionally perform a standard "CBC 2.0", which profiles the whole transcriptome of a patient, and potentially further omics layers. This data would be stored locally, for instance at a cooperating hospital with the respective IT infrastructure, which is collaborating via an international Swarm Network, thereby receiving regular updates on diverse disease models. For example, one could imagine a Swarm Network collaboratively training a pan-cancer classification model, which can predict the most likely type of cancer based on the transcriptome. Ideally, the GP would have easy access to these scores via a SL portal, could get information on these disease scores and match it with other parameters that they assessed manually, such as the overall condition of the patient. Like this, onward referral of a patient could be accelerated. Next, the hematologist could equally profit from accessing the Swarm Portal and could, in addition to standard tools such as assessing cytomorphology, cytogenetics and immunophenotyping, have access to individualized therapy response predictions based on the molecular profile of the patient. Finally, after the disease subtype is defined and a therapy plan is established, it would be

essential for such a workflow to include a follow-up evaluation e.g. one and five years after treatment.

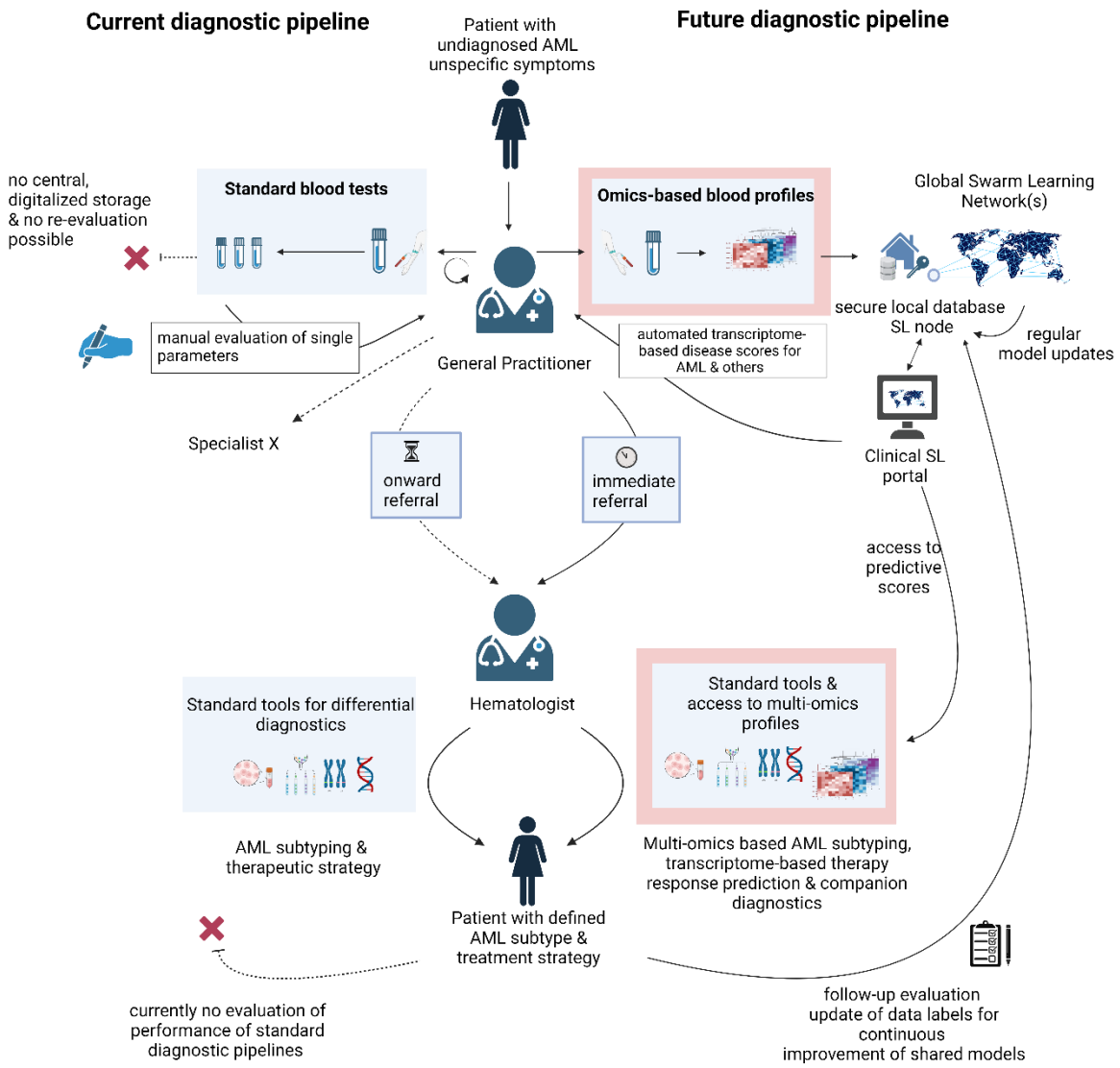


Figure 8: *Schema on current workflow of AML primary and differential diagnostics in comparison to a future, multi-omics based workflow that which is connected to a continuously learning Swarm Learning ecosystem. Like this, world-wide knowledge can be accessed and translated into local treatment decisions.*

Like this, the patient data on the finally assigned disease subtype, the response to the chosen therapy regimen, remission status and survival could be added to the secure database and these expert-proven labeled data could in turn improve the globally shared diagnostic model, without compromising sensitive information. Like this, classifier performance could be continuously evaluated, and it can be immediately intervened in case the performance is not sufficient. This stands in stark contrast to current practice, where performance measures of state-of-the-art

diagnostic methods are simply not measured or reported. Clearly, such an alternative diagnostic workflow would, besides the proposed solution for knowledge sharing without data sharing, require fundamental changes in the infrastructure and mechanisms for data collection and storage, which are beyond the scope of this work.

Finally, as with any AI-based system meant to support clinical decision making, additional ethical, regulatory, and practical considerations would need to be addressed for deployment. Importantly, the approach proposed here is not aiming to put AI in a position to outcompete human decision making, as it may regularly be framed as "human vs. AI". The perspective is rather to elucidate whether a collaborative setup between humans and AI is possible, where humans still have oversight and are integral part of the decision process ("human in the loop") (Rajpurkar et al. 2022). This is also reflected by current legal frameworks, since ultimately AI systems will not be subject to liability. In that sense, AI-guided diagnostics would be an additional, yet very powerful, set in the diagnostic toolbox, which would be no means free human from responsibility that comes with any diagnostic test, not the user, the physician, nor the producer of the application, who are liable for the performance of any diagnostic tests provided already today. In case of harm caused by AI systems, the legal responsibility would be allocated between either the producers or the users of AI, but not the AI itself. (Buiten, de Streel, and Peitz 2023). This having said, the concrete status of AI-based technology in routine diagnostics however would also be dependent on the perception and the usability for a concrete medical use-case. For example, whether results of AI algorithms are communicated as probabilities, text recommendations or by highlighting areas of interest in an image. As outlined above, the next step would be to develop prospective trials, which systematically access the diagnostic utility of the proposed approach, ideally in comparison to the current diagnostic standards, for which unfortunately no information on detection performance is reported. Standard guidelines that aim to improve completeness of reporting of clinical trials and their protocols have however been extended to also cover AI interventions (Cruz Rivera et al. 2020; Xiaoxuan Liu et al. 2020), making a systematic evaluation of such a concept possible.

Overall, the concept presented here could be translated to many other settings than AML diagnostics, it would enable improved detection of rare diseases, it could be beneficial e.g. in pandemic scenarios, and it would enable the development of system-level medical diagnostics based on global cooperation.

6 References

Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. “Deep Learning with Differential Privacy.” In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS’16*, 308–18. New York, New York, USA: ACM Press. https://doi.org/10.1145/2976749.2978318.

Abdul Aziz, Nurul Ainin, Norfilza M Mokhtar, Roslan Harun, Md Manir Hossain Mollah, Isa Mohamed Rose, Ismail Sagap, Azmi Mohd Tamil, Wan Zurinah Wan Ngah, and Rahman Jamal. 2016. “A 19-Gene Expression Signature as a Predictor of Survival in Colorectal Cancer.” *BMC Medical Genomics* 9 (1): 58. https://doi.org/10.1186/s12920-016-0218-1.

Abràmoff, Michael D, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, et al. 2013. “Automated Analysis of Retinal Images for Detection of Referable Diabetic Retinopathy.” *JAMA Ophthalmology* 131 (3): 351–57. https://doi.org/10.1001/jamaophthalmol.2013.1743.

Abràmoff, Michael D, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. 2018. “Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices.” *Npj Digital Medicine* 1 (August): 39. https://doi.org/10.1038/s41746-018-0040-6.

Acosta, Julián N, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. “Multimodal Biomedical AI.” *Nature Medicine* 28 (9): 1773–84. https://doi.org/10.1038/s41591-022-01981-2.

Acquaviva, Massimo, Ramesh Menon, Marco Di Dario, Gloria Dalla Costa, Marzia Romeo, Francesca Sangalli, Bruno Colombo, et al. 2020. “Inferring Multiple Sclerosis Stages from the Blood Transcriptome via Machine Learning.” *Cell Reports. Medicine* 1 (4): 100053. https://doi.org/10.1016/j.xcrm.2020.100053.

Alizadeh, A A, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, et al. 2000. “Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling.” *Nature* 403 (6769): 503–11. https://doi.org/10.1038/35000501.

Allam, Ahmed, Mate Nagy, George Thoma, and Michael Krauthammer. 2019. “Neural Networks versus Logistic Regression for 30 Days All-Cause Readmission Prediction.” *Scientific Reports* 9 (1): 9277. https://doi.org/10.1038/s41598-019-45685-z.

Allen, Naomi E, Cathie Sudlow, Tim Peakman, Rory Collins, and UK Biobank. 2014. “UK Biobank Data: Come and Get It.” *Science Translational Medicine* 6 (224): 224ed4. https://doi.org/10.1126/scitranslmed.3008601.

All of Us Research Program Investigators, J C Denny, J L Rutter, D B Goldstein, A Philippakis, J W Smoller, G Jenkins, and E Dishman. 2019. “The ‘All of Us’ Research Program.” *The New England Journal of Medicine* 381 (7): 668–76. https://doi.org/10.1056/NEJMsr1809937.

Almazaydeh, Laiali, Khaled Elleithy, and Miad Faezipour. 2012. “Obstructive Sleep Apnea Detection Using SVM-Based Classification of ECG Signal Features.” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2012: 4938–41. https://doi.org/10.1109/EMBC.2012.6347100.

Alpert, Jeffrey P, Allen Greiner, and Sandra Hall. 2004. “Health Fair Screening: The Clinical Utility of the Comprehensive Metabolic Profile.” *Family Medicine* 36 (7): 514–19.

Alquicira-Hernandez, Jose, Anuja Sathe, Hanlee P Ji, Quan Nguyen, and Joseph E Powell. 2019. "ScPred: Accurate Supervised Method for Cell-Type Classification from Single-Cell RNA-Seq Data." *Genome Biology* 20 (1): 264. https://doi.org/10.1186/s13059-019-1862-5.

Altman, Matthew C, Darawan Rinchai, Nicole Baldwin, Mohammed Toufiq, Elizabeth Whalen, Mathieu Garand, Basirudeen Syed Ahamed Kabeer, et al. 2021. "Development of a Fixed Module Repertoire for the Analysis and Interpretation of Blood Transcriptome Data." *Nature Communications* 12 (1): 4385. https://doi.org/10.1038/s41467-021-24584-w.

Amrollahi, Fatemeh, Supreeth P Shashikumar, Andre L Holder, and Shamim Nemati. 2022. "Leveraging Clinical Data across Healthcare Institutions for Continual Learning of Predictive Risk Models." *Scientific Reports* 12 (1): 8380. https://doi.org/10.1038/s41598-022-12497-7.

Andersson, A, C Ritz, D Lindgren, P Edén, C Lassen, J Heldrup, T Olofsson, et al. 2007. "Microarray-Based Classification of a Consecutive Series of 121 Childhood Acute Leukemias: Prediction of Leukemic and Genetic Subtype as Well as of Minimal Residual Disease Status." *Leukemia* 21 (6): 1198–1203. https://doi.org/10.1038/sj.leu.2404688.

Angermueller, Christof, Heather J Lee, Wolf Reik, and Oliver Stegle. 2017. "DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning." *Genome Biology* 18 (1): 67. https://doi.org/10.1186/s13059-017-1189-z.

Apweiler, Rolf, Tim Beissbarth, Michael R Berthold, Nils Blüthgen, Yvonne Burmeister, Olaf Dammann, Andreas Deutsch, et al. 2018. "Whither Systems Medicine?" *Experimental & Molecular Medicine* 50 (3): e453. https://doi.org/10.1038/emm.2017.290.

Araujo, Leonardo S de, Lea A I Vaas, Marcelo Ribeiro-Alves, Robert Geffers, Fernanda C Q Mello, Alexandre S de Almeida, Adriana da S R Moreira, et al. 2016. "Transcriptomic Biomarkers for Tuberculosis: Evaluation of DOCK9. EPHA4, and NPC2 MRNA Expression in Peripheral Blood." *Frontiers in Microbiology* 7 (October): 1586. https://doi.org/10.3389/fmicb.2016.01586.

Arber, Daniel A, Attilio Orazi, Robert P Hasserjian, Michael J Borowitz, Katherine R Calvo, Hans-Michael Kvasnicka, Sa A Wang, et al. 2022. "International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: Integrating Morphologic, Clinical, and Genomic Data." *Blood* 140 (11): 1200–1228. https://doi.org/10.1182/blood.2022015850.

Ardila, Diego, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, et al. 2019. "End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography." *Nature Medicine* 25 (6): 954–61. https://doi.org/10.1038/s41591-019-0447-x.

Aronson, Samuel J, and Heidi L Rehm. 2015. "Building the Foundation for Genomics in Precision Medicine." *Nature* 526 (7573): 336–42. https://doi.org/10.1038/nature15816.

Asakura, Keisuke, Tsukasa Kadota, Juntaro Matsuzaki, Yukihiro Yoshida, Yusuke Yamamoto, Kazuo Nakagawa, Satoko Takizawa, et al. 2020. "A MiRNA-Based Diagnostic Model Predicts Resectable Lung Cancer in Humans with High Accuracy." *Communications Biology* 3 (1): 134. https://doi.org/10.1038/s42003-020-0863-y.

Ashley, Euan A. 2016. "Towards Precision Medicine." *Nature Reviews. Genetics* 17 (9): 507–22. https://doi.org/10.1038/nrg.2016.86.

Attia, Zachi I, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, et al. 2019. "An Artificial Intelligence-Enabled ECG Algorithm for the Identification of Patients with Atrial Fibrillation during Sinus Rhythm: A

Retrospective Analysis of Outcome Prediction." *The Lancet* 394 (10201): 861–67. https://doi.org/10.1016/S0140-6736(19)31721-0.

Awada, Hassan, Arda Durmaz, Carmelo Gurnari, Ashwin Kishtagari, Manja Meggendorfer, Cassandra M Kerr, Teodora Kuzmanovic, et al. 2021. "Machine Learning Integrates Genomic Signatures for Subclassification beyond Primary and Secondary Acute Myeloid Leukemia." *Blood* 138 (19): 1885–95. https://doi.org/10.1182/blood.2020010603.

Ayaz, Muhammad, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, and Deris Stiawan. 2021. "The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities." *JMIR Medical Informatics* 9 (7): e21929. https://doi.org/10.2196/21929.

Babenko, Boris, Akinori Mitani, Ilana Traynis, Naho Kitade, Preeti Singh, April Y Maa, Jorge Cuadros, et al. 2022. "Detection of Signs of Disease in External Photographs of the Eyes via Deep Learning." *Nature Biomedical Engineering* 6 (12): 1370–83. https://doi.org/10.1038/s41551-022-00867-5.

Badgeley, Marcus A, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. 2019. "Deep Learning Predicts Hip Fracture Using Confounding Patient and Healthcare Variables." *Npj Digital Medicine* 2 (April): 31. https://doi.org/10.1038/s41746-019-0105-1.

Bae, Sean, Silviu Borac, Yunus Emre, Jonathan Wang, Jiang Wu, Mehr Kashyap, Si-Hyuck Kang, et al. 2022. "Prospective Validation of Smartphone-Based Heart Rate and Respiratory Rate Measurement Algorithms." *Communications Medicine* 2 (April): 40. https://doi.org/10.1038/s43856-022-00102-x.

Barakat, Nahla H, Andrew P Bradley, and Mohamed Nabil H Barakat. 2010. "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus." *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society* 14 (4): 1114–20. https://doi.org/10.1109/TITB.2009.2039485.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2018. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology* 37 (December): 38–44. https://doi.org/10.1038/nbt.4314.

Ben Azzouz, Fadoua, Bertrand Michel, Hamza Lasla, Wilfried Gouraud, Anne-Flore François, Fabien Girka, Théo Lecointre, et al. 2021. "Development of an Absolute Assignment Predictor for Triple-Negative Breast Cancer Subtyping Using Machine Learning Approaches." *Computers in Biology and Medicine* 129 (February): 104171. https://doi.org/10.1016/j.compbiomed.2020.104171.

Benjamens, Stan, Pranavsingh Dhunnoo, and Bertalan Meskó. 2020. "The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database." *Npj Digital Medicine* 3 (1): 118. https://doi.org/10.1038/s41746-020-00324-0.

Bernardes, Joana P, Neha Mishra, Florian Tran, Thomas Bahmer, Lena Best, Johanna I Blase, Dora Bordoni, et al. 2020. "Longitudinal Multi-Omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19." *Immunity* 53 (6): 1296-1314.e9. https://doi.org/10.1016/j.immuni.2020.11.017.

Bhalla, Sherry, Kumardeep Chaudhary, Ritesh Kumar, Manika Sehgal, Harpreet Kaur, Suresh Sharma, and Gajendra P S Raghava. 2017. "Gene Expression-Based Biomarkers for Discriminating Early and Late Stage of Clear Cell Renal Cancer." *Scientific Reports* 7 (March): 44997. https://doi.org/10.1038/srep44997.

6 References

Bibi, Nighat, Misba Sikandar, Ikram Ud Din, Ahmad Almogren, and Sikandar Ali. 2020. “IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning.” *Journal of Healthcare Engineering* 2020 (December): 6648574. https://doi.org/10.1155/2020/6648574.

Biswas, Dhruva, Nicolai J Birkbak, Rachel Rosenthal, Crispin T Hiley, Emilia L Lim, Krisztian Papp, Stefan Boeing, et al. 2019. “A Clonal Expression Biomarker Associates with Lung Cancer Mortality.” *Nature Medicine* 25 (10): 1540–48. https://doi.org/10.1038/s41591-019-0595-z.

Blanes-Vidal, Victoria, Katrine P Lindvig, Maja Thiele, Esmaeil S Nadimi, and Aleksander Krag. 2022. “Artificial Intelligence Outperforms Standard Blood-Based Scores in Identifying Liver Fibrosis Patients in Primary Care.” *Scientific Reports* 12 (1): 2914. https://doi.org/10.1038/s41598-022-06998-8.

Boland, B J, P C Wollan, and M D Silverstein. 1996. “Yield of Laboratory Tests for Case-Finding in the Ambulatory General Medical Examination.” *The American Journal of Medicine* 101 (2): 142–52. https://doi.org/10.1016/s0002-9343(96)80068-4.

Bonaguro, Lorenzo, Jonas Schulte-Schrepping, Caterina Carraro, Laura L Sun, Benedikt Reiz, Ioanna Gemünd, Adem Saglam, et al. 2022. “Human Variation in Population-Wide Gene Expression Data Predicts Gene Perturbation Phenotype.” *IScience* 25 (11): 105328. https://doi.org/10.1016/j.isci.2022.105328.

Bonaguro, Lorenzo, Jonas Schulte-Schrepping, Thomas Ulas, Anna C Aschenbrenner, Marc Beyer, and Joachim L Schultze. 2022. “A Guide to Systems-Level Immunomics.” *Nature Immunology* 23 (10): 1412–23. https://doi.org/10.1038/s41590-022-01309-9.

Boyle, Evan A, Yang I Li, and Jonathan K Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* 169 (7): 1177–86. https://doi.org/10.1016/j.cell.2017.05.038.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning*.

Bringsjord, Selmer, and Naveen Sundar Govindarajulu. 2022. “Artificial Intelligence.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2022. Metaphysics Research Lab, Stanford University.

Brodin, Petter, Darragh Duffy, and Lluis Quintana-Murci. 2019. “A Call for Blood-In Human Immunology.” *Immunity* 50 (6): 1335–36. https://doi.org/10.1016/j.immuni.2019.05.012.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” *ArXiv*. https://doi.org/10.48550/arxiv.2303.12712.

Buchanan, B, and G Sutherland. 1969. “Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses.” *Organic Chemistry*.

Buchanan, B G, and E H Shortliffe. 1984. “Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.”

Buchard, Albert, and Jonathan G. Richens. 2022. “Artificial Intelligence for Medical Decisions.” In *Artificial Intelligence in Medicine*, edited by Niklas Lidströmer and Hutan Ashrafian, 159–79. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_28.

Bühlmann, P, and S Van De Geer. 2011. “Statistics for High-Dimensional Data: Methods, Theory and Applications.” *Statistics for High-Dimensional Data: Methods, Theory and Applications*.

Buiten, Miriam, Alexandre de Streel, and Martin Peitz. 2023. “The Law and Economics of AI Liability.” *Computer Law & Security Review* 48 (April): 105794. https://doi.org/10.1016/j.clsr.2023.105794.

Bullinger, Lars, Konstanze Döhner, Eric Bair, Stefan Fröhling, Richard F Schlenk, Robert Tibshirani, Hartmut Döhner, and Jonathan R Pollack. 2004. “Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia.” *The New England Journal of Medicine* 350 (16): 1605–16. https://doi.org/10.1056/NEJMoa031046.

Bumgarner, Roger. 2013. “Overview of DNA Microarrays: Types, Applications, and Their Future.” *Current Protocols in Molecular Biology* Chapter 22 (January): Unit 22.1. https://doi.org/10.1002/0471142727.mb2201s101.

Butz, Martin V. 2021. “Towards Strong AI.” *KI - Künstliche Intelligenz* 35 (1): 91–101. https://doi.org/10.1007/s13218-021-00705-x.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. “Statistics versus Machine Learning.” *Nature Methods* 15 (4): 233–34. https://doi.org/10.1038/nmeth.4642.

Callaway, Ewen. 2020. “‘It Will Change Everything’: DeepMind’s AI Makes Gigantic Leap in Solving Protein Structures.” *Nature* 588 (7837): 203–4. https://doi.org/10.1038/d41586-020-03348-4.

Cancer Genome Atlas Network. 2015. “Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas.” *Nature* 517 (7536): 576–82. https://doi.org/10.1038/nature14129.

Cancer Genome Atlas Research Network, Timothy J Ley, Christopher Miller, Li Ding, Benjamin J Raphael, Andrew J Mungall, A Gordon Robertson, et al. 2013. “Genomic and Epigenomic Landscapes of Adult de Novo Acute Myeloid Leukemia.” *The New England Journal of Medicine* 368 (22): 2059–74. https://doi.org/10.1056/NEJMoa1301689.

Cancer Genome Atlas Research Network. 2015. “The Molecular Taxonomy of Primary Prostate Cancer.” *Cell* 163 (4): 1011–25. https://doi.org/10.1016/j.cell.2015.10.025.

“Can We Stop AI Outsmarting Humanity? .” 2019. The Guardian. 2019. https://www.theguardian.com/technology/2019/mar/28/can-we-stop-robots-outsmarting-humanity-artificial-intelligence-singularity.

Carraro, Caterina, Lorenzo Bonaguro, Jonas Schulte-Schrepping, Arik Horne, Marie Oestreich, Stefanie Warnat-Herresthal, Tim Helbing, et al. 2022. “Decoding Mechanism of Action and Sensitivity to Drug Candidates from Integrated Transcriptome and Chromatin State.” *ELife* 11 (August). https://doi.org/10.7554/eLife.78012.

Chan, Justin, Sharat Raju, Rajalakshmi Nandakumar, Randall Bly, and Shyamnath Gollakota. 2019. “Detecting Middle Ear Fluid Using Smartphones.” *Science Translational Medicine* 11 (492). https://doi.org/10.1126/scitranslmed.aav1102.

Chaussabel, Damien, Virginia Pascual, and Jacques Banchereau. 2010. “Assessing the Human Immune System through Blood Transcriptomics.” *BMC Biology* 8 (July): 84. https://doi.org/10.1186/1741-7007-8-84.

Chaussabel, Damien. 2015. “Assessment of Immune Status Using Blood Transcriptomics and Potential Implications for Global Health.” *Seminars in Immunology* 27 (1): 58–66. https://doi.org/10.1016/j.smim.2015.03.002.

Chen, Runpu, Le Yang, Steve Goodison, and Yijun Sun. 2020. “Deep-Learning Approach to Identifying Cancer Subtypes Using High-Dimensional Genomic Data.” *Bioinformatics* 36 (5): 1476–83. https://doi.org/10.1093/bioinformatics/btz769.

Cherlin, Svetlana, Myles J Lewis, Darren Plant, Nisha Nair, Katriona Goldmann, Evan Tzanis, Michael R Barnes, et al. 2020. “Investigation of Genetically Regulated Gene Expression and Response to Treatment in Rheumatoid Arthritis Highlights an Association between IL18RAP Expression and Treatment Response.” *Annals of the Rheumatic Diseases* 79 (11): 1446–52. https://doi.org/10.1136/annrheumdis-2020-217204.

Cheung, Carol Y, An Ran Ran, Shujun Wang, Victor T T Chan, Kaiser Sham, Saima Hilal, Narayanaswamy Venketasubramanian, et al. 2022. “A Deep Learning Model for Detection of Alzheimer’s Disease Based on Retinal Photographs: A Retrospective, Multicentre Case-Control Study.” *The Lancet. Digital Health* 4 (11): e806–15. https://doi.org/10.1016/S2589-7500(22)00169-8.

Cheung, Carol Y, Dejiang Xu, Ching-Yu Cheng, Charumathi Sabanayagam, Yih-Chung Tham, Marco Yu, Tyler Hyungtaek Rim, et al. 2021. “A Deep-Learning System for the Assessment of Cardiovascular Disease Risk via the Measurement of Retinal-Vessel Calibre.” *Nature Biomedical Engineering* 5 (6): 498–508. https://doi.org/10.1038/s41551-020-00626-4.

Chhatwal, Jagpreet, Oguzhan Alagoz, Mary J Lindstrom, Charles E Kahn, Katherine A Shaffer, and Elizabeth S Burnside. 2009. “A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis.” *American Journal of Roentgenology* 192 (4): 1117–27. https://doi.org/10.2214/AJR.07.3345.

Choi, Edward, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2017. “Using Recurrent Neural Network Models for Early Detection of Heart Failure Onset.” *Journal of the American Medical Informatics Association* 24 (2): 361–70. https://doi.org/10.1093/jamia/ocw112.

Claassen, Jan, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez, Joshua U Okonkwo, et al. 2019. “Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury.” *The New England Journal of Medicine* 380 (26): 2497–2505. https://doi.org/10.1056/NEJMoa1812757.

Codella, N C F, Q B Nguyen, S Pankanti, D A Gutman, B Helba, A C Halpern, and J R Smith. 2017. “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images.” *IBM Journal of Research and Development* 61 (4): 5:1-5:15. https://doi.org/10.1147/JRD.2017.2708299.

Cooper, Thomas A, Lili Wan, and Gideon Dreyfuss. 2009. “RNA and Disease.” *Cell* 136 (4): 777–93. https://doi.org/10.1016/j.cell.2009.02.011.

Corman, Victor M, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel Kw Chu, Tobias Bleicker, et al. 2020. “Detection of 2019 Novel Coronavirus (2019-NCoV) by Real-Time RT-PCR.” *Euro Surveillance* 25 (3). https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

Cruz Rivera, Samantha, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, Melanie J Calvert, SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, and SPIRIT-AI and CONSORT-AI Consensus Group. 2020. “Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension.” *Nature Medicine* 26 (9): 1351–63. https://doi.org/10.1038/s41591-020-1037-7.

Dai, Yu-Ting, Fan Zhang, Hai Fang, Jian-Feng Li, Gang Lu, Lu Jiang, Bing Chen, et al. 2022. “Transcriptome-Wide Subtyping of Pediatric and Adult T Cell Acute Lymphoblastic Leukemia in an

International Study of 707 Cases." *Proceedings of the National Academy of Sciences of the United States of America* 119 (15): e2120787119. https://doi.org/10.1073/pnas.2120787119.

Das, Biplab Kanti, and Himadri Sekhar Dutta. 2019. "Infection Level Identification for Leukemia Detection Using Optimized Support Vector Neural Network." *The Imaging Science Journal* 67 (8): 417–33. https://doi.org/10.1080/13682199.2019.1701172.

Davis, Mark M, Cristina M Tato, and David Furman. 2017. "Systems Immunology: Just Getting Started." *Nature Immunology* 18 (7): 725–32. https://doi.org/10.1038/ni.3768.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. 2021. "AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal." *Nature Machine Intelligence*, May. https://doi.org/10.1038/s42256-021-00338-7.

Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. 2006. "Gene Selection and Classification of Microarray Data Using Random Forest." *BMC Bioinformatics* 7 (January): 3. https://doi.org/10.1186/1471-2105-7-3.

DiStefano, Johanna K. 2018. "The Emerging Role of Long Noncoding Rnas in Human Disease." *Methods in Molecular Biology* 1706: 91–110. https://doi.org/10.1007/978-1-4939-7471-9_6.

Döhner, Hartmut, Elihu Estey, David Grimwade, Sergio Amadori, Frederick R Appelbaum, Thomas Büchner, Hervé Dombret, et al. 2017. "Diagnosis and Management of AML in Adults: 2017 ELN Recommendations from an International Expert Panel." *Blood* 129 (4): 424–47. https://doi.org/10.1182/blood-2016-08-733196.

Döhner, Hartmut, Andrew H Wei, Frederick R Appelbaum, Charles Craddock, Courtney D DiNardo, Hervé Dombret, Benjamin L Ebert, et al. 2022. "Diagnosis and Management of AML in Adults: 2022 Recommendations from an International Expert Panel on Behalf of the ELN." *Blood* 140 (12): 1345–77. https://doi.org/10.1182/blood.2022016867.

Dong, X, S Sarker, and L Qian. 2022. "Integrating Human-in-the-Loop into Swarm Learning for Decentralized Fake News Detection." *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 46.

Dove, Edward S, Yann Joly, Anne-Marie Tassé, Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, and Bartha M Knoppers. 2015. "Genomic Cloud Computing: Legal and Ethical Points to Consider." *European Journal of Human Genetics* 23 (10): 1271–78. https://doi.org/10.1038/ejhg.2014.196.

Duchmann, Matthieu, Orianne Wagner-Ballon, Thomas Boyer, Meyling Cheok, Elise Fournier, Estelle Guerin, Laurène Fenwarth, et al. 2022. "Machine Learning Identifies the Independent Role of Dysplasia in the Prediction of Response to Chemotherapy in AML." *Leukemia* 36 (3): 656–63. https://doi.org/10.1038/s41375-021-01435-7.

Dumeaux, Vanessa, Josie Ursini-Siegel, Arnar Flatberg, Hans E Fjosne, Jan-Ole Frantzen, Marit Muri Holmen, Enno Rodegerdts, Ellen Schlichting, and Eiliv Lund. 2015. "Peripheral Blood Cells Inform on the Presence of Breast Cancer: A Population-Based Case-Control Study." *International Journal of Cancer* 136 (3): 656–67. https://doi.org/10.1002/ijc.29030.

Duò, Angelo, Mark D Robinson, and Charlotte Soneson. 2018. "A Systematic Performance Evaluation of Clustering Methods for Single-Cell RNA-Seq Data." *F1000Research* 7 (July): 1141. https://doi.org/10.12688/f1000research.15666.3.

Dybowski, Richard. 2022. “Emergence of Deep Machine Learning in Medicine.” In *Artificial Intelligence in Medicine*, edited by Niklas Lidströmer and Hutan Ashrafian, 449–57. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_26.

Echle, Amelie, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. 2021. “Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers.” *British Journal of Cancer* 124 (4): 686–96. https://doi.org/10.1038/s41416-020-01122-x.

Eckardt, Jan-Niklas, Martin Bornhäuser, Karsten Wendt, and Jan Moritz Middeke. 2020. “Application of Machine Learning in the Management of Acute Myeloid Leukemia: Current Practice and Future Prospects.” *Blood Advances* 4 (23): 6077–85. https://doi.org/10.1182/bloodadvances.2020002997.

Elshafeey, Nabil, Aikaterini Kotrotsou, Ahmed Hassan, Nancy Elshafei, Islam Hassan, Sara Ahmed, Srishti Abrol, et al. 2019. “Multicenter Study Demonstrates Radiomic Features Derived from Magnetic Resonance Perfusion Images Identify Pseudoprogression in Glioblastoma.” *Nature Communications* 10 (1): 3170. https://doi.org/10.1038/s41467-019-11007-0.

El Emam, Khaled. 2011. “Methods for the De-Identification of Electronic Health Records for Genomic Research.” *Genome Medicine* 3 (4): 25. https://doi.org/10.1186/gm239.

Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J Theis. 2019. “Deep Learning: New Computational Modelling Techniques for Genomics.” *Nature Reviews. Genetics* 20 (7): 389–403. https://doi.org/10.1038/s41576-019-0122-6.

Esmeral, Laura Carolina Martinez, and Andreas Uhl. 2022. “Patient Identification Methods Based on Medical Imagery and Their Impact on Patient Privacy and Open Medical Data.” In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 406–11. IEEE. https://doi.org/10.1109/CBMS55023.2022.00079.

Esteva, Andre, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. “A Guide to Deep Learning in Healthcare.” *Nature Medicine* 25 (1): 24–29. https://doi.org/10.1038/s41591-018-0316-z.

Falini, Brunangelo, and Maria Paola Martelli. 2023. “Comparison of the International Consensus and 5th WHO Edition Classifications of Adult Myelodysplastic Syndromes and Acute Myeloid Leukemia.” *American Journal of Hematology* 98 (3): 481–92. https://doi.org/10.1002/ajh.26812.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, et al. 2015. “Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants.” *Nature* 518 (7539): 337–43. https://doi.org/10.1038/nature13835.

Fatma, Mashiat, and Jaya Sharma. 2014. “Identification and Classification of Acute Leukemia Using Neural Network.” In *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 142–45. IEEE. https://doi.org/10.1109/MedCom.2014.7005992.

FDA. 2018. “FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems.” April 11, 2018. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

Figgett, William A, Katherine Monaghan, Milica Ng, Monther Alhamdoosh, Eugene Maraskovsky, Nicholas J Wilson, Alberta Y Hoi, Eric F Morand, and Fabienne Mackay. 2019. “Machine Learning Applied to Whole-Blood RNA-Sequencing Data Uncovers Distinct Subsets of Patients with Systemic Lupus Erythematosus.” *Clinical & Translational Immunology* 8 (12): e01093. https://doi.org/10.1002/cti2.1093.

Frid-Adar, Maayan, Rula Amer, Ophir Gozes, Jannette Nassar, and Hayit Greenspan. 2021. “COVID-19 in CXR: From Detection and Severity Scoring to Patient Disease Monitoring.” *IEEE Journal of Biomedical and Health Informatics* 25 (6): 1892–1903. https://doi.org/10.1109/JBHI.2021.3069169.

Fridman, Lex, Daniel E. Brown, Michael Glazer, William Angell, Spencer Dodd, Benedikt Jenik, Jack Terwilliger, et al. 2019. “MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation.” *IEEE Access : Practical Innovations, Open Solutions* 7: 102021–38. https://doi.org/10.1109/ACCESS.2019.2926040.

Froelicher, David, Juan R Troncoso-Pastoriza, Jean Louis Raisaro, Michel A Cuendet, Joao Sa Sousa, Hyunghoon Cho, Bonnie Berger, Jacques Fellay, and Jean-Pierre Hubaux. 2021. “Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption.” *Nature Communications* 12 (1): 5910. https://doi.org/10.1038/s41467-021-25972-y.

Frost, H Robert, and Christopher I Amos. 2018. “A Multi-Omics Approach for Identifying Important Pathways and Genes in Human Cancer.” *BMC Bioinformatics* 19 (1): 479. https://doi.org/10.1186/s12859-018-2476-8.

Fu, Yu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. 2020. “Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis.” *Nature Cancer* 1 (8): 800–810. https://doi.org/10.1038/s43018-020-0085-8.

Future of Life Institute. 2023. “Pause Giant AI Experiments: An Open Letter.” 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Galen, Peter van, Volker Hovestadt, Marc H Wadsworth Ii, Travis K Hughes, Gabriel K Griffin, Sofia Battaglia, Julia A Verga, et al. 2019. “Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity.” *Cell* 176 (6): 1265-1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.

“General Data Protection Regulation.” 2023. 2023. https://gdpr.eu/tag/gdpr/.

George-Gay, Beverly, and Katherine Parker. 2003. “Understanding the Complete Blood Count with Differential.” *Journal of Perianesthesia Nursing : Official Journal of the American Society of PeriAnesthesia Nurses / American Society of PeriAnesthesia Nurses* 18 (2): 96–114; quiz 115. https://doi.org/10.1053/jpan.2003.50013.

Ghosh, Debashis, and Arul M Chinnaiyan. 2005. “Classification and Selection of Biomarkers in Genomic Data Using LASSO.” *Journal of Biomedicine & Biotechnology* 2005 (2): 147–54. https://doi.org/10.1155/JBB.2005.147.

Golub, T R, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, et al. 1999. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science* 286 (5439): 531–37. https://doi.org/10.1126/science.286.5439.531.

Gong, Dexin, Lianlian Wu, Jun Zhang, Ganggang Mu, Lei Shen, Jun Liu, Zhengqiang Wang, et al. 2020. "Detection of Colorectal Adenomas with a Real-Time Computer-Aided System (ENDOANGEL): A Randomised Controlled Study." *The Lancet. Gastroenterology & Hepatology* 5 (4): 352–61. https://doi.org/10.1016/S2468-1253(19)30413-3.

Goodfellow, I, Y Bengio, and A Courville. 2016. "Deep Learning." *Deep Learning*.

Goswami, Rashmi S, Mahadeo A Sukhai, Mariam Thomas, Patricia P Reis, and Suzanne Kamel-Reid. 2009. "Applications of Microarray Technology to Acute Myelogenous Leukemia." *Cancer Informatics* 7: 13–28.

Guha Roy, Abhijit, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, et al. 2022. "Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions." *Medical Image Analysis* 75 (January): 102274. https://doi.org/10.1016/j.media.2021.102274.

Gunčar, Gregor, Matjaž Kukar, Mateja Notar, Miran Brvar, Peter Černelč, Manca Notar, and Marko Notar. 2018. "An Application of Machine Learning to Haematological Diagnosis." *Scientific Reports* 8 (1): 411. https://doi.org/10.1038/s41598-017-18564-8.

Haenssle, H A, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, et al. 2018. "Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists." *Annals of Oncology* 29 (8): 1836–42. https://doi.org/10.1093/annonc/mdy166.

Haferlach, T, A Kohlmann, U Bacher, S Schnittger, C Haferlach, and W Kern. 2007. "Gene Expression Profiling for the Diagnosis of Acute Leukaemia." *British Journal of Cancer* 96 (4): 535–40. https://doi.org/10.1038/sj.bjc.6603495.

Haferlach, Torsten, Alexander Kohlmann, Susanne Schnittger, Martin Dugas, Wolfgang Hiddemann, Wolfgang Kern, and Claudia Schoch. 2005. "Global Approach to the Diagnosis of Leukemia Using Gene Expression Profiling." *Blood* 106 (4): 1189–98. https://doi.org/10.1182/blood-2004-12-4938.

Haferlach, Torsten, Alexander Kohlmann, Lothar Wieczorek, Giuseppe Basso, Geertruy Te Kronnie, Marie-Christine Béné, John De Vos, et al. 2010. "Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report from the International Microarray Innovations in Leukemia Study Group." *Journal of Clinical Oncology* 28 (15): 2529–37. https://doi.org/10.1200/JCO.2009.23.4732.

Halbeisen, R E, A Galgano, T Scherrer, and A P Gerber. 2008. "Post-Transcriptional Gene Regulation: From Genome-Wide Studies to Principles." *Cellular and Molecular Life Sciences* 65 (5): 798–813. https://doi.org/10.1007/s00018-007-7447-6.

Hendrycks, Dan, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, et al. 2021. "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization." In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8320–29. IEEE. https://doi.org/10.1109/ICCV48922.2021.00823.

Hinton, Geoffrey, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The

Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29 (6): 82–97. https://doi.org/10.1109/MSP.2012.2205597.

"HIPAA and Administrative Simplification." 2022. April 25, 2022. https://www.cms.gov/regulations-and-guidance/administrative-simplification/hipaa-aca.

Hoadley, Katherine A, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, et al. 2018. "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer." *Cell* 173 (2): 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

Hou, Jun, Gertine van Oord, Zwier M A Groothuismink, Mark A A Claassen, Kim Kreefft, Fatiha Zaaraoui-Boutahar, Wilfred F J van IJcken, et al. 2014. "Gene Expression Profiling to Predict and Assess the Consequences of Therapy-Induced Virus Eradication in Chronic Hepatitis C Virus Infection." *Journal of Virology* 88 (21): 12254–64. https://doi.org/10.1128/JVI.00775-14.

Huang, Peng, Cheng T Lin, Yuliang Li, Martin C Tammemagi, Malcolm V Brock, Sukhinder Atkar-Khattra, Yanxun Xu, et al. 2019. "Prediction of Lung Cancer Risk at Follow-up Screening with Low-Dose CT: A Training and Validation Study of a Deep Learning Method." *The Lancet. Digital Health* 1 (7): e353–62. https://doi.org/10.1016/S2589-7500(19)30159-1.

Huang, Shih-Cheng, Akshay S Chaudhari, Curtis P Langlotz, Nigam Shah, Serena Yeung, and Matthew P Lungren. 2022. "Developing Medical Imaging AI for Emerging Infectious Diseases." *Nature Communications* 13 (1): 7060. https://doi.org/10.1038/s41467-022-34234-4.

Huang, Shujun, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. 2018. "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics." *Cancer Genomics & Proteomics* 15 (1): 41–51. https://doi.org/10.21873/cgp.20063.

Hu, Yongli, Takeshi Hase, Hui Peng Li, Shyam Prabhakar, Hiroaki Kitano, See Kiong Ng, Samik Ghosh, and Lawrence Jin Kiat Wee. 2016. "A Machine Learning Approach for the Identification of Key Markers Involved in Brain Development from Single-Cell Transcriptomic Data." *BMC Genomics* 17 (Suppl 13): 1025. https://doi.org/10.1186/s12864-016-3317-7.

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Irizarry, R A, B Hobbs, F Collin, Y D Beazer-Barclay, K J Antonellis, U Scherf, and T P Speed. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics* 4 (2): 249–64. https://doi.org/10.1093/biostatistics/4.2.249.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021a. "Statistical Learning." In *An Introduction to Statistical Learning: With Applications in R*, 15–57. Springer Texts in Statistics. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1418-1_2.

———. 2021b. "Statistical Learning." In *An Introduction to Statistical Learning: With Applications in R*, 15–57. Springer Texts in Statistics. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1418-1_2.

Javed, Abdul Rehman, Habib Ullah Khan, Mohammad Kamel Bader Alomari, Muhammad Usman Sarwar, Muhammad Asim, Ahmad S Almadhor, and Muhammad Zahid Khan. 2023. "Toward Explainable AI-Empowered Cognitive Health Assessment." *Frontiers in Public Health* 11 (March): 1024195. https://doi.org/10.3389/fpubh.2023.1024195.

Jha, Krishna Kumar, Prasanta Das, and Himadri Sekhar Dutta. 2020. “FAB Classification Based Leukemia Identification and Prediction Using Machine Learning.” In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–6. IEEE. https://doi.org/10.1109/ICSCAN49426.2020.9262388.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Ju, Jingyue, Dae Hyun Kim, Lanrong Bi, Qinglin Meng, Xiaopeng Bai, Zengmin Li, Xiaoxu Li, et al. 2006. “Four-Color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (52): 19635–40. https://doi.org/10.1073/pnas.0609513103.

Kaissis, Georgios, Sebastian Ziegelmayer, Fabian Lohöfer, Katja Steiger, Hana Algül, Alexander Muckenhuber, Hsi-Yu Yen, et al. 2019. “A Machine Learning Algorithm Predicts Molecular Subtypes in Pancreatic Ductal Adenocarcinoma with Differential Response to Gemcitabine-Based versus FOLFIRINOX Chemotherapy.” *Plos One* 14 (10): e0218642. https://doi.org/10.1371/journal.pone.0218642.

Kaissis, Georgios A., Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. 2020. “Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging.” *Nature Machine Intelligence*, June. https://doi.org/10.1038/s42256-020-0186-1.

Kandul, Serhiy, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, François Fleuret, and Markus Christen. 2023. “Explainable AI: A Review of the Empirical Literature.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4325219.

Keane, Pearse A, and Eric J Topol. 2018. “With an Eye to AI and Autonomous Diagnosis.” *Npj Digital Medicine* 1 (August): 40. https://doi.org/10.1038/s41746-018-0048-y.

Kelley, David R, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. 2018. “Sequential Regulatory Activity Prediction across Chromosomes with Convolutional Neural Networks.” *Genome Research* 28 (5): 739–50. https://doi.org/10.1101/gr.227819.117.

Kermany, Daniel S, Michael Goldbaum, Wenjia Cai, Carolina C S Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, et al. 2018. “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.” *Cell* 172 (5): 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.

Khoury, Joseph D, Eric Solary, Oussama Abla, Yassmine Akkari, Rita Alaggio, Jane F Apperley, Rafael Bejar, et al. 2022. “The 5th Edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms.” *Leukemia* 36 (7): 1703–19. https://doi.org/10.1038/s41375-022-01613-1.

Kildow, Beau J, Vasili Karas, Elizabeth Howell, Cynthia L Green, William T Baumgartner, Colin T Penrose, Michael P Bolognesi, and Thorsten M Seyler. 2018. “The Utility of Basic Metabolic Panel Tests after Total Joint Arthroplasty.” *The Journal of Arthroplasty* 33 (9): 2752–58. https://doi.org/10.1016/j.arth.2018.05.003.

Kiyasseh, Dani, Tingting Zhu, and David Clifton. 2021. “A Clinical Deep Learning Framework for Continually Learning from Cardiac Signals across Diseases, Time, Modalities, and Institutions.” *Nature Communications* 12 (1): 4221. https://doi.org/10.1038/s41467-021-24483-0.

Knottnerus, J A. 1991. “Medical Decision Making by General Practitioners and Specialists.” *Family Practice* 8 (4): 305–7. https://doi.org/10.1093/fampra/8.4.305.

Ko, Emily R, Casandra W Philipson, Thomas W Burke, Regina Z Cer, Kimberly A Bishop-Lilly, Logan J Voegtly, Ephraim L Tsalik, Christopher W Woods, Danielle V Clark, and Kevin L Schully. 2019. “Direct-from-Blood RNA Sequencing Identifies the Cause of Post-Bronchoscopy Fever.” *BMC Infectious Diseases* 19 (1): 905. https://doi.org/10.1186/s12879-019-4462-9.

Konečný, Jakub, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. “Federated Optimization: Distributed Machine Learning for On-Device Intelligence.” *ArXiv*, October.

Krämer, Benjamin, Rainer Knoll, Lorenzo Bonaguro, Michael ToVinh, Jan Raabe, Rosario Astaburuaga-García, Jonas Schulte-Schrepping, et al. 2021. “Early IFN-α Signatures and Persistent Dysfunction Are Distinguishing Features of NK Cells in Severe COVID-19.” *Immunity* 54 (11): 2650-2669.e14. https://doi.org/10.1016/j.immuni.2021.09.002.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

Kuiper, R, A Broyl, Y de Knegt, M H van Vliet, E H van Beers, B van der Holt, L el Jarari, et al. 2012. “A Gene Expression Signature for High-Risk Multiple Myeloma.” *Leukemia* 26 (11): 2406–13. https://doi.org/10.1038/leu.2012.127.

Kukar, Matjaž, Gregor Gunčar, Tomaž Vovko, Simon Podnar, Peter Černelč, Miran Brvar, Mateja Zalaznik, Mateja Notar, Sašo Moškon, and Marko Notar. 2021. “COVID-19 Diagnosis by Routine Blood Tests Using Machine Learning.” *Scientific Reports* 11 (1): 10738. https://doi.org/10.1038/s41598-021-90265-9.

Kukurba, Kimberly R, and Stephen B Montgomery. 2015. “RNA Sequencing and Analysis.” *Cold Spring Harbor Protocols* 2015 (11): 951–69. https://doi.org/10.1101/pdb.top084970.

Kulenovic, Adnan, and Azra Lagumdzija-Kulenovic. 2022. “Using Logistic Regression to Predict Long COVID Conditions in Chronic Patients.” *Studies in Health Technology and Informatics* 295 (June): 265–68. https://doi.org/10.3233/SHTI220713.

Laguarta, Jordi, Ferran Hueto, and Brian Subirana. 2020. “COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings.” *IEEE Open Journal of Engineering in Medicine and Biology* 1 (September): 275–81. https://doi.org/10.1109/OJEMB.2020.3026928.

Lakhani, Paras, and Baskaran Sundaram. 2017. “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks.” *Radiology* 284 (2): 574–82. https://doi.org/10.1148/radiol.2017162326.

Langmead, Ben, and Abhinav Nellore. 2018. “Cloud Computing for Genomic Data Analysis and Collaboration.” *Nature Reviews. Genetics* 19 (4): 208–19. https://doi.org/10.1038/nrg.2017.113.

Lauritsen, Simon Meyer, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. 2020. “Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Electronic Health Records.” *Nature Communications* 11 (1): 3852. https://doi.org/10.1038/s41467-020-17431-x.

LeCun, Y, Y Bengio, and G Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” *Bioinformatics* 36 (4): 1234–40. https://doi.org/10.1093/bioinformatics/btz682.

Lee, Su-In, Safiye Celik, Benjamin A Logsdon, Scott M Lundberg, Timothy J Martins, Vivian G Oehler, Elihu H Estey, et al. 2018. “A Machine Learning Approach to Integrate Big Data for Precision Medicine in Acute Myeloid Leukemia.” *Nature Communications* 9 (1): 42. https://doi.org/10.1038/s41467-017-02465-5.

Leng, Dongjin, Linyi Zheng, Yuqi Wen, Yunhao Zhang, Lianlian Wu, Jing Wang, Meihong Wang, Zhongnan Zhang, Song He, and Xiaochen Bo. 2022. “A Benchmark Study of Deep Learning-Based Multi-Omics Data Fusion Methods for Cancer.” *Genome Biology* 23 (1): 171. https://doi.org/10.1186/s13059-022-02739-2.

Leo, Patrick, Andrew Janowczyk, Robin Elliott, Nafiseh Janaki, Kaustav Bera, Rakesh Shiradkar, Xavier Farré, et al. 2021. “Computer Extracted Gland Features from H&E Predicts Prostate Cancer Recurrence Comparably to a Genomic Companion Diagnostic Test: A Large Multi-Site Study.” *NPJ Precision Oncology* 5 (1): 35. https://doi.org/10.1038/s41698-021-00174-3.

Leong, Samantha, Yue Zhao, Noyal M Joseph, Natasha S Hochberg, Sonali Sarkar, Jane Pleskunas, David Hom, et al. 2018. “Existing Blood Transcriptional Classifiers Accurately Discriminate Active Tuberculosis from Latent Infection in Individuals from South India.” *Tuberculosis* 109 (January): 41–51. https://doi.org/10.1016/j.tube.2018.01.002.

Libbrecht, Maxwell W, and William Stafford Noble. 2015. “Machine Learning Applications in Genetics and Genomics.” *Nature Reviews. Genetics* 16 (6): 321–32. https://doi.org/10.1038/nrg3920.

Li, Yang I, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. 2016. “RNA Splicing Is a Primary Link between Genetic Variation and Disease.” *Science* 352 (6285): 600–604. https://doi.org/10.1126/science.aad9417.

Lidströmer, Niklas, Federica Aresu, and Hutan Ashrafian. 2022. “Basic Concepts of Artificial Intelligence: Primed for Clinicians.” In *Artificial Intelligence in Medicine*, edited by Niklas Lidströmer and Hutan Ashrafian, 3–20. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_1.

Linmans, Jasper, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. 2023. “Predictive Uncertainty Estimation for Out-of-Distribution Detection in Digital Pathology.” *Medical Image Analysis* 83 (January): 102655. https://doi.org/10.1016/j.media.2022.102655.

Lin, Steven Y, Megan R Mahoney, and Christine A Sinsky. 2019. “Ten Ways Artificial Intelligence Will Transform Primary Care.” *Journal of General Internal Medicine* 34 (8): 1626–30. https://doi.org/10.1007/s11606-019-05035-1.

Liu, Xiaoxuan, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, and SPIRIT-AI and CONSORT-AI Working Group. 2020. “Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension.” *Nature Medicine* 26 (9): 1364–74. https://doi.org/10.1038/s41591-020-1034-x.

Li, Xiaoxiao, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. 2020. “Multi-Site FMRI Analysis Using Privacy-Preserving Federated Learning and Domain Adaptation: ABIDE Results.” *Medical Image Analysis* 65 (October): 101765. https://doi.org/10.1016/j.media.2020.101765.

Li, Zejuan, Tobias Herold, Chunjiang He, Peter J M Valk, Ping Chen, Vindi Jurinovic, Ulrich Mansmann, et al. 2013. “Identification of a 24-Gene Prognostic Signature That Improves the European LeukemiaNet Risk Classification of Acute Myeloid Leukemia: An International Collaborative Study.” *Journal of Clinical Oncology* 31 (9): 1172–81. https://doi.org/10.1200/JCO.2012.44.3184.

Liu, Xuan, Emily Speranza, César Muñoz-Fontela, Sam Haldenby, Natasha Y Rickett, Isabel Garcia-Dorival, Yongxiang Fang, et al. 2017. “Transcriptomic Signatures Differentiate Survival from Fatal Outcomes in Humans Infected with Ebola Virus.” *Genome Biology* 18 (1): 4. https://doi.org/10.1186/s13059-016-1137-3.

Liu, Yi-Rong, Yi-Zhou Jiang, Xiao-En Xu, Ke-Da Yu, Xi Jin, Xin Hu, Wen-Jia Zuo, et al. 2016. “Comprehensive Transcriptome Analysis Identifies Novel Molecular Subtypes and Subtype-Specific RNAs of Triple-Negative Breast Cancer.” *Breast Cancer Research* 18 (1): 33. https://doi.org/10.1186/s13058-016-0690-8.

Lockhart, D J, H Dong, M C Byrne, M T Follettie, M V Gallo, M S Chee, M Mittmann, et al. 1996. “Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays.” *Nature Biotechnology* 14 (13): 1675–80. https://doi.org/10.1038/nbt1296-1675.

Luo, Yuan, Peter Szolovits, Anand S Dighe, and Jason M Baron. 2016. “Using Machine Learning to Predict Laboratory Test Results.” *American Journal of Clinical Pathology* 145 (6): 778–88. https://doi.org/10.1093/ajcp/aqw064.

Lu, Haonan, Mubarik Arshad, Andrew Thornton, Giacomo Avesani, Paula Cunnea, Ed Curry, Fahdi Kanavati, et al. 2019. “A Mathematical-Descriptor of Tumor-Mesoscopic-Structure from Computed-Tomography Images Annotates Prognostic- and Molecular-Phenotypes of Epithelial Ovarian Cancer.” *Nature Communications* 10 (1): 764. https://doi.org/10.1038/s41467-019-08718-9.

Magnuson, Brian, Karan Bedi, and Mats Ljungman. 2016. “Genome Stability versus Transcript Diversity.” *DNA Repair* 44 (August): 81–86. https://doi.org/10.1016/j.dnarep.2016.05.010.

Mahajan, Prashant, Nathan Kuppermann, Asuncion Mejias, Nicolas Suarez, Damien Chaussabel, T Charles Casper, Bennett Smith, et al. 2016. “Association of RNA Biosignatures with Bacterial Infections in Febrile Infants Aged 60 Days or Younger.” *The Journal of the American Medical Association* 316 (8): 846–57. https://doi.org/10.1001/jama.2016.9207.

MAQC Consortium, Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, et al. 2006. “The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements.” *Nature Biotechnology* 24 (9): 1151–61. https://doi.org/10.1038/nbt1239.

Maringe, Camille, James Spicer, Melanie Morris, Arnie Purushotham, Ellen Nolte, Richard Sullivan, Bernard Rachet, and Ajay Aggarwal. 2020. “The Impact of the COVID-19 Pandemic on Cancer Deaths Due to Delays in Diagnosis in England, UK: A National, Population-Based, Modelling Study.” *The Lancet Oncology* 21 (8): 1023–34. https://doi.org/10.1016/S1470-2045(20)30388-0.

Marshall, Eliot. 2004. “Getting the Noise out of Gene Arrays.” *Science* 306 (5696): 630–31. https://doi.org/10.1126/science.306.5696.630.

Mårtensson, Gustav, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, et al. 2020. “The Reliability of a Deep Learning Model in Clinical Out-of-Distribution MRI Data: A Multicohort Study.” *Medical Image Analysis* 66 (December): 101714. https://doi.org/10.1016/j.media.2020.101714.

McCall, Becky. 2018. “What Does the GDPR Mean for the Medical Community?” *The Lancet* 391 (10127): 1249–50. https://doi.org/10.1016/S0140-6736(18)30739-6.

McCullagh, P, and J Nelder. 1989. “Generalized Linear Models Second Edition Chapman & Hall.”

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. “International Evaluation of an AI System for Breast Cancer Screening.” *Nature* 577 (7788): 89–94. https://doi.org/10.1038/s41586-019-1799-6.

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera Arcas. 2016a. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” *ArXiv*, February.

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016b. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” *ArXiv*. https://doi.org/10.48550/arxiv.1602.05629.

Mei, Xueyan, Hao-Chih Lee, Kai-Yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, et al. 2020. “Artificial Intelligence-Enabled Rapid Diagnosis of Patients with COVID-19.” *Nature Medicine* 26 (8): 1224–28. https://doi.org/10.1038/s41591-020-0931-3.

Mellors, Theodore, Johanna B. Withers, Asher Ameli, Alex Jones, Mengran Wang, Lixia Zhang, Helia N. Sanchez, et al. 2020. “Clinical Validation of a Blood-Based Predictive Test for Stratification of Response to Tumor Necrosis Factor Inhibitor Therapies in Rheumatoid Arthritis Patients.” *Network and Systems Medicine* 3 (1): 91–104. https://doi.org/10.1089/nsm.2020.0007.

Meskó, Bertalan, and Marton Görög. 2020. “A Short Guide for Medical Professionals in the Era of Artificial Intelligence.” *Npj Digital Medicine* 3 (September): 126. https://doi.org/10.1038/s41746-020-00333-z.

“Microarray Probe Mapping.” 2023. 2023. https://www.ensembl.org/info/genome/microarray_probe_set_mapping.html.

Miklos, George L Gabor, and Ryszard Maleszka. 2004. “Microarray Reality Checks in the Context of a Complex Disease.” *Nature Biotechnology* 22 (5): 615–21. https://doi.org/10.1038/nbt965.

Miller, Ruth R, Vincent Montoya, Jennifer L Gardy, David M Patrick, and Patrick Tang. 2013. “Metagenomics for Pathogen Detection in Public Health.” *Genome Medicine* 5 (9): 81. https://doi.org/10.1186/gm485.

Montgomery, Stephen B, Jonathan A Bernstein, and Matthew T Wheeler. 2022. “Toward Transcriptomics as a Primary Tool for Rare Disease Investigation.” *Molecular Case Studies* 8 (2). https://doi.org/10.1101/mcs.a006198.

Mukhtorov, Doniyorjon, Madinakhon Rakhmonova, Shakhnoza Muksimova, and Young-Im Cho. 2023. “Endoscopic Image Classification Based on Explainable Deep Learning.” *Sensors (Basel, Switzerland)* 23 (6). https://doi.org/10.3390/s23063176.

Müller, Vincent C. 2021. “Ethics of Artificial Intelligence and Robotics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University.

Mullis, K, F Faloona, S Scharf, R Saiki, G Horn, and H Erlich. 1986. “Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction.” *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1: 263–73. https://doi.org/10.1101/SQB.1986.051.01.032.

Nakauma-González, J Alberto, Maud Rijnders, Job van Riet, Michiel S van der Heijden, Jens Voortman, Edwin Cuppen, Niven Mehra, et al. 2022. “Comprehensive Molecular Characterization Reveals Genomic and Transcriptomic Subtypes of Metastatic Urothelial Carcinoma.” *European Urology* 81 (4): 331–36. https://doi.org/10.1016/j.eururo.2022.01.026.

Nembrini, Stefano, Inke R König, and Marvin N Wright. 2018. “The Revival of the Gini Importance?” *Bioinformatics* 34 (21): 3711–18. https://doi.org/10.1093/bioinformatics/bty373.

Nichita, C, L Ciarloni, S Monnier-Benoit, S Hosseinian, G Dorta, and C Rüegg. 2014. “A Novel Gene Expression Signature in Peripheral Blood Mononuclear Cells for Early Detection of Colorectal Cancer.” *Alimentary Pharmacology & Therapeutics* 39 (5): 507–17. https://doi.org/10.1111/apt.12618.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819346.

Novakovsky, Gherman, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. 2023. “Obtaining Genetics Insights from Deep Learning via Explainable Artificial Intelligence.” *Nature Reviews. Genetics* 24 (2): 125–37. https://doi.org/10.1038/s41576-022-00532-2.

Obermeyer, Ziad, and Ezekiel J Emanuel. 2016. “Predicting the Future - Big Data, Machine Learning, and Clinical Medicine.” *The New England Journal of Medicine* 375 (13): 1216–19. https://doi.org/10.1056/NEJMp1606181.

“ODELIA .” 2023. ODELIA Revolutionizing Medical AI With Swarm Learning. 2023. https://odelia.ai/.

Oelen, Roy, Dylan H de Vries, Harm Brugge, M Grace Gordon, Martijn Vochteloo, single-cell eQTLGen consortium, BIOS Consortium, et al. 2022. “Single-Cell RNA-Sequencing of Peripheral Blood Mononuclear Cells Reveals Widespread, Context-Specific Gene Expression Regulation upon Pathogenic Exposure.” *Nature Communications* 13 (1): 3267. https://doi.org/10.1038/s41467-022-30893-5.

Oestreich, Marie, Dingfan Chen, Joachim L Schultze, Mario Fritz, and Matthias Becker. 2021. “Privacy Considerations for Sharing Genomics Data.” *EXCLI Journal* 20 (July): 1243–60. https://doi.org/10.17179/excli2021-4002.

Pan, Liyan, Guangjian Liu, Fangqin Lin, Shuling Zhong, Huimin Xia, Xin Sun, and Huiying Liang. 2017. “Machine Learning Applications for Prediction of Relapse in Childhood Acute Lymphoblastic Leukemia.” *Scientific Reports* 7 (1): 7402. https://doi.org/10.1038/s41598-017-07408-0.

Papaemmanuil, Elli, Moritz Gerstung, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Nicola D Roberts, Nicola E Potter, et al. 2016. “Genomic Classification and Prognosis in Acute Myeloid Leukemia.” *The New England Journal of Medicine* 374 (23): 2209–21. https://doi.org/10.1056/NEJMoa1516192.

Pati, Sarthak, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, et al. 2022. “Federated Learning Enables Big Data for Rare Cancer Boundary Detection.” *Nature Communications* 13 (1): 7346. https://doi.org/10.1038/s41467-022-33407-5.

Perez, Richard K, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, et al. 2022. “Single-Cell RNA-Seq Reveals Cell Type-Specific Molecular and Genetic Associations to Lupus.” *Science* 376 (6589): eabf1970. https://doi.org/10.1126/science.abf1970.

Piasecka, Barbara, Darragh Duffy, Alejandra Urrutia, Hélène Quach, Etienne Patin, Céline Posseme, Jacob Bergstedt, et al. 2018. “Distinctive Roles of Age, Sex, and Genetics in Shaping Transcriptional Variation of Human Immune Responses to Microbial Challenges.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (3): E488–97. https://doi.org/10.1073/pnas.1714765115.

Piazza, Ilaria, Nigel Beaton, Roland Bruderer, Thomas Knobloch, Crystel Barbisan, Lucie Chandat, Alexander Sudau, et al. 2020. “A Machine Learning-Based Chemoproteomic Approach to Identify Drug Targets and Binding Sites in Complex Proteomes.” *Nature Communications* 11 (1): 4200. https://doi.org/10.1038/s41467-020-18071-x.

Piccialli, Francesco, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. 2021. “A Survey on Deep Learning in Medicine: Why, How and When?” *Information Fusion* 66 (February): 111–37. https://doi.org/10.1016/j.inffus.2020.09.006.

Ping, Peipei, Henning Hermjakob, Jennifer S Polson, Panagiotis V Benos, and Wei Wang. 2018. “Biomedical Informatics on the Cloud: A Treasure Hunt for Advancing Cardiovascular Medicine.” *Circulation Research* 122 (9): 1290–1301. https://doi.org/10.1161/CIRCRESAHA.117.310967.

Podnar, Simon, Matjaž Kukar, Gregor Gunčar, Mateja Notar, Nina Gošnjak, and Marko Notar. 2019. “Diagnosing Brain Tumours by Routine Blood Tests Using Machine Learning.” *Scientific Reports* 9 (1): 14481. https://doi.org/10.1038/s41598-019-51147-3.

Porumb, Mihaela, Saverio Stranges, Antonio Pescapè, and Leandro Pecchia. 2020. “Precision Medicine and Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events Detection Based on ECG.” *Scientific Reports* 10 (1): 170. https://doi.org/10.1038/s41598-019-56927-5.

Potter, S Steven. 2018. “Single-Cell RNA Sequencing for the Study of Development, Physiology and Disease.” *Nature Reviews. Nephrology* 14 (8): 479–92. https://doi.org/10.1038/s41581-018-0021-7.

Preuer, Kristina, Richard P I Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. 2018. “DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning.” *Bioinformatics* 34 (9): 1538–46. https://doi.org/10.1093/bioinformatics/btx806.

Price, W Nicholson, and I Glenn Cohen. 2019. “Privacy in the Age of Medical Big Data.” *Nature Medicine* 25 (1): 37–43. https://doi.org/10.1038/s41591-018-0272-7.

Prior, Fred, Kirk Smith, Ashish Sharma, Justin Kirby, Lawrence Tarbox, Ken Clark, William Bennett, Tracy Nolan, and John Freymann. 2017. “The Public Cancer Radiology Imaging Collections of The Cancer Imaging Archive.” *Scientific Data* 4 (September): 170124. https://doi.org/10.1038/sdata.2017.124.

Rajewsky, Nikolaus, Geneviève Almouzni, Stanislaw A Gorski, Stein Aerts, Ido Amit, Michela G Bertero, Christoph Bock, et al. 2020. “LifeTime and Improving European Healthcare through Cell-Based Interceptive Medicine.” *Nature* 587 (7834): 377–86. https://doi.org/10.1038/s41586-020-2715-9.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, et al. 2018. “Scalable and Accurate Deep Learning with Electronic Health Records.” *Npj Digital Medicine* 1 (May): 18. https://doi.org/10.1038/s41746-018-0029-1.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. “AI in Health and Medicine.” *Nature Medicine* 28 (1): 31–38. https://doi.org/10.1038/s41591-021-01614-0.

Ramtohul, Toulsie, Luc Cabel, Xavier Paoletti, Laurent Chiche, Pauline Moreau, Aurélien Noret, Perrine Vuagnat, et al. 2020. “Quantitative CT Extent of Lung Damage in COVID-19 Pneumonia Is an Independent Risk Factor for Inpatient Mortality in a Population of Cancer Patients: A Prospective Study.” *Frontiers in Oncology* 10 (September): 1560. https://doi.org/10.3389/fonc.2020.01560.

Ravanidis, Stylianos, Anastasia Bougea, Nikolaos Papagiannakis, Matina Maniati, Christos Koros, Athina-Maria Simitsi, Maria Bozi, et al. 2020. “Circulating Brain-Enriched MicroRNAs for Detection and Discrimination of Idiopathic and Genetic Parkinson’s Disease.” *Movement Disorders* 35 (3): 457–67. https://doi.org/10.1002/mds.27928.

Reddy, Sandeep. 2022. “Explainability and Artificial Intelligence in Medicine.” *The Lancet. Digital Health* 4 (4): e214–15. https://doi.org/10.1016/S2589-7500(22)00029-2.

Richens, Jonathan G., and Albert Buchard. 2022. “Artificial Intelligence for Medical Diagnosis.” In *Artificial Intelligence in Medicine*, edited by Niklas Lidströmer and Hutan Ashrafian, 181–201. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_29.

Richter, A, C Schwager, S Hentze, W Ansorge, M W Hentze, and M Muckenthaler. 2002. “Comparison of Fluorescent Tag DNA Labeling Methods Used for Expression Analysis by DNA Microarrays.” *Biotechniques* 33 (3): 620–28, 630. https://doi.org/10.2144/02333rr05.

Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, et al. 2020. “The Future of Digital Health with Federated Learning.” *Npj Digital Medicine* 3 (September): 119. https://doi.org/10.1038/s41746-020-00323-1.

Roberts, Michael, Derek Driggs, Matthew Thorpe, Julian Gilbey, AIX-COVNET, Michael Yeung, Stephan Ursprung, et al. 2021. “Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans.” *Nature Machine Intelligence*, March. https://doi.org/10.1038/s42256-021-00307-0.

Robinson, Mark D, and Terence P Speed. 2007. “A Comparison of Affymetrix Gene Expression Arrays.” *BMC Bioinformatics* 8 (November): 449. https://doi.org/10.1186/1471-2105-8-449.

Röglin, Julia, Katharina Ziegeler, Jana Kube, Franziska König, Kay-Geert Hermann, and Steffen Ortmann. 2022. “Improving Classification Results on a Small Medical Dataset Using a GAN; An Outlook for Dealing with Rare Disease Datasets.” *Frontiers of Computer Science* 4 (August). https://doi.org/10.3389/fcomp.2022.858874.

Rood, Jennifer E, Aidan Maartens, Anna Hupalowska, Sarah A Teichmann, and Aviv Regev. 2022. “Impact of the Human Cell Atlas on Medicine.” *Nature Medicine* 28 (12): 2486–96. https://doi.org/10.1038/s41591-022-02104-7.

Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review* 65 (6): 386–408. https://doi.org/10.1037/h0042519.

Russell, Stuart, and Peter Norvig. 2016. *Artificial Intelligence. A Modern Approach*. Pearson Education Limited.

Russell, S J. 2010. "Artificial Intelligence a Modern Approach." *Artificial Intelligence a Modern Approach*.

Rüttimann, S, and D Clémençon. 1994. "Usefulness of Routine Urine Analysis in Medical Outpatients." *Journal of Medical Screening* 1 (2): 84–87. https://doi.org/10.1177/096914139400100204.

Rüttimann, Sigmund. 1992. "Usefulness of Complete Blood Counts as a Case-Finding Tool in Medical Outpatients." *Annals of Internal Medicine* 116 (1): 44. https://doi.org/10.7326/0003-4819-116-1-44.

Sabanayagam, Charumathi, Dejiang Xu, Daniel S W Ting, Simon Nusinovici, Riswana Banu, Haslina Hamzah, Cynthia Lim, et al. 2020. "A Deep Learning Algorithm to Detect Chronic Kidney Disease from Retinal Photographs in Community-Based Populations." *The Lancet. Digital Health* 2 (6): e295–302. https://doi.org/10.1016/S2589-7500(20)30063-7.

Sabry, Farida, Tamer Eltaras, Wadha Labda, Khawla Alzoubi, and Qutaibah Malluhi. 2022. "Machine Learning for Healthcare Wearable Devices: The Big Picture." *Journal of Healthcare Engineering* 2022 (April): 4653923. https://doi.org/10.1155/2022/4653923.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics* 23 (19): 2507–17. https://doi.org/10.1093/bioinformatics/btm344.

Saldanha, Oliver Lester, Hannah Sophie Muti, Heike I Grabsch, Rupert Langer, Bastian Dislich, Meike Kohlruss, Gisela Keller, et al. 2023. "Direct Prediction of Genetic Aberrations from Pathology Images in Gastric Cancer with Swarm Learning." *Gastric Cancer* 26 (2): 264–74. https://doi.org/10.1007/s10120-022-01347-0.

Saldanha, Oliver Lester, Philip Quirke, Nicholas P West, Jacqueline A James, Maurice B Loughrey, Heike I Grabsch, Manuel Salto-Tellez, et al. 2022. "Swarm Learning for Decentralized Artificial Intelligence in Cancer Histopathology." *Nature Medicine* 28 (6): 1232–39. https://doi.org/10.1038/s41591-022-01768-5.

Sammut, Stephen-John, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, et al. 2022. "Multi-Omic Machine Learning Predictor of Breast Cancer Therapy Response." *Nature* 601 (7894): 623–29. https://doi.org/10.1038/s41586-021-04278-5.

Sanchez, Robersy, and Sally A Mackenzie. 2020. "Integrative Network Analysis of Differentially Methylated and Expressed Genes for Biomarker Identification in Leukemia." *Scientific Reports* 10 (1): 2123. https://doi.org/10.1038/s41598-020-58123-2.

Sanger, F, S Nicklen, and A R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Schauer, Cameron, Michael Chieng, Michael Wang, Michelle Neave, Sarah Watson, Marius Van Rijnsoever, Russell Walmsley, and Ali Jafer. 2022. "Artificial Intelligence Improves Adenoma Detection Rate during Colonoscopy." *The New Zealand Medical Journal* 135 (1561): 22–30.

Schneider, Howard. 2022. “Applying Principles from Medicine Back to Artificial Intelligence.” In *Artificial Intelligence in Medicine*, edited by Niklas Lidströmer and Hutan Ashrafian, 21–35. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_289.

Schrödinger, Erwin. 1944. “What Is Life? The Physical Aspect of the Living Cell.”

Schulte-Schrepping, Jonas, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, et al. 2020a. “Suppressive Myeloid Cells Are a Hallmark of Severe COVID-19.” *MedRxiv*, June. https://doi.org/10.1101/2020.06.03.20119818.

———. 2020b. “Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment.” *Cell* 182 (6): 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001.

Schwalbe, Nina, and Brian Wahl. 2020. “Artificial Intelligence and the Future of Global Health.” *The Lancet* 395 (10236): 1579–86. https://doi.org/10.1016/S0140-6736(20)30226-9.

Severe Covid-19 GWAS Group, David Ellinghaus, Frauke Degenhardt, Luis Bujanda, Maria Buti, Agustín Albillos, Pietro Invernizzi, et al. 2020. “Genomewide Association Study of Severe Covid-19 with Respiratory Failure.” *The New England Journal of Medicine* 383 (16): 1522–34. https://doi.org/10.1056/NEJMoa2020283.

Shafique, Sarmad, and Samabia Tehsin. 2018. “Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks.” *Technology in Cancer Research & Treatment* 17 (January): 1533033818802789. https://doi.org/10.1177/1533033818802789.

Sharifi, Zahra, Mahmood Talkhabi, and Sara Taleahmad. 2022. “Identification of Potential MicroRNA Diagnostic Panels and Uncovering Regulatory Mechanisms in Breast Cancer Pathogenesis.” *Scientific Reports* 12 (1): 20135. https://doi.org/10.1038/s41598-022-24347-7.

Shilo, Smadar, Hagai Rossman, and Eran Segal. 2020. “Axes of a Revolution: Challenges and Promises of Big Data in Healthcare.” *Nature Medicine* 26 (1): 29–38. https://doi.org/10.1038/s41591-019-0727-5.

Shimony, Shai, Maximilian Stahl, and Richard M Stone. 2023. “Acute Myeloid Leukemia: 2023 Update on Diagnosis, Risk-Stratification, and Management.” *American Journal of Hematology* 98 (3): 502–26. https://doi.org/10.1002/ajh.26822.

Shi, Huwenbo, Gleb Kichaev, and Bogdan Pasaniuc. 2016. “Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data.” *American Journal of Human Genetics* 99 (1): 139–53. https://doi.org/10.1016/j.ajhg.2016.05.013.

Shi, Leming, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, et al. 2010. “The MicroArray Quality Control (MAQC)-II Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models.” *Nature Biotechnology* 28 (8): 827–38. https://doi.org/10.1038/nbt.1665.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. “A Survey on Image Data Augmentation for Deep Learning.” *Journal of Big Data* 6 (1): 60. https://doi.org/10.1186/s40537-019-0197-0.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529 (7587): 484–89. https://doi.org/10.1038/nature16961.

Sivesind, Torunn Elise, Taylor Runion, Megan Branda, Lisa M Schilling, and Robert P Dellavalle. 2022. "Dermatologic Research Potential of the Observational Health Data Sciences and Informatics (OHDSI) Network." *Dermatology* 238 (1): 44–52. https://doi.org/10.1159/000514536.

Smart Blood Analytics. 2023a. "MySmartBlood." 2023. https://www.smartbloodanalytics.com/en/mysmartblood.

———. 2023b. "SBAS Software." 2023. https://www.smartbloodanalytics.com/en/sbas-software.

Sovrano, Francesco, Fabio Vitali, and Monica Palmirani. 2020. "Modelling GDPR-Compliant Explanations for Trustworthy AI." In *Electronic Government and the Information Systems Perspective: 9th International Conference, EGOVIS 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings*, edited by Andrea Kő, Enrico Francesconi, Gabriele Kotsis, A Min Tjoa, and Ismail Khalil, 12394:219–33. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58957-8_16.

Speake, C, E Whalen, V H Gersuk, D Chaussabel, J M Odegard, and C J Greenbaum. 2017. "Longitudinal Monitoring of Gene Expression in Ultra-Low-Volume Blood Samples Self-Collected at Home." *Clinical and Experimental Immunology* 188 (2): 226–33. https://doi.org/10.1111/cei.12916.

Statello, Luisa, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. 2021. "Gene Regulation by Long Non-Coding RNAs and Its Biological Functions." *Nature Reviews. Molecular Cell Biology* 22 (2): 96–118. https://doi.org/10.1038/s41580-020-00315-9.

Statnikov, Alexander, Ioannis Tsamardinos, Yerbolat Dosbayev, and Constantin F Aliferis. 2005. "GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data." *International Journal of Medical Informatics* 74 (7–8): 491–503. https://doi.org/10.1016/j.ijmedinf.2005.05.002.

Sud, Amit, Bethany Torr, Michael E Jones, John Broggio, Stephen Scott, Chey Loveday, Alice Garrett, et al. 2020. "Effect of Delays in the 2-Week-Wait Cancer Referral Pathway during the COVID-19 Pandemic on Cancer Survival in the UK: A Modelling Study." *The Lancet Oncology* 21 (8): 1035–44. https://doi.org/10.1016/S1470-2045(20)30392-2.

Summerton, Nick, and Martin Cansdale. 2019. "Artificial Intelligence and Diagnosis in General Practice." *The British Journal of General Practice* 69 (684): 324–25. https://doi.org/10.3399/bjgp19X704165.

Su, Andrew I, Michael P Cooke, Keith A Ching, Yaron Hakak, John R Walker, Tim Wiltshire, Anthony P Orth, et al. 2002. "Large-Scale Analysis of the Human and Mouse Transcriptomes." *Proceedings of the National Academy of Sciences of the United States of America* 99 (7): 4465–70. https://doi.org/10.1073/pnas.012025199.

Su, Andrew I, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, et al. 2004. "A Gene Atlas of the Mouse and Human Protein-Encoding Transcriptomes." *Proceedings of the National Academy of Sciences of the United States of America* 101 (16): 6062–67. https://doi.org/10.1073/pnas.0400782101.

Sweeney, L. 2000. "Simple Demographics Often Identify People Uniquely." *Health (San Francisco)* 671 (2000): 1–34.

Thompson, Ethan G, Ying Du, Stephanus T Malherbe, Smitha Shankar, Jackie Braun, Joe Valvo, Katharina Ronacher, et al. 2017. "Host Blood RNA Signatures Predict the Outcome of Tuberculosis Treatment." *Tuberculosis* 107 (December): 48–58. https://doi.org/10.1016/j.tube.2017.08.004.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology*.

Tigchelaar, Ettje F, Alexandra Zhernakova, Jackie A M Dekens, Gerben Hermes, Agnieszka Baranska, Zlatan Mujagic, Morris A Swertz, et al. 2015. “Cohort Profile: LifeLines DEEP, a Prospective, General Population Cohort Study in the Northern Netherlands: Study Design and Baseline Characteristics.” *BMJ Open* 5 (8): e006772. https://doi.org/10.1136/bmjopen-2014-006772.

Toosi, Amirhosein, Andrea Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim. 2021. “A Brief History of AI: How to Prevent Another Winter (a Critical Review).” *ArXiv*. https://doi.org/10.48550/arxiv.2109.01517.

Topalovic, Marko, Nilakash Das, Pierre-Régis Burgel, Marc Daenen, Eric Derom, Christel Haenebalcke, Rob Janssen, et al. 2019. “Artificial Intelligence Outperforms Pulmonologists in the Interpretation of Pulmonary Function Tests.” *The European Respiratory Journal* 53 (4). https://doi.org/10.1183/13993003.01660-2018.

Topol, Eric J. 2022. “Foreword to Artificial Intelligence in Medicine.” In *Artificial Intelligence in Medicine*.

Topol, E. 2019. “The Topol Review.” *Preparing the Healthcare Workforce to Deliver the Digital Future*, 1–48.

Turing, A M. 1950. “Computing Machinery and Intelligence.” *Mind; a Quarterly Review of Psychology and Philosophy* LIX (236): 433. https://doi.org/10.1093/mind/LIX.236.433.

Ulas, Thomas, Lea Seep, Jona Schulte-Schrepping, Elena De Domenico, Simachew Mengiste, Heidi Theis, Michael Kraut, et al. 2020. “Disease Severity-Specific Neutrophil Signatures in Blood Transcriptomes Stratify COVID-19 Patients.” *MedRxiv*, July. https://doi.org/10.1101/2020.07.07.20148395.

Valk, Peter J M, Roel G W Verhaak, M Antoinette Beijen, Claudia A J Erpelinck, Sahar Barjesteh van Waalwijk van Doorn-Khosrovani, Judith M Boer, H Berna Beverloo, et al. 2004. “Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia.” *The New England Journal of Medicine* 350 (16): 1617–28. https://doi.org/10.1056/NEJMoa040465.

Vandereyken, Katy, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. “Methods and Applications for Single-Cell and Spatial Multi-Omics.” *Nature Reviews. Genetics*, March, 1–22. https://doi.org/10.1038/s41576-023-00580-2.

Van den Berge, Koen, Katharina M. Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I. Love, Rob Patro, and Mark D. Robinson. 2019. “RNA Sequencing Data: Hitchhiker’s Guide to Expression Analysis.” *Annual Review of Biomedical Data Science* 2 (1). https://doi.org/10.1146/annurev-biodatasci-072018-021255.

Veer, Laura J van ’t, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, et al. 2002. “Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer.” *Nature* 415 (6871): 530–36. https://doi.org/10.1038/415530a.

Verma, Sheetal, Peicheng Du, Damalie Nakanjako, Sabine Hermans, Jessica Briggs, Lydia Nakiyingi, Jerrold J Ellner, Yukari C Manabe, and Padmini Salgame. 2018. “‘Tuberculosis in Advanced HIV Infection Is Associated with Increased Expression of IFNγ and Its Downstream Targets’.” *BMC Infectious Diseases* 18 (1): 220. https://doi.org/10.1186/s12879-018-3127-4.

Vidić, Igor, Liv Egnell, Neil P Jerome, Jose R Teruel, Torill E Sjøbakk, Agnes Østlie, Hans E Fjøsne, Tone F Bathen, and Pål Erik Goa. 2018. “Support Vector Machine for Breast Cancer Classification Using Diffusion-Weighted MRI Histogram Features: Preliminary Study.” *Journal of Magnetic Resonance Imaging* 47 (5): 1205–16. https://doi.org/10.1002/jmri.25873.

Vogeser, Michael, and Anne K Bendt. 2023. “From Research Cohorts to the Patient - a Role for ‘Omics’ in Diagnostics and Laboratory Medicine?” *Clinical Chemistry and Laboratory Medicine*, January. https://doi.org/10.1515/cclm-2022-1147.

Wang, Pu, Xiaogang Liu, Tyler M Berzin, Jeremy R Glissen Brown, Peixi Liu, Chao Zhou, Lei Lei, et al. 2020. “Effect of a Deep-Learning Computer-Aided Detection System on Adenoma Detection during Colonoscopy (CADe-DB Trial): A Double-Blind Randomised Study.” *The Lancet. Gastroenterology & Hepatology* 5 (4): 343–51. https://doi.org/10.1016/S2468-1253(19)30411-X.

Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.” In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–71. IEEE. https://doi.org/10.1109/CVPR.2017.369.

Wang, Yeping, Zuo Wang, and Hongping Zhang. 2018. “Identification of Diagnostic Biomarker in Patients with Gestational Diabetes Mellitus Based on Transcriptome-Wide Gene Expression and Pattern Recognition.” *Journal of Cellular Biochemistry*, August. https://doi.org/10.1002/jcb.27279.

Wang, Zhilin, and Qin Hu. 2021. “Blockchain-Based Federated Learning: A Comprehensive Survey.” *ArXiv*. https://doi.org/10.48550/arxiv.2110.02182.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Wan, Jonathan C M, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. 2017. “Liquid Biopsies Come of Age: Towards Implementation of Circulating Tumour DNA.” *Nature Reviews. Cancer* 17 (4): 223–38. https://doi.org/10.1038/nrc.2017.7.

Wardi, Gabriel, Morgan Carlile, Andre Holder, Supreeth Shashikumar, Stephen R Hayden, and Shamim Nemati. 2021. “Predicting Progression to Septic Shock in the Emergency Department Using an Externally Generalizable Machine-Learning Algorithm.” *Annals of Emergency Medicine* 77 (4): 395–406. https://doi.org/10.1016/j.annemergmed.2020.11.007.

Watson, Jessica, Luke Mounce, Sarah Er Bailey, Sharon L Cooper, and Willie Hamilton. 2019. “Blood Markers for Cancer.” *BMJ (Clinical Research Ed.)* 367 (October): l5774. https://doi.org/10.1136/bmj.l5774.

Watson, J D, and F H Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

Weiner, Michael W, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, et al. 2013. “The Alzheimer’s Disease Neuroimaging Initiative: A Review of Papers Published since Its Inception.” *Alzheimer’s & Dementia* 9 (5): e111-94. https://doi.org/10.1016/j.jalz.2013.05.1769.

Wilkinson, Jack, Kellyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, et al. 2020. “Time to Reality Check the Promises of Machine Learning-Powered Precision Medicine.” *The Lancet. Digital Health* 2 (12): e677–80. https://doi.org/10.1016/S2589-7500(20)30200-4.

Wisser, Dirk, Klaus van Ackern, Ernst Knoll, Hermann Wisser, and Thomas Bertsch. 2003. “Blood Loss from Laboratory Tests.” *Clinical Chemistry* 49 (10): 1651–55. https://doi.org/10.1373/49.10.1651.

Wouters, Bas J, Bob Löwenberg, Claudia A J Erpelinck-Verschueren, Wim L J van Putten, Peter J M Valk, and Ruud Delwel. 2009. “Double CEBPA Mutations, but Not Single CEBPA Mutations, Define a Subgroup of Acute Myeloid Leukemia with a Distinctive Gene Expression Profile That Is Uniquely Associated with a Favorable Outcome.” *Blood* 113 (13): 3088–91. https://doi.org/10.1182/blood-2008-09-179895.

Wu, Nan, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, et al. 2020. “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening.” *IEEE Transactions on Medical Imaging* 39 (4): 1184–94. https://doi.org/10.1109/TMI.2019.2945514.

Wübken, Magdalena, Jana Oswald, and Antonius Schneider. 2013. “[Dealing with Diagnostic Uncertainty in General Practice].” *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen* 107 (9–10): 632–37. https://doi.org/10.1016/j.zefq.2013.10.017.

Xiao, Jing, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, and Zhibin Ye. 2019. “Comparison and Development of Machine Learning Tools in the Prediction of Chronic Kidney Disease Progression.” *Journal of Translational Medicine* 17 (1): 119. https://doi.org/10.1186/s12967-019-1860-0.

Xiong, Hui Y, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan K C Yuen, Yimin Hua, et al. 2015. “RNA Splicing. The Human Splicing Code Reveals New Insights into the Genetic Determinants of Disease.” *Science* 347 (6218): 1254806. https://doi.org/10.1126/science.1254806.

Xu, Jie, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. 2021. “Federated Learning for Healthcare Informatics.” *Journal of Healthcare Informatics Research* 5 (1): 1–19. https://doi.org/10.1007/s41666-020-00082-4.

Yadav, Samir S., and Shivajirao M. Jadhav. 2019. “Deep Convolutional Neural Network Based Medical Image Classification for Disease Diagnosis.” *Journal of Big Data* 6 (1): 113. https://doi.org/10.1186/s40537-019-0276-2.

Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, et al. 2022. “A Saturated Map of Common Genetic Variants Associated with Human Height.” *Nature* 610 (7933): 704–12. https://doi.org/10.1038/s41586-022-05275-y.

Yu, Wei, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J Khoury. 2010. “Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes.” *BMC Medical Informatics and Decision Making* 10 (March): 16. https://doi.org/10.1186/1472-6947-10-16.

Zak, Daniel E, Adam Penn-Nicholson, Thomas J Scriba, Ethan Thompson, Sara Suliman, Lynn M Amon, Hassan Mahomed, et al. 2016. “A Blood RNA Signature for Tuberculosis Disease Risk: A Prospective Cohort Study.” *The Lancet* 387 (10035): 2312–22. https://doi.org/10.1016/S0140-6736(15)01316-1.

Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. “Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study.” *PLoS Medicine* 15 (11): e1002683. https://doi.org/10.1371/journal.pmed.1002683.

Zemmour, Christophe, François Bertucci, Pascal Finetti, Bernard Chetrit, Daniel Birnbaum, Thomas Filleron, and Jean-Marie Boher. 2015. “Prediction of Early Breast Cancer Metastasis from DNA Microarray Data Using High-Dimensional Cox Regression Models.” *Cancer Informatics* 14 (Suppl 2): 129–38. https://doi.org/10.4137/CIN.S17284.

Zhang, Zhen, Lin Zhao, Xijin Wei, Qiang Guo, Xiaoxiao Zhu, Ran Wei, Xunqiang Yin, Yunhong Zhang, Bin Wang, and Xia Li. 2018. “Integrated Bioinformatic Analysis of Microarray Data Reveals Shared Gene Signature between MDS and AML.” *Oncology Letters* 16 (4): 5147–59. https://doi.org/10.3892/ol.2018.9237.

Zhan, Fenghuang, Bart Barlogie, Varant Arzoumanian, Yongsheng Huang, David R Williams, Klaus Hollmig, Mauricio Pineda-Roman, et al. 2007. “Gene-Expression Signature of Benign Monoclonal Gammopathy Evident in Multiple Myeloma Is Linked to Good Prognosis.” *Blood* 109 (4): 1692–1700. https://doi.org/10.1182/blood-2006-07-037077.

Zhao, S, Y Zhang, R Gamini, B Zhang, and D von Schack. 2018. “Evaluation of Two Main RNA-Seq Approaches for Gene Quantification in Clinical RNA Sequencing: PolyA+ Selection versus RRNA Depletion.” *Scientific Reports* 8 (1): 4781. https://doi.org/10.1038/s41598-018-23226-4.

Zheng, Jie, and Ke Wang. 2019. “Emerging Deep Learning Methods for Single-Cell RNA-Seq Data Analysis.” *Quantitative Biology* 7 (4): 247–54. https://doi.org/10.1007/s40484-019-0189-2.

Zhu, Xudong, Fan Zhang, and Hui Li. 2022. “Swarm Deep Reinforcement Learning for Robotic Manipulation.” *Procedia Computer Science* 198: 472–79. https://doi.org/10.1016/j.procs.2021.12.272.

Zijlema, Wilma L, Nynke Smidt, Bart Klijs, David W Morley, John Gulliver, Kees de Hoogh, Salome Scholtens, Judith G M Rosmalen, and Ronald P Stolk. 2016. “The LifeLines Cohort Study: A Resource Providing New Opportunities for Environmental Epidemiology.” *Archives of Public Health = Archives Belges de Sante Publique* 74 (August): 32. https://doi.org/10.1186/s13690-016-0144-x.

Zimmerer, David, Peter M Full, Fabian Isensee, Paul Jager, Tim Adler, Jens Petersen, Gregor Kohler, et al. 2022. “MOOD 2020: A Public Benchmark for Out-of-Distribution Detection and Localization on Medical Images.” *IEEE Transactions on Medical Imaging* 41 (10): 2728–38. https://doi.org/10.1109/TMI.2022.3170077.

Zoabi, Yazeed, Shira Deri-Rozov, and Noam Shomron. 2021. “Machine Learning-Based Prediction of COVID-19 Diagnosis Based on Symptoms.” *Npj Digital Medicine* 4 (1): 3. https://doi.org/10.1038/s41746-020-00372-6.

Zwiener, Isabella, Barbara Frisch, and Harald Binder. 2014. “Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures.” *Plos One* 9 (1): e85150. https://doi.org/10.1371/journal.pone.0085150.

7 Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Dr. Joachim L. Schultze for his trust, unwavering support, guidance, and encouragement from when I started by lab rotation in 2015 until the finish line of my PhD journey.

I am also very grateful to the members of my dissertation committee, Prof. Dr. Jan Hasenauer, Prof. Dr. Waldemar Kolanus and Prof. Dr. Christian Bauckhage for taking the time to carefully evaluate my thesis and giving their feedback to my work.

I am deeply grateful to all present and former colleagues and friends in the Systems Medicine Department, who have provided me with a supportive and stimulating intellectual community throughout the whole journey from starting a lab rotation in the group to finally finishing this PhD. Lorenzo, Jonas, Caterina, Tal, Kevin, Patrick, Matthias, Thomas, Anna, Marc, Lisa, Charlotte, Pawel, Kathrin, Jil, Kristian, Elke, Heidi, Michel, and everyone I have missed - your friendship, encouragement, and intellectual engagement have always been an invaluable source of inspiration and motivation. I am grateful for having had the opportunity to work in this fantastic group and for having such wonderful people in my life to share this journey with.

Finally, I would like to express my deepest gratitude to my family, who have provided me with support, love, and encouragement throughout the last years. Especially I would like to thank Jannis who has been my biggest supporter in this process and clearly I would not have been able to finalize the papers nor this thesis without you. I would also like to thank my mother for believing in my capabilities and supporting me in all my pursuits and my father for his patience and trust that all will work out fine. Last but certainly not least, I would like to sincerely thank my children Jona, Valentin and Simon, for their sacrifices and their constant support in what I am doing, even though it must seem abstract (“she is looking at blood, but not with a microscope”).

Thank you for making this journey all the more meaningful!

8 Appendix

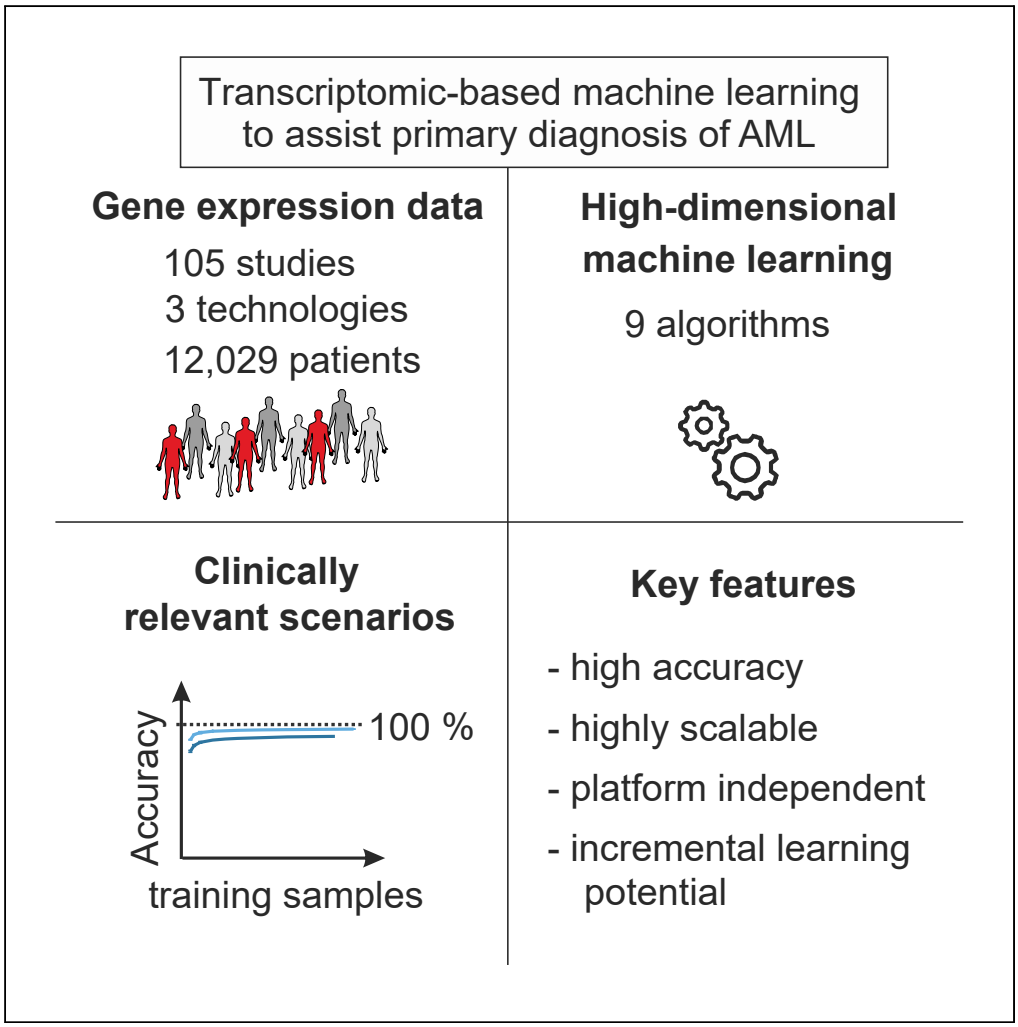
- **Publication I:**
  Warnat-Herresthal S, Perrakis K, Taschler B, Becker M, Baßler K, Beyer M, Günther P, Schulte-Schrepping J, Seep L, Klee K, Ulas T, Haferlach T, Mukherjee S, Schultze JL. Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. iScience. 2020 Jan 24;23(1):100780. doi: 10.1016/j.isci.2019.100780.
- **Publication II:**
  Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, Sarveswara R, Händler K, Pickkers P, Aziz NA, Ktena S, Tran F, Bitzer M, Ossowski S, Casadei N, Herr C, Petersheim D, Behrends U, Kern F, Fehlmann T, Schommers P, Lehmann C, Augustin M, Rybniker J, Altmüller J, Mishra N, Bernardes JP, Krämer B, Bonaguro L, Schulte-Schrepping J, De Domenico E, Siever C, Kraut M, Desai M, Monnet B, Saridaki M, Siegel CM, Drews A, Nuesch-Germano M, Theis H, Heyckendorf J, Schreiber S, Kim-Hellmuth S; COVID-19 Aachen Study (COVAS); Nattermann J, Skowasch D, Kurth I, Keller A, Bals R, Nürnberg P, Rieß O, Rosenstiel P, Netea MG, Theis F, Mukherjee S, Backes M, Aschenbrenner AC, Ulas T; Deutsche COVID-19 Omics Initiative (DeCOI); Breteler MMB, Giamarellos-Bourboulis EJ, Kox M, Becker M, Cheran S, Woodacre MS, Goh EL, Schultze JL. Swarm Learning for decentralized and confidential clinical machine learning. Nature. 2021 Jun;594(7862):265-270. doi: 10.1038/s41586-021-03583-3.

iScience

CellPress

Article

Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics



Stefanie Warnat-Herresthal, Konstantinos Perrakis, Bernd Taschler, ..., Torsten Haferlach, Sach Mukherjee, Joachim L. Schultze

sach.mukherjee@dzne.de (S.M.)
j.schultze@uni-bonn.de (J.L.S.)

HIGHLIGHTS

Study presents one of the largest transcriptomics datasets to date for AML prediction

Effective classifiers can be obtained by high-dimensional machine learning

Accuracy increases with dataset size

Includes challenging scenarios such as cross-study and cross-technology

DATA AND CODE AVAILABILITY

GSE122517
GSE122505
GSE122511
GSE122515

Warnat-Herresthal et al., iScience *23*, 100780
January 24, 2020
https://doi.org/10.1016/j.isci.2019.100780

iScience

CellPress

Article

Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics

Stefanie Warnat-Herresthal,[1,6] Konstantinos Perrakis,[2,6] Bernd Taschler,[2] Matthias Becker,[4] Kevin Baßler,[1] Marc Beyer,[3,4] Patrick Günther,[1] Jonas Schulte-Schrepping,[1] Lea Seep,[1] Kathrin Klee,[1] Thomas Ulas,[1] Torsten Haferlach,[5] Sach Mukherjee,[2,7,*] and Joachim L. Schultze[1,4,7,8,*]

SUMMARY

Acute myeloid leukemia (AML) is a severe, mostly fatal hematopoietic malignancy. We were interested in whether transcriptomic-based machine learning could predict AML status without requiring expert input. Using 12,029 samples from 105 different studies, we present a large-scale study of machine learning-based prediction of AML in which we address key questions relating to the combination of machine learning and transcriptomics and their practical use. We find data-driven, high-dimensional approaches—in which multivariate signatures are learned directly from genome-wide data with no prior knowledge—to be accurate and robust. Importantly, these approaches are highly scalable with low marginal cost, essentially matching human expert annotation in a near-automated workflow. Our results support the notion that transcriptomics combined with machine learning could be used as part of an integrated -omics approach wherein risk prediction, differential diagnosis, and subclassification of AML are achieved by genomics while diagnosis could be assisted by transcriptomic-based machine learning.

INTRODUCTION

Recommendations for the diagnosis and management of malignant diseases are organized by international expert panels. For example, the first edition of the European LeukemiaNet (ELN) recommendations for the diagnosis and management of acute myeloid leukemia (AML) in adults was published in 2010 (Döhner et al., 2010) and recently revised in 2017 (Döhner et al., 2017). Based on recent DNA sequencing results, such as those derived from The Cancer Genome Atlas, AML can be subdivided into multiple subclasses (Arber et al., 2016; Ding et al., 2012; Ley et al., 2008, 2010; Loriaux et al., 2008; Papaemmanuil et al., 2016; The Cancer Genome Atlas Research Network (TCGA) et al., 2013; Welch et al., 2012; Yan et al., 2011). Leukemias are characterized by strong transcriptomic signals, as seen in a pioneering study almost two decades ago by Golub et al. (Golub et al., 1999) and a rich body of subsequent work (Debernardi et al., 2003; Kohlmann et al., 2003; Ross et al., 2004; Schoch et al., 2002; Virtaneva et al., 2001). These findings led to the suggestion that gene expression profiling (GEP) could be utilized to define leukemia subtypes and derive useful predictive gene signatures (Andersson et al., 2007; Bullinger et al., 2004). Nevertheless, according to the ELN recommendations primary diagnosis still relies on classical approaches including assessment of morphology, immunophenotyping, cytochemistry, and cytogenetics (Döhner et al., 2017). Although undoubtedly effective in detecting disease, these existing diagnostic approaches rely on large investments in human expertise (training and employment of specialists) and physical infrastructure, whose costs scale with the number of samples. This has implications for accessibility (e.g., in rural areas or outside developed regions) and on cost and logistical grounds alone limits the scope to consider alternatives to the overall decision pipeline. In contrast to classical diagnostic pipelines that are centered on interpretation of results by human experts, artificial intelligence- (AI) and machine learning- (ML) based approaches have the potential for low marginal cost (i.e., cost per additional sample once the system is trained) (Esteva et al., 2017), and this key aspect of AI and ML is widely appreciated in the economics literature (see, e.g., Brynjolfsson and McAfee, 2014).

The potential of GEP for leukemia diagnosis has been recognized. A decade after the pioneering work of Golub et al., the International Microarray Innovations in Leukemia Study Group proposed GEP by microarray analysis to be a robust technology for the diagnosis of hematologic malignancies with high accuracy

[1]LIMES-Institute, Department for Genomics and Immunoregulation, University of Bonn, Carl-Troll-Str. 31, 53115 Bonn, Germany

[2]Statistics and Machine Learning, German Center for Neurodegenerative Diseases, Venusberg-Campus 1, Building 99, 53127 Bonn, Germany

[3]Molecular Immunology in Neurodegeneration, German Center for Neurodegenerative Diseases, Venusberg-Campus 1, Building 99, 53127 Bonn, Germany

[4]PRECISE Platform for Single Cell Genomics and Epigenomics, German Center for Neurodegenerative Diseases and the University of Bonn, Venusberg-Campus 1, Building 99, 53127 Bonn, Germany

[5]MLL, Münchner Leukämielabor GmbH, Max-Lebsche-Platz 31, 81377 München, Germany

[6]These authors contributed equally

[7]These authors contributed equally

[8]Lead Contact

*Correspondence: sach.mukherjee@dzne.de (S.M.), j.schultze@uni-bonn.de (J.L.S.)

https://doi.org/10.1016/j.isci.2019.100780

(Haferlach et al., 2010). The utility of GEP by RNA sequencing (RNA-seq) has been also demonstrated for other tumor entities, for example, breast cancer (Ciriello et al., 2015; Kristensen et al., 2012; Parker et al., 2009), bladder cancer, or lung cancer (Hoadley et al., 2014; Robertson et al., 2017). Furthermore, in AML research large RNA-seq datasets have been described in the meantime (Garzon et al., 2014; Lavallee et al., 2016; Lavallée et al., 2015; Macrae et al., 2013; Pabst et al., 2016).

In parallel, a series of advances in ML, AI, and computational statistics have transformed our understanding of prediction using high-dimensional data. A variety of approaches are now an established part of the toolkit, and for some models (including sparse linear and generalized linear models), there is a rich mathematical theory concerning their performance in the high-dimensional setting (Bühlmann and van de Geer, 2011). In a nutshell, the body of empirical and theoretical research has shown that learning predictive models over large numbers of variables is often feasible and remarkably effective. In applied ML, there has been a deepening understanding of practical issues, e.g., relating to the transferability of predictions across contexts (Quiñonero-Candela et al., 2009), that is very relevant to the clinical setting.

Based on these developments in the data sciences and the increasing availability of GEP data derived from peripheral blood including AML, we sought to develop near-automated approaches in which ML tools automatically learn suitable patterns directly from the global transcriptomic data without pre-selection of genes. To this end, we built the probably largest reference blood GEP dataset comprising 105 individual studies with, in total, more than 12,000 patient samples. We applied high-dimensional ML approaches to build genome-wide predictors in an unbiased, entirely data-driven manner and tested predictive accuracy in held-out data. We emphasize that our goal was not to outperform classical diagnostic methods, but to ask whether we could match human annotation in a near-automated and scalable manner. This aim is common to a number of recent efforts to use ML and AI advances in the diagnostic setting (see, e.g., Esteva et al., 2017) wherein human-derived labels are used to guide learning. We did not address the question of subclassification of leukemic disease, where the mutation status of the leukemic cells is currently the dominant approach (Arber et al., 2016; Heath et al., 2017; Papaemmanuil et al., 2016; The Cancer Genome Atlas Research Network (TCGA) et al., 2013), but rather focused on primary diagnosis, which continues to rely mostly on classical approaches (morphology, immunophenotyping, cytochemistry). We carried out extensive tests designed to address specific concerns relevant to practical use, including the case of transferring predictive models between entirely disjoint studies (that could be subject to batch effects or other unwanted variation) and even between transcriptomic platforms. Our results show that combining ML and blood transcriptomics can yield highly effective and robust classifiers. This supports the notion that transcriptomic-based ML could be used to assist AML diagnostics, particularly in settings wherein hematological expertise is not sufficiently available and/or costly.

RESULTS

Establishment of a Unique GEP Dataset for Classifier Development

We hypothesized that the determination and comprehensive evaluation of GEP- and ML-based AML classifiers requires large datasets, should include samples from many sources to mimic the situation in real-world deployment, and should include several technical platforms to better understand their influence on classifier performance. To achieve these goals, we wanted to include the largest number of peripheral blood mononuclear cells (PBMC) or bone marrow samples possible and therefore systematically searched the National Center for Biotechnology Gene Expression Omnibus (GEO; Edgar, 2002) database for PBMC and bone marrow studies (Figure 1A). We identified 153,922 datasets, of which 111,632 contained human samples. To include only whole sample series and to avoid duplicate samples, we filtered for GEO series (GSE) and excluded the so-called super series, which resulted in 2,715 studies. We then focused the analysis on studies with samples analyzed on one of three platforms including the HG-U133A microarray, the HG-U133 2.0 microarray, and Illumina RNA-seq. Next, duplicated samples and studies working with pre-filtered cell subsets were excluded. This study search strategy resulted in 105 studies with a total of 12,029 samples (Figure 1) including 2,500 samples assessed by HG-U133A microarray (Dataset 1), 8,348 samples by HG-U133 2.0 microarray (Dataset 2), and 1,181 samples by RNA-seq (Dataset 3). In total, the dataset contained 4,145 AML samples of diverse disease subtypes and 7,884 other samples derived from healthy controls (n = 904), patients with acute lymphocytic leukemia (ALL, n = 3,466), chronic myeloid leukemia (CML, n = 162), chronic lymphocytic leukemia (CLL, n = 770), myelodysplastic syndrome (MDS, n = 267), and other non-leukemic diseases (n = 2,312) (Figures 1B, S1, and S2). Unless otherwise noted, all samples derived from patients with AML are referred to as cases and non-AML samples as controls. We

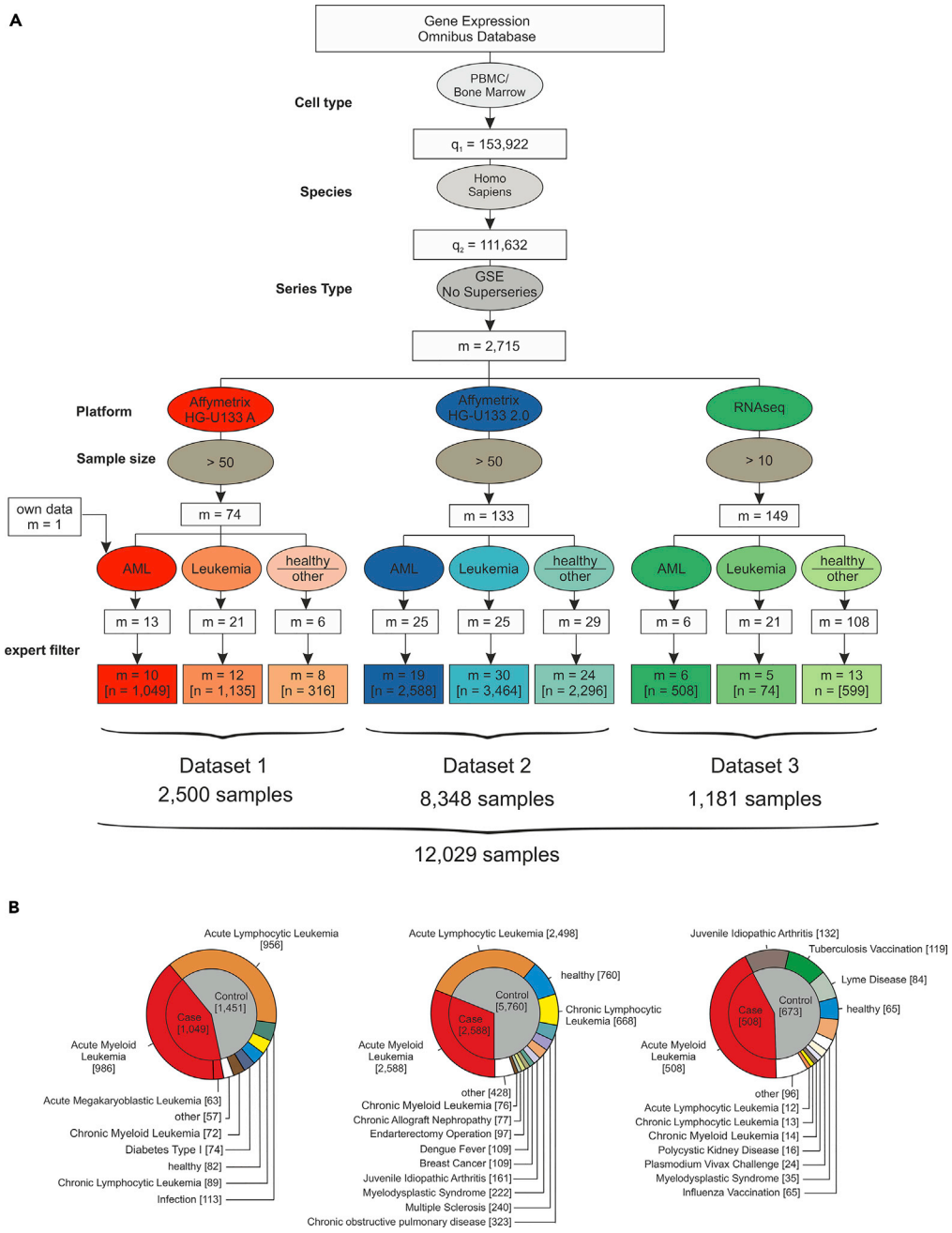


Figure 1. Establishing Datasets for the Largest AML Meta-study to Date

(A) Flowchart for the inclusion of studies. The gene expression omnibus (GEO) database was systematically searched for GEO Series of human PBMC and bone marrow samples processed with microarray platforms (Affymetrix HG-U133A and HG-U133 2.0) or next-generation RNA sequencing (RNA-seq) data. These data were filtered for inclusion of AML samples, samples of other leukemia, and healthy samples or other diseases. After manual revision and exclusion of duplicates and experiments using sorted cell populations ("expert filter"), the data were combined and normalized independently for each dataset.

(B) Detailed overview of the three datasets established in this study after filtering as given in (A).

See also Figure S1.

additionally considered a differential diagnosis-like setting, in which case the controls comprised non-AML leukemias. According to the three platform types, the whole sample cohort was divided into three datasets referred to as datasets 1, 2, and 3 (Table S1, Figure S1).

Effective AML Classification Using High-Dimensional Models

Here, we sought to assess classification of AML versus non-AML. Microarray data were RMA normalized using the R package affy (Gautier et al., 2004), whereas RNA-seq data were normalized as implemented in the R package DESeq2 (Love et al., 2014). For further analysis and better comparison between the different datasets, we trimmed the data to 12,708 genes, which were annotated within all datasets. No filtering of low expressed genes was performed (Figure 2A). The size of the test set was 20% of the total sample size, and random sampling of training and test sets was repeated 100 times. As main performance metrics, we considered (held-out) accuracy, sensitivity, and specificity. Classification was performed using l_1-regularized logistic regression (the lasso; see also later).

First, we included all non-AML samples, consisting of healthy controls and non-leukemic diseases, among the controls (Figures 2B, 2D, and 2F, light blue lines, Table S2). The goal was to classify unseen samples as AML or control. To understand how much data is needed in this setting, we plotted learning curves showing the test set accuracy as a function of training sample size n_{train}. For each gene expression platform, this was done by randomly subsampling n_{train} samples and testing on held-out test data with fixed sample size n_{test} (as shown). We see that prediction in this setting is already highly effective with a small number of training samples, although accuracy still increases with increasing n_{train} (note that the total number of samples and hence range of n_{train} differs by platform).

In many clinical settings, the control group does not contain healthy controls, but rather related diseases. To test effectiveness in a differential diagnosis setting, we repeated the experiments but with controls sampled only from other leukemic diseases, such as ALL, CLL, CML, and MDS (Figures 2C, 2E, S3–S5, and S13). We observed similar prediction results, which indicated that prediction accuracy is not only due to large differences between AML and non-leukemic conditions.

In additional experiments we considered performance of nine different classification methods (Figures S3–S5, Table S4). We could predict AML with good accuracy with all tested classification algorithms on microarray platforms (Figures S3 and S4). For RNA-seq data, the lasso, *k* nearest neighbors, linear support vector machines, linear discriminant analysis, and random forests were able to predict with high sensitivity and specificity (Figure S5, for details on used packages see Transparent Methods). Lasso-type methods have several advantages, including extensive theoretical support and interpretability, so we focused on these as our main predictive tool. Deep neural networks provided similar prediction performance to the lasso (Figure 2G) on dataset 2. We preferred the latter in this setting due to interpretability, because the lasso provides explicit variable selection, facilitating model interpretation.

Evaluation of Positive Predictive Value under Various Prevalence Scenarios

For diagnostic utility, the positive predictive value (PPV; the probability of disease given a positive test result) is an important quantity. The PPV depends not only on sensitivity and specificity but also on prevalence, as it is harder to achieve a high PPV for a condition that is rare in the population of interest. This has implications for any change to the effective threshold at which a potential case enters the diagnostic pipeline. As this threshold is relaxed, the prevalence (in the tested population) decreases, which in turn reduces the PPV. Thus, although we found high accuracy, sensitivity, and specificity already at moderate n_{train}, depending on the use case, this could still imply that large training sample sizes would be useful to reach acceptable PPVs. For example, the predictive gains in increasing n_{train} from the lowest to the highest values indicated in Figure 2C, which is for the dataset with largest total sample size, correspond to a *doubling* of PPV from ~20% to ~40% at an assumed prevalence of 1% (Figure 3). This illustrates the fact that although after a certain point increasing n_{train} tends to increase accuracy only slowly, the gains, even if small in absolute terms, can be highly relevant with respect to PPV in low-prevalence settings.

Assessing the Effect of Cross-Study Variation on Predictive Performance

Microarray data and data generated by high-throughput sequencing are both known to be susceptible to batch effects (Leek et al., 2010). More generally, diverse study-specific effects and sources of study-to-study variation can pose problems in the context of predictive tests for clinical applications. Predictors that perform well within one study may perform worse when applied to data from new studies (Hornung et al., 2017) with implications for practical generalizability.

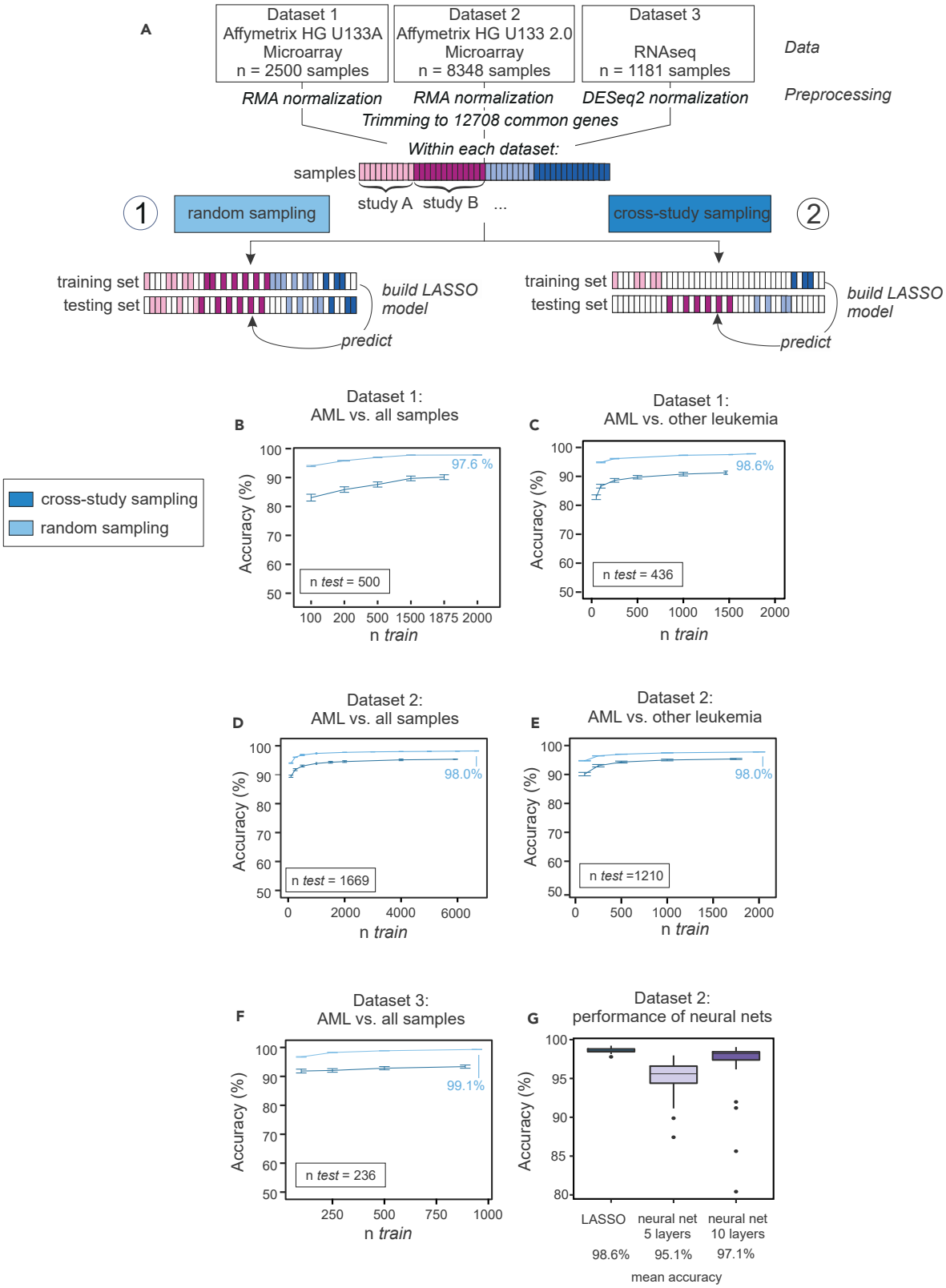

A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict
cross-study sampling
random sampling
B
Dataset 1:
AML vs. all samples
Accuracy (%)
97.6 %
n test = 500
n train
C
Dataset 1:
AML vs. other leukemia
Accuracy (%)
98.6%
n test = 436
n train
D
Dataset 2:
AML vs. all samples
Accuracy (%)
98.0%
n test = 1669
n train
E
Dataset 2:
AML vs. other leukemia
Accuracy (%)
98.0%
n test =1210
n train
F
Dataset 3:
AML vs. all samples
Accuracy (%)
99.1%
n test = 236
n train
G
Dataset 2:
performance of neural nets
Accuracy (%)
LASSO
neural net 5 layers
neural net 10 layers
98.6%
95.1%
97.1%
mean accuracy

Figure 2. Prediction of AML in Random and Cross-study Sampling Scenarios

(A) Schema illustrating the approach to predict AML in random and cross-study sampling scenarios.

(B–D) AML classification accuracies based on the lasso model of AML versus all other samples and for both sampling strategies are shown for dataset 1 (B), dataset 2 (C), and dataset 3 (D).

(E and F) Classification accuracies for the differential diagnosis case (AML versus other leukemic samples, namely, AML, ALL, CML, CLL, and MDS) for both sampling strategies are shown for dataset 1 (E) and dataset 2 (F). Mean accuracies of the lasso models are shown as a function of the training sample size n_{train}. Results are over 100 random training and test sets, with error bars indicating the standard deviation.

(G) Comparison of the performance of the LASSO models introduced in panels A to F with a neural network approach using either 5 or 10 layers. Error bars indicate the standard deviation.

See also Figures S3–S8 and S13, and Tables S2 and S4.

The aforementioned results spanned data from multiple heterogeneous studies. Provided training and test data are sampled in the same way, such heterogeneity does not necessarily pose problems for classification, as evidenced earlier. However, if the training and test data are from entirely different sites/studies (rather than randomly sampled from a shared pool), then the impact of batch/study effects may be more serious. We took advantage of the large number of studies in our dataset to sample training and testing sets in such a way that they were mutually disjoint with respect to studies. That is, any individual study from which any sample was included into the training dataset was entirely absent from the test set, and *vice versa,* and we use the term *cross-study* to refer to this strictly disjoint case. Results are shown in Figures 2B, 2D, and 2F (dark blue lines). As expected, performance was worse in the cross-study setting than under entirely random sampling (light blue lines). However, in the dataset with the largest sample size (dataset 2, platform HG-U133 2.0; Figure 2D) we see that the performance in the cross-study case gradually catches up to the random sampling case with only a small gap at the largest n_{train}. The other two datasets have smaller total sample sizes, so they never reach comparable training sample sizes. Note that we did not carry out any batch effect removal using tools such as combat (Johnson et al., 2007), SVA (Leek et al., 2012), or RUV (Jacob et al., 2016), and in that sense our results are conservative. Despite the availability of these and other tools for batch effect correction, it is difficult to be fully assured of the removal of unwanted variation in practice. Our intention here was not to remove between-study variation but rather to (conservatively) quantify its effects on accuracy.

Owing to the large number of studies included in our analysis, we were able to carry out an entirely disjoint cross-study analysis also for the differential diagnosis case. These results are shown in Figures 2C and 2E (dark blue lines; cross-study sampling for differential diagnosis was not possible using dataset 3 due to lack of samples, see Figure S1) and are broadly similar, also across different classification algorithms (Figures S6–S8).

However, even in this strict cross-study sampling scenario, where samples from studies of the training and testing sets are entirely disjoint, the predictor matrices are still normalized together, meaning that the prediction rule still depends to some extent on features (not labels) in the test set. To address this issue, we performed addon RMA normalization (Hornung et al., 2017) as implemented in the R package bapred (Hornung et al., 2016). We split dataset 1 in training and testing data in a strict cross-study setting as in Figure 2A, performed RMA normalization on the training data, and then performed addon normalization of the test data onto the training data, meaning that the normalization of the training data does not in any sense depend on the testing data (Figure S9A). Accuracy, sensitivity, and specificity of this setting compare well to the "classical" cross-study setting described earlier (Figures 2 and S6A).

Classification Accuracy and AML Subtypes

Next, we sought to understand whether the accuracy of the classifiers depended on specific AML subtypes. As only a limited number of samples in our data were already annotated according to the new World Health Organization (WHO) classification, we utilized the French-American-British (FAB) classification of AML. The FAB classification was available for a total of 616 samples of dataset 1 and 1,269 samples in dataset 2. We utilized results from train/test splits of datasets 1 and 2 to quantify accuracy for each individual sample (Figures 4A and 4D). No particular AML subtype dominated classification accuracy in either dataset 1 or 2 (Figures 4C and 4F). Prediction accuracy was also consistent when broken down by non-AML disease category (Figures 4B and 4E). In dataset 1, 8 MDS samples and 10 samples from patients with Down syndrome transient myeloproliferative disorder were misclassified. However, both are diseases closely related to AML and represented by a very limited sample size in dataset 1. For dataset 2, correct classification of MDS appeared to depend on the individual sample, potentially reflecting disease heterogeneity.

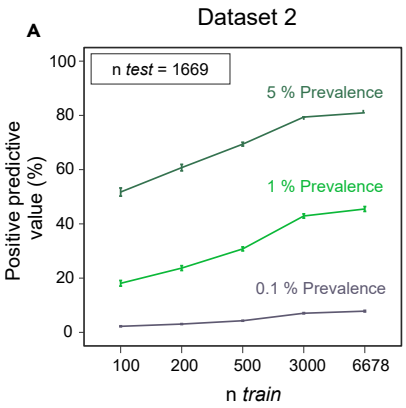


Figure 3. Positive Predictive Value

Positive predictive value as a function of n_{train} corresponding to the setting as in Figure 2C and assumed prevalence of 0.1%, 1%, or 5% is shown (see text). Error bars depict the standard deviation.

While our main focus is on diagnosis, we asked whether the transcriptomic data could contribute to classifying AML subtypes. To exemplify this aspect, we focused on AML subtype M3, also named *acute promyelocytic leukemia*, as this is the only genetically defined subtype of the FAB classification that is also part of the WHO classification. Using dataset 2, we used a train/test approach, drawing subsets of dataset 2 with approximately the same class balance as in main results (here, one-third AML-M3 cases in every subset) (Figure 4G). M3 was distinguished from non-M3 AML with high accuracy, sensitivity, and specificity (Figure 4H). Although the data here do not allow rigorous testing of transcriptomics combined with genomics in an integrated fashion for subtype classification, and we would not recommend at this stage the use of a purely transcriptomic classifier for subtyping, these initial results suggest that it may be useful to further study the potential value of testing scalable ML- and GEP-based methodology in the area of subclassification as well.

Translation of Classifiers across Technical Platforms

Over the long term, clinical pipelines must cope with changes in technological platforms. It is therefore relevant to understand to what extent predictors can generalize not only between studies but also between different platforms. In other words, is it possible to take a model learned on data from platform *A* and deploy it using unseen data from platform *B*? To address this question, we constructed AML versus non-AML training and test sets in a *cross-platform* manner, i.e., training on one platform and testing on another (Figure 5A). That is, a model was learned using independently normalized data from one platform and then this model, used "as is," with no further fine-tuning, was used to make predictions using expression data from a different platform. We see that classification accuracy varies greatly. Classifiers that were trained on HG-U133 A (dataset 1) work well when tested using data generated with the more advanced microarray HG-U133 2.0 (dataset 2) (Figures 5B and S10) and models trained on HG-U133 2.0 data can predict well using RNA-seq data (dataset 3) (Figures 5D and S11). However, models trained naively on HG-U133 A data cannot predict using RNA-seq data (Figures 5F, S12, and S13, Table S4).

To explore the utility of simple transformations in this context, we then performed a rank transformation to normality on all datasets (see Transparent Methods). This is among the simplest and best known data transformations, has previously been shown to increase the performance of prognostic gene expression signatures, and can even outperform more complex variance-stabilizing approaches (Zwiener et al., 2014). With this approach, we reached very good overall performance across all platforms under study (Figures 5C, 5E, and 5G). This is particularly interesting for the prediction of dataset 3, which fails when the model is trained on the untransformed dataset 1 (Figures 5F, 5G, S11, and S12) and performs *worse* (on dataset 3) as n_{train} increases. This is because as n_{train} increases, the models learn a pattern that is increasingly fine-tuned to the data type in the training set. However, because the test set is from a *different* platform, test performance suffers. This is most likely not classical "overfitting," because as shown in previous figures test error is well-behaved *within* dataset 1, but rather an example of a transfer learning/distribution-shift type problem, which in this case is solved simply by rank transformation. Note that the transformation is simply applied to each dataset independently and could be easily deployed in any practical use case without any need for prior input into, e.g., cross-platform designs such as inclusion of control samples.

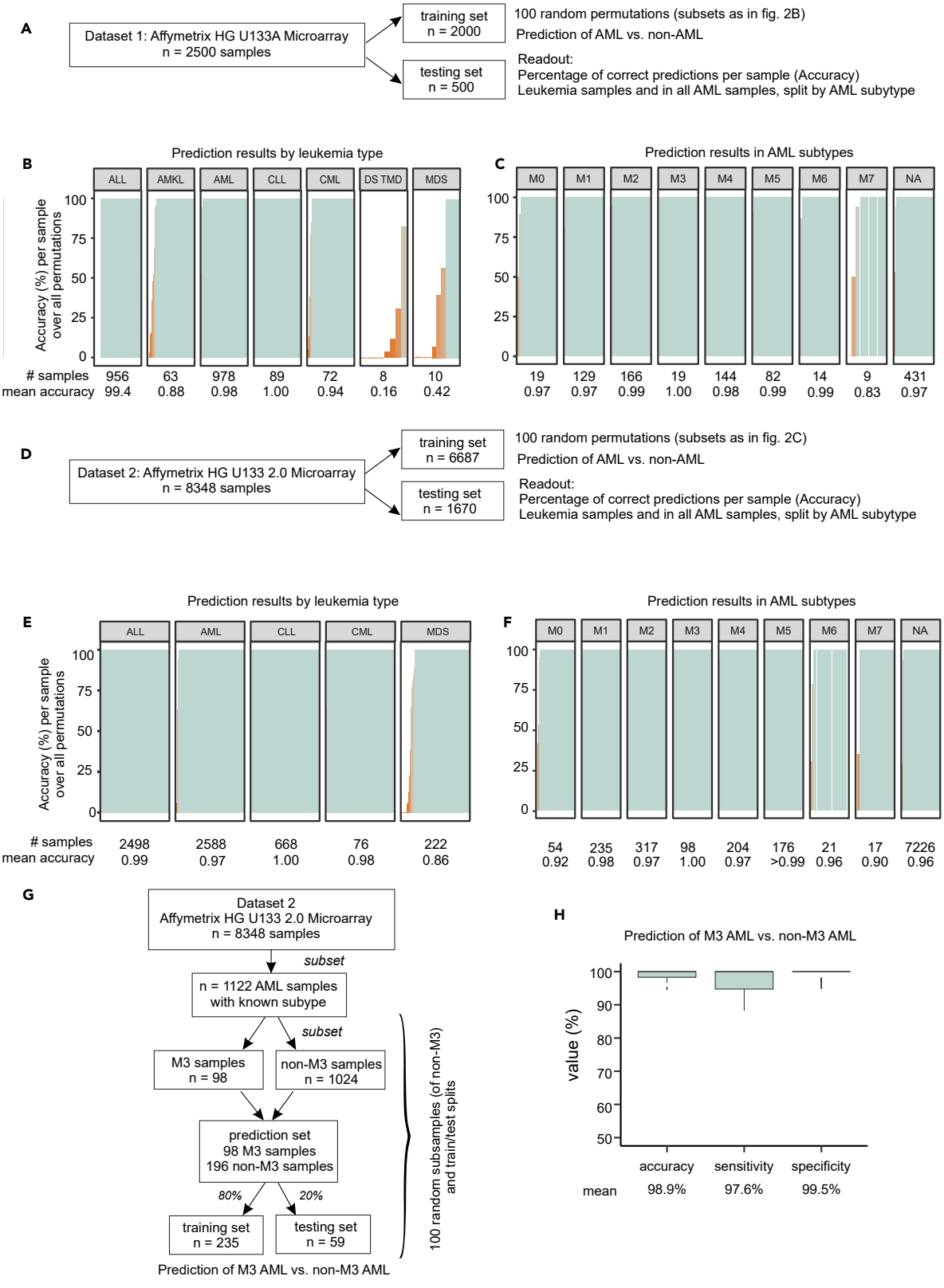

A
Dataset 1: Affymetrix HG U133A Microarray
n = 2500 samples
training set
n = 2000
testing set
n = 500
100 random permutations (subsets as in fig. 2B)
Prediction of AML vs. non-AML
Readout:
Percentage of correct predictions per sample (Accuracy)
Leukemia samples and in all AML samples, split by AML subytype
B
Prediction results by leukemia type
ALL AMKL AML CLL CML DS TMD MDS
Accuracy (%) per sample over all permutations
samples 956 63 978 89 72 8 10
mean accuracy 99.4 0.88 0.98 1.00 0.94 0.16 0.42
C
Prediction results in AML subtypes
M0 M1 M2 M3 M4 M5 M6 M7 NA
19 129 166 19 144 82 14 9 431
0.97 0.97 0.99 1.00 0.98 0.99 0.99 0.83 0.97
D
Dataset 2: Affymetrix HG U133 2.0 Microarray
n = 8348 samples
training set
n = 6687
testing set
n = 1670
100 random permutations (subsets as in fig. 2C)
Prediction of AML vs. non-AML
Readout:
Percentage of correct predictions per sample (Accuracy)
Leukemia samples and in all AML samples, split by AML subytype
E
Prediction results by leukemia type
ALL AML CLL CML MDS
Accuracy (%) per sample over all permutations
samples 2498 2588 668 76 222
mean accuracy 0.99 0.97 1.00 0.98 0.86
F
Prediction results in AML subtypes
M0 M1 M2 M3 M4 M5 M6 M7 NA
54 235 317 98 204 176 21 17 7226
0.92 0.98 0.97 1.00 0.97 >0.99 0.96 0.90 0.96
G
Dataset 2
Affymetrix HG U133 2.0 Microarray
n = 8348 samples
subset
n = 1122 AML samples
with known subype
subset
M3 samples
n = 98
non-M3 samples
n = 1024
prediction set
98 M3 samples
196 non-M3 samples
80%
20%
training set
n = 235
testing set
n = 59
100 random subsamples (of non-M3) and train/test splits
Prediction of M3 AML vs. non-M3 AML
H
Prediction of M3 AML vs. non-M3 AML
value (%)
accuracy sensitivity specificity
mean 98.9% 97.6% 99.5%

Figure 4. Accuracy of AML Classification in Different Leukemia Types and AML Subclasses

(A) Schema for determining accuracy for leukemia types and AML subclasses in dataset 1.

(B–D) Normalized dataset 1 was randomly split into training and test sets 100 times (same permutations as in Figure 2B), and prediction accuracy is reported for each individual sample. The bars in the figure correspond to individual samples broken down by leukemia type (B) and AML subtype (C). (D) Schema for determining accuracy for leukemia types and AML subclasses in dataset 2.

(E–H) Normalized dataset 2 was randomly split into training and test sets 100 times (D) and prediction accuracy is reported for each individual sample, listed by leukemia type (E) and AML subtype (F). Workflow for M3 subtype prediction using dataset 2 (G) Boxplots of prediction accuracy, sensitivity, and specificity over 100 train/test splits (H). Error bars depict the standard deviation.

Furthermore, we used the rich resource of the present dataset to explore whether prediction across leukemic diseases would be possible as well. For this, we trained a multilabel-classifier on dataset 2 using both datasets 1 and 3 as independent validation sets (Figure S14A). We found good prediction accuracy, sensitivity, and specificity over most tested diseases (Figure S14B); however, a rigorous study over all leukemic conditions would clearly require the inclusion of more training samples for CLL and CML.

Predictive Signatures and AML Biology

The predictive models derived from the lasso and used earlier are sparse in the sense that they automatically select a small number of genes to drive the prediction. The genes are selected in a unified global analysis, rather than by differential expression (DE) on a gene-by-gene basis. From a statistical point of view, global sparsity patterns for prediction and gene-by-gene DE are different criteria. Differentially expressed genes are those that individually have different levels between the groups, whereas genes selected for prediction are those that together perform well in a predictive sense. For the lasso, the selected set of genes also typically includes false-positives with respect to the truly relevant predictors. Furthermore, a good set of genes for prediction need not be mechanistic (in the sense of constituting causal drivers of the disease state). We therefore sought to understand the relationship between DE, known mechanisms, and predictive gene signatures.

Using dataset 2 (the largest dataset) we compared DE and the sparse predictive models (Figure S15). We performed DE analysis using the whole dataset and compared the results with the set of genes in the lasso model ("lasso genes") based on the same data (Figure 6A). A total of 506 genes was differentially expressed ("DE genes"), of which 26 were associated with the disease ontology term or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway for AML ("AML-related genes"). Of the 141 lasso genes, 7 genes were leukemia related and 46 were DE genes, meaning that many of the lasso genes were not differentially expressed, as clearly seen when overlaying the lasso gene selection on a volcano plot (Figure 6A). This underlines the fact that DE and predictive value in a signature sense are different criteria.

Next, we extended this analysis to focus on DE and lasso genes whose selection was robust to data subsampling. This was done by subsampling half the dataset randomly 100 times and in each such subsample carrying out the full DE and lasso analyses. For the lasso this type of approach has been studied under the name stability selection (Meinshausen and Bühlmann, 2010). DE and lasso genes were then scored according to the frequency with which they appeared among the 100 rounds of selection (Figure 6B). Thus, an inclusion score of 100% for a DE gene means that the gene is selected as differentially expressed in all 100 iterations, and similarly for the lasso genes. In total, 669 genes passed the DE cutoffs in at least 50% of the iterations, whereas 80 genes were called in at least 50% of the iterations by the lasso model (Figure 6B). Of these genes, 35 were called according to both criteria. The above-mentioned results show that even among the genes that are included in the lasso models with high frequency (i.e., those genes that are robustly selected for prediction), many are not differentially expressed.

Next, we excluded the 155 known AML genes that are associated with the disease ontology term or KEGG pathway for AML from the prediction, which did not affect disease prediction at all (Figure 6C), highlighting the strong robustness of the classifier. To better understand the potential biological relevance for AML of the 35 genes that were robustly called under both DE and lasso criteria (Figure 6B), we visualized the top-ranked genes over all 8,348 samples within the dataset by hierarchical clustering of *z*-transformed expression values (Figure 6D). We identified one distinct cluster of genes with the majority of genes being elevated in AML compared with other leukemias and non-leukemic samples (cluster 1, $n = 29$). Although we identified several well-known AML-related genes (gene name in red color) such as the KIT Proto-Oncogene Receptor Tyrosine Kinase (KIT) (Gao et al., 2015; Heo et al., 2017; Ikeda et al., 1991), RUNX2 (Kuo et al., 2009), and FLT3 (Bullinger et al., 2008; Carow et al., 1996) in this cluster, many genes have not yet been

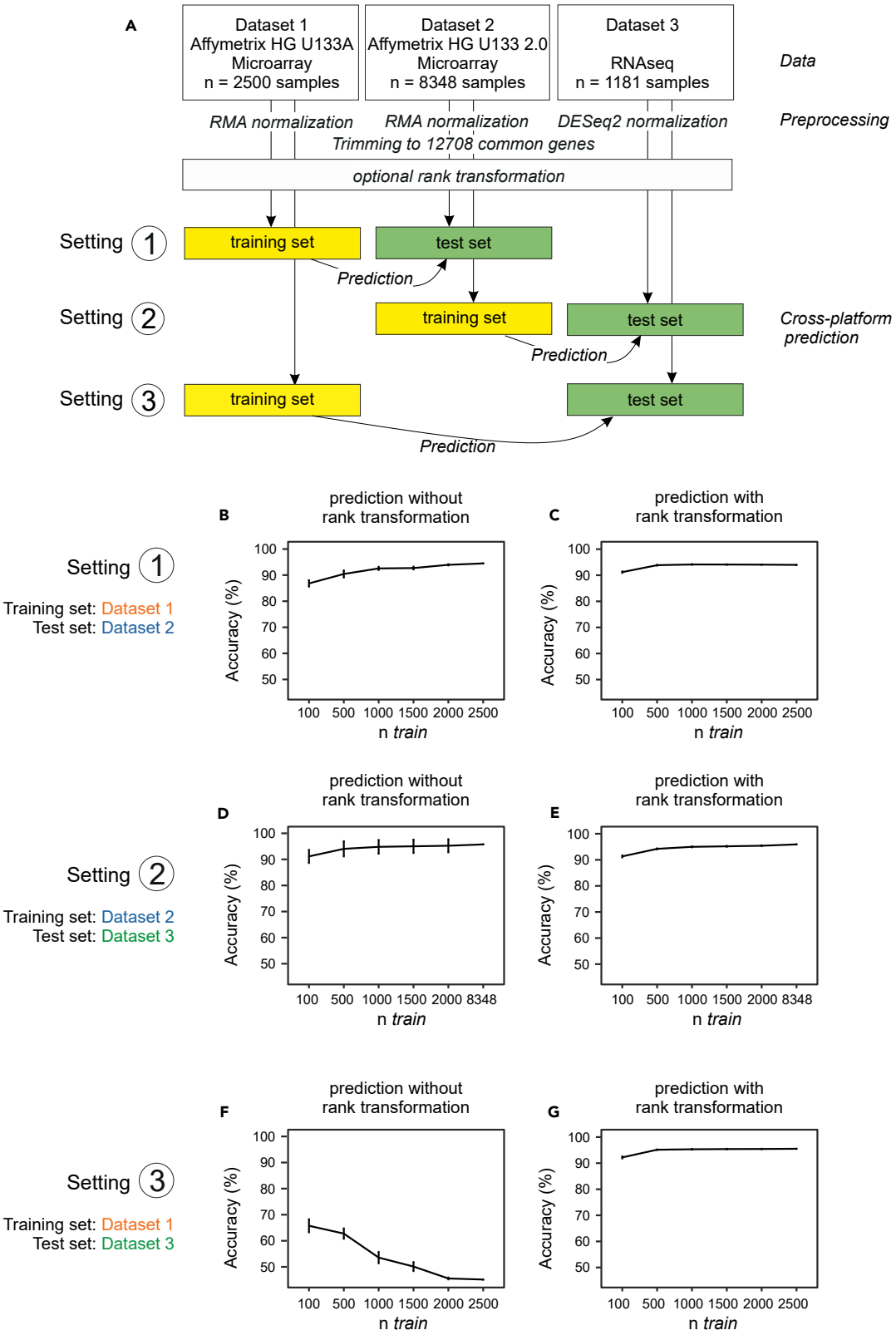

A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
optional rank transformation
Setting 1
training set
test set
Prediction
Setting 2
training set
test set
Cross-platform prediction
Prediction
Setting 3
training set
test set
Prediction
B
prediction without rank transformation
C
prediction with rank transformation
Setting 1
Training set: Dataset 1
Test set: Dataset 2
Accuracy (%)
n train
D
prediction without rank transformation
E
prediction with rank transformation
Setting 2
Training set: Dataset 2
Test set: Dataset 3
F
prediction without rank transformation
G
prediction with rank transformation
Setting 3
Training set: Dataset 1
Test set: Dataset 3

Figure 5. Translating Predictive Signatures across Technological Platforms

(A) Schema of signature translation across platforms. Datasets were normalized individually and trimmed to 12,708 common genes. The classifiers were trained on subsamples of different sizes on one platform and tested on all samples of another platform.

(B–G) Classification accuracies are shown as a function of training sample size (n_{train}) without rank transformation (B, D, and F) and with rank transformation (C, E, and G). For the latter case, the training and test datasets (from different platforms) were separately rank transformed (see text for details). Error bars depict the standard deviation.

See also Figures S10–S13.

linked to AML biology, and, although not the focus of the present article, further study of these genes may be interesting from a mechanistic point of view. Within the other cluster (cluster 2, n = 6 genes), genes had reduced expression values in AML compared with other leukemias and two of these genes have been linked to other types of leukemias (gene names in orange color).

DISCUSSION

Despite the pioneering studies by Golub and others (Debernardi et al., 2003; Kohlmann et al., 2003; Ross et al., 2004; Schoch et al., 2002; Virtaneva et al., 2001) suggesting high potential value of GEP for primary AML diagnosis and differential diagnosis, current recommendations for diagnosing this disease currently center on classical approaches including assessment of morphology, immunophenotyping, cytochemistry, and cytogenetics (Döhner et al., 2017). Analyzing more than 12,000 samples from more than 100 individual studies, we provide evidence that combining large transcriptomic data with ML allows for the development of robust disease classifiers. Such classifiers could, in the future, potentially assist in primary diagnosis of this deadly disease particularly in settings where hematological expertise is not sufficiently available and/or costly. Considering the increased utilization of whole-genome and whole-transcriptome sequencing in the management of patients with cancer, we propose that application of GEP- and ML-based classifiers for diagnosis needs to be re-evaluated. This is in line with previous suggestions by the International Microarray Innovations in Leukemia Study Group (Haferlach et al., 2010). Furthermore, we suggest that similar analyses may be useful for other diseases when analyzing whole blood or PBMC-derived gene expression profiles, or for multiple conditions in parallel (see later).

We sought to understand and address some of the bottlenecks in the way of clinical deployment of transcriptomic-based ML tools for diagnosis. To this end, we considered a range of practical scenarios, including cross-study issues and prediction across different technological platforms. We found that accurate prediction is possible across a range of scenarios and, in many cases, with relatively few training samples. However, we also showed that depending on the use case and the associated prevalence, large training sets may be required to reach accuracies high enough to yield acceptable PPVs.

Our results show that with existing technologies it is potentially possible to achieve good performance in a near-automated fashion. An ML-plus-genomics approach can be run at very low marginal cost: the RNA assays can already be done at <$100 (and this continues to fall), and in the long-term these costs will drop still further. To our knowledge, this is already in a cost range that is lower than the combined use of morphology, immunophenotyping, and cytochemistry for primary AML diagnosis. Furthermore, the sparse models we considered, once trained, require only a small subset of the genome, hence custom sequencing pipelines could be used. Marginal cost is important precisely because it opens up the possibility of a truly scalable detection/diagnosis strategy. One example of a recently developed, very-low-cost whole-transcriptomics protocol is BRB-seq which allows generating genome-wide transcriptomic data at a similar cost as profiling four genes using RT-qPCR (Alpern et al., 2019), which could be a candidate for future clinical development. Furthermore, recent developments in nanopore sequencing (Byrne et al., 2017) suggest that in the future, delivery of transcriptomic assays could be greatly simplified, and this, combined with cloud- or local-device-based ML prediction, would represent a paradigm shift in terms of scalability and accessibility. Such transcriptome-based ML might therefore also be utilized at an earlier time point in the disease course, when patients present with non-specific symptoms to their primary care physician. Here, ML-based diagnostics might assist a faster transfer of the patient to specialized hematology centers for complete diagnostics and therapeutic management.

The next steps toward better understanding ML-based diagnosis for AML would include prospective studies specifically aimed at assessing diagnostic utility. Before any development in the future pivotal clinical trials for approval with the respective regulatory bodies would be required. Naturally, any such development would require additional, independent studies with the development of deployment-ready

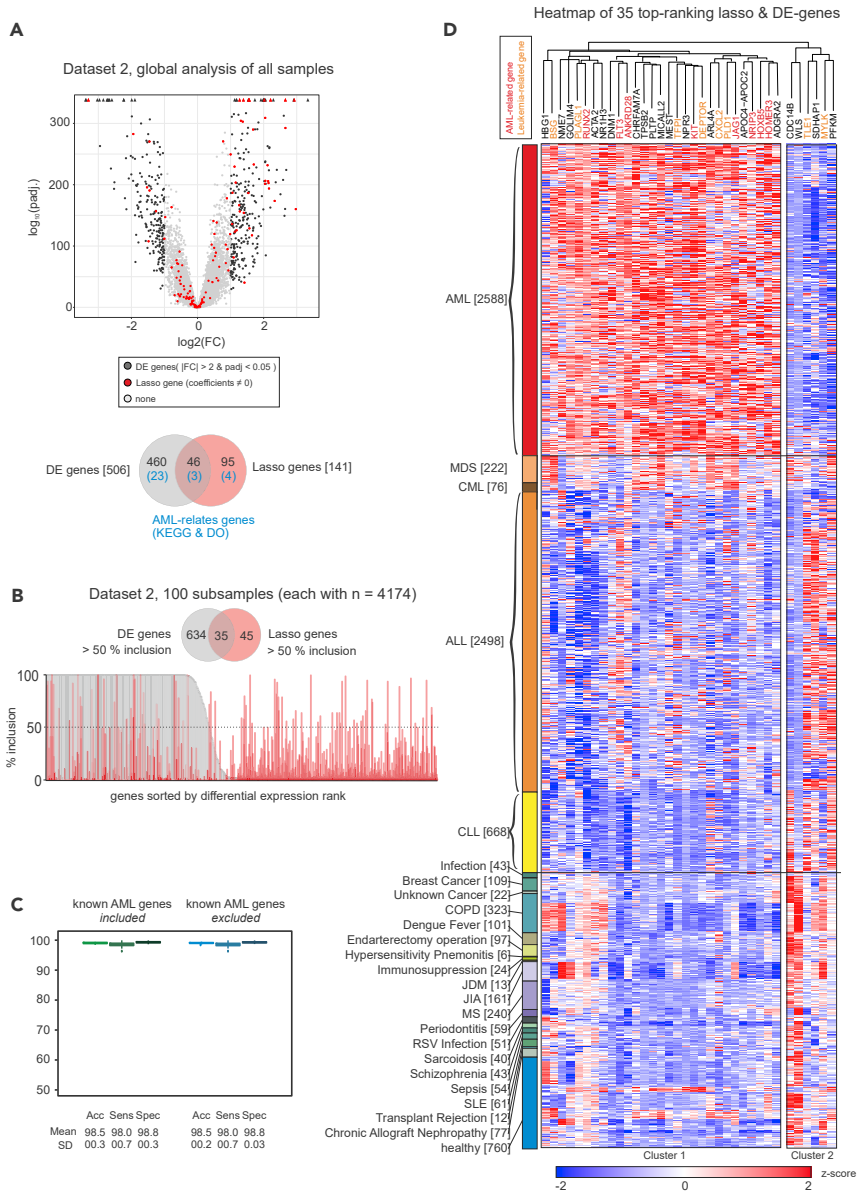


Figure 6. Predictive Signatures and AML Biology

(A) Volcano plot of global differentially expressed (DE) genes and genes of the lasso model ("lasso genes") in dataset 2, and Venn diagram indicating the overlap of both gene sets and the genes included in the KEGG pathway or the disease ontology term "AML."

(B) Inclusion plot of DE genes and lasso genes in 100 random permutations of dataset 2. The plot is sorted according to DE gene rank, and a Venn diagram shows the overlap between genes with a minimum of 50% inclusion.

(C) Boxplot of accuracy, sensitivity, and specificity of the predictive model trained and tested on random subsets of dataset 2 with inclusion of all genes of the dataset and without 155 genes known to be relevant for AML biology. Error bars depict the standard deviation.

(D) Heatmap and hierarchical clustering of z-scaled expression values of 35 genes with >50% inclusion both in lasso and DE genes, as shown in (B). Genes with known associations with AML are marked red; genes associated with other types of leukemia are labeled in orange.

See also Figure S15.

pipelines, which by itself is a nontrivial undertaking (as discussed in Keane and Topol, 2018). However, initial prospective studies have already been started, such as the 5000 genomes project (https://www.mll.com/en/science/5000-genome-project.html), which also performs RNA-seq to develop such a classifier

for the clinics. It is also important to emphasize that just as regulatory standards have evolved for classical diagnostics, so too will new regulatory frameworks be needed for ML-assisted diagnostics in the future (Keane and Topol, 2018).

An additional point concerns explicit and implicit thresholds at which a suspected case is entered into the pipeline in the first place. A lower threshold for entry could lower false-negatives and reduce the risk of delayed treatment (which has been associated with worse outcomes, notably in younger patients; Sekeres et al., 2008). Using current diagnostic systems any such change would dramatically increase the overall costs; in contrast, more efficient solutions would allow thresholds to be optimized for patient benefit while keeping the overall costs controlled. Naturally any modification to the overall diagnostic strategy would need a full health economic and decision analysis (accounting in particular for a necessarily higher false-positive rate) and case-by-case assessment. For some diseases it may be the case that earlier entry into a diagnostic pipeline would overall *not* be beneficial, a point that is widely appreciated in the context of population-level screening (see, e.g., Jacobs et al., 2016). Nevertheless, the point is that scalable diagnostic strategies increase the scope for optimization of decision making for patient benefit.

We saw also encouraging results across other conditions. Although the data used in the present study do not allow rigorous study of diagnosis across multiple conditions, we conjecture that diagnosis of multiple conditions from blood transcriptomes may be possible, opening up the possibility of training multi-class classifiers on blood transcriptomic data. Note that this would allow diagnosis of several conditions at essentially the *same* marginal cost per additional sample, bolstering the economic case outlined earlier. Rigorous study would require new pan-disease study designs, but we think that such approaches could lead to large efficiency gains in the future.

All our models were learned in an unbiased manner, directly from the full transcriptome data with no prior biological knowledge or any pre-selection of genes. We showed that genes relevant for prediction were often not differentially expressed and that prediction was robust to removal of known AML-related genes. These observations illustrate two points of relevance to clinical applications. First, for prediction it can be more fruitful to consider signatures derived in data-driven, genome-wide fashion than to think in terms of single genes or DE. Second, high-dimensional analyses, although complex relative to more classical methods, can be highly predictive as well as robust to the presence or absence of specific genes. Taken together, our results underline the immense value of making GEP data publicly available, allowing for new and large-scale multi-study analyses. Furthermore, we support the notion that the application of ML approaches based on sequencing data to identify gene signatures for certain diseases such as AML will become part of recommendations for diagnosis and management of AML. We envision that combining whole-genome and whole-transcriptome analysis based on ML algorithms will ultimately allow early detection, diagnosis, differential diagnosis, subclassification, and outcome prediction in an integrated fashion.

Limitations of the Study

It is important to note that the data used here were pooled from multiple studies with different designs and goals. Further work, including suitably designed prospective studies, would be needed to better understand the diagnostic utility of an ML-plus-transcriptomics approach. Site- and study-specific effects may be relevant for clinical applications. This is because a classifier once learned might be deployed in a range of new settings (sites, regions) that could lead in a number of ways to unwanted variation. If training and test sets are very different, this can impact performance. In clinical applications of predictive models it will be important to continually track performance even after deployment and the possibility of distributional shifts that require more complex analyses cannot be ruled out.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

Processed data can be accessed via the SuperSeries GSE122517 or via the individual SubSeries GSE122505 (dataset 1), GSE122511 (dataset 2), and GSE122515 (dataset 3). The code for preprocessing and for predictions can be found at GitHub (https://github.com/schultzelab/aml_classifer). In addition, all data and

package versions are stored in a docker container on Docker Hub (https://hub.docker.com/r/schultzelab/aml_classifier, Table S3).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.100780.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC2151—390873048. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733100. J.L.S is member of the Helmholtz network Sparse2Big.

AUTHOR CONTRIBUTIONS

Conceptualization, J.L.S. and S.M.; Methodology, S.M., J.L.S., and S.W.-H.; Software, S.W.-H., K.P., and B.T.; Validation, B.T.; Formal Analysis, S.W.-H., T.U., K.P., J.S.-S., K.K., M.B., L.S.; Investigation, S.W.-H., T.U., P.G., K.B.; Resources, T.H., S.W.-H.; Data Curation, S.W.-H.; Writing – Original Draft, J.L.S., S.M., S.W.-H.; Visualization, S.W.-H., Supervision, J.L.S., S.M., M.B.

DECLARATION OF INTERESTS

There are no competing interests.

Received: August 26, 2019
Revised: December 3, 2019
Accepted: December 12, 2019
Published: January 24, 2020

REFERENCES

Alpern, D., Gardeux, V., Russeil, J., Mangeat, B., Meireles-Filho, A.C.A., Breysse, R., Hacker, D., and Deplancke, B. (2019). BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biol. *20*, 71.

Andersson, A., Ritz, C., Lindgren, D., Edén, P., Lassen, C., Heldrup, J., Olofsson, T., Råde, J., Fontes, M., Porwit-MacDonald, A., et al. (2007). Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. Leukemia *21*, 1198–1203.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405.

Brynjolfsson, E., and McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (W W Norton & Co).

Bühlmann, P., and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer).

Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R.F., Tibshirani, R., Döhner, H., and Pollack, J.R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N. Engl. J. Med. *350*, 1605–1616.

Bullinger, L., Döhner, K., Kranz, R., Stirner, C., Fröhling, S., Scholl, C., Kim, Y.H., Schlenk, R.F., Tibshirani, R., Döhner, H., et al. (2008). An FLT3 gene-expression signature predicts clinical outcome in normal karyotype AML. Blood *111*, 4490–4495.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. *8*, 16027.

Carow, C.E., Levenstein, M., Kaufmann, S.H., Chen, J., Amin, S., Rockwell, P., Witte, L., Borowitz, M.J., Civin, C.I., and Small, D. (1996). Expression of the hematopoietic growth factor receptor FLT3 (STK-UFlk2) in human leukemias. Blood *87*, 1089–1096.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. Cell *163*, 506–519.

Debernardi, S., Lillington, D.M., Chaplin, T., Tomlinson, S., Amess, J., Rohatiner, A., Lister, T.A., and Young, B.D. (2003). Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. Genes Chromosomes Cancer *37*, 149–158.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Döhner, H., Estey, E.H., Amadori, S., Appelbaum, F.R., Buchner, T., Burnett, A.K., Dombret, H., Fenaux, P., Grimwade, D., Larson, R.A., et al. (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood *115*, 453–474.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F.R., Büchner, T., Dombret, H., Ebert, B.L., Fenaux, P., Larson, R.A., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood *129*, 424–447.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Gao, X., Lin, J., Gao, L., Deng, A., Lu, X., Li, Y., Wang, L., and Yu, L. (2015). High expression of c-kit mRNA predicts unfavorable outcome in adult patients with t(8;21) acute myeloid leukemia. PLoS One *10*, e0124241.

Garzon, R., Volinia, S., Papaioannou, D., Nicolet, D., Kohlschmidt, J., Yan, P.S., Mrózek, K., Bucci, D., Carroll, A.J., Baer, M.R., et al. (2014). Expression and prognostic impact of lncRNAs in acute myeloid leukemia. Proc. Natl. Acad. Sci. U S A *111*, 18679–18684.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Béné, M.-C., De Vos, J., Hernández, J.M., Hofmann, W.-K., Mills, K.I., et al. (2010). Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. J. Clin. Oncol. *28*, 2529–2537.

Heath, E.M., Chan, S.M., Minden, M.D., Murphy, T., Shlush, L.I., and Schimmer, A.D. (2017). Biological and clinical consequences of NPM1 mutations in AML. Leukemia *31*, 798–807.

Heo, S.-K., Noh, E.-K., Kim, J.Y., Jeong, Y.K., Jo, J.-C., Choi, Y., Koh, S., Baek, J.H., Min, Y.J., and Kim, H. (2017). Targeting c-KIT (CD117) by dasatinib and radotinib promotes acute myeloid leukemia cell death. Sci. Rep. *7*, 15278.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Hornung, R., Boulesteix, A.-L., and Causeur, D. (2016). Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. BMC Bioinformatics *17*, 27.

Hornung, R., Causeur, D., Bernau, C., and Boulesteix, A.-L. (2017). Improving cross-study prediction through addon batch effect adjustment or addon normalization. Bioinformatics *33*, 397–404.

Ikeda, H., Kanakura, Y., Tamaki, T., Kuriu, A., Kitayama, H., Ishikawa, J., Kanayama, Y., Yonezawa, T., Tarui, S., and Griffin, J. (1991). Expression and functional role of the proto-oncogene c-kit in acute myeloblastic leukemia cells. Blood *78*, 2962–2968.

Jacob, L., Gagnon-Bartsch, J.A., and Speed, T.P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Biostatistics *17*, 16–28.

Jacobs, I.J., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., Kalsi, J.K., Amso, N.N., Apostolidou, S., Benjamin, E., Cruickshank, D., et al. (2016). Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet *387*, 945–956.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Keane, P.A., and Topol, E.J. (2018). With an eye to AI and autonomous diagnosis. NPJ Digit. Med. *1*, 40.

Kohlmann, A., Schoch, C., Schnittger, S., Dugas, M., Hiddemann, W., Kern, W., and Haferlach, T. (2003). Molecular characterization of acute leukemias by use of microarray technology. Genes Chromosomes Cancer *37*, 396–405.

Kristensen, V.N., Vaske, C.J., Ursini-Siegel, J., Van Loo, P., Nordgard, S.H., Sachidanandam, R., Sorlie, T., Warnberg, F., Haakensen, V.D., Helland, A., et al. (2012). Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. Proc. Natl. Acad. Sci. U S A *109*, 2802–2807.

Kuo, Y.-H., Zaidi, S.K., Gornostaeva, S., Komori, T., Stein, G.S., and Castilla, L.H. (2009). Runx2 induces acute myeloid leukemia in cooperation with Cbfbeta-SMMHC in mice. Blood *113*, 3323–3332.

Lavallee, V.-P., Lemieux, S., Boucher, G., Gendron, P., Boivin, I., Armstrong, R.N., Sauvageau, G., and Hebert, J. (2016). RNA-sequencing analysis of core binding factor AML identifies recurrent ZBTB7A mutations and defines RUNX1-CBFA2T3 fusion signature. Blood *127*, 2498–2501.

Lavallée, V.-P., Baccelli, I., Krosl, J., Wilhelm, B., Barabé, F., Gendron, P., Boucher, G., Lemieux, S., Marinier, A., Meloche, S., et al. (2015). The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. Nat. Genet. *47*, 1030–1037.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. *11*, 733–739.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66–72.

Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. N. Engl. J. Med. *363*, 2424–2433.

Loriaux, M.M., Levine, R.L., Tyner, J.W., Fröhling, S., Scholl, C., Stoffregen, E.P., Wernig, G., Erickson, H., Eide, C.A., Berger, R., et al. (2008). High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia. Blood *111*, 4788–4796.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Macrae, T., Sargeant, T., Lemieux, S., Hébert, J., Deneault, E., and Sauvageau, G. (2013). RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. PLoS One *8*, e72884.

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. J. R. Stat. Soc. *72*, 417–473.

Pabst, C., Bergeron, A., Lavallee, V.-P., Yeh, J., Gendron, P., Norddahl, G.L., Krosl, J., Boivin, I., Deneault, E., Simard, J., et al. (2016). GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. Blood *127*, 2018–2027.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N. Engl. J. Med. *374*, 2209–2221.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised Risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. (2009). Dataset Shift in Machine Learning (MIT Press).

Robertson, A.G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A.D., Hinoue, T., Laird, P.W., Hoadley, K.A., Akbani, R., et al. (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell *171*, 540–556.e25.

Ross, M.E., Mahfouz, R., Onciu, M., Liu, H.-C., Zhou, X., Song, G., Shurtleff, S.A., Pounds, S., Cheng, C., Ma, J., et al. (2004). Gene expression profiling of pediatric acute myelogenous leukemia. Blood *104*, 3679–3687.

Schoch, C., Kohlmann, A., Schnittger, S., Brors, B., Dugas, M., Mergenthaler, S., Kern, W., Hiddemann, W., Eils, R., and Haferlach, T. (2002). Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. Proc. Natl. Acad. Sci. U S A *99*, 10008–10013.

Sekeres, M.A., Elson, P., Kalaycio, M.E., Advani, A.S., Copelan, E.A., Faderl, S., Kantarjian, H.M., and Estey, E. (2008). Time from diagnosis to treatment initiation predicts survival in younger, but not older, acute myeloid leukemia patients. Blood *113*, 28–36.

The Cancer Genome Atlas Research Network (TCGA), Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., de La Chapelle, A., and Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid

leukemia with isolated trisomy 8 and normal cytogenetics. Proc. Natl. Acad. Sci. U S A *98*, 1124–1129.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in Acute Myeloid Leukemia. Cell *150*, 264–278.

Yan, X.-J., Xu, J., Gu, Z.-H., Pan, C.-M., Lu, G., Shen, Y., Shi, J.-Y., Zhu, Y.-M., Tang, L., Zhang, X.-W., et al. (2011). Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat. Genet. *43*, 309–315.

Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. PLoS One *9*, e85150.

iScience, Volume *23*

Supplemental Information

Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics

Stefanie Warnat-Herresthal, Konstantinos Perrakis, Bernd Taschler, Matthias Becker, Kevin Baßler, Marc Beyer, Patrick Günther, Jonas Schulte-Schrepping, Lea Seep, Kathrin Klee, Thomas Ulas, Torsten Haferlach, Sach Mukherjee, and Joachim L. Schultze

Figure S1

A

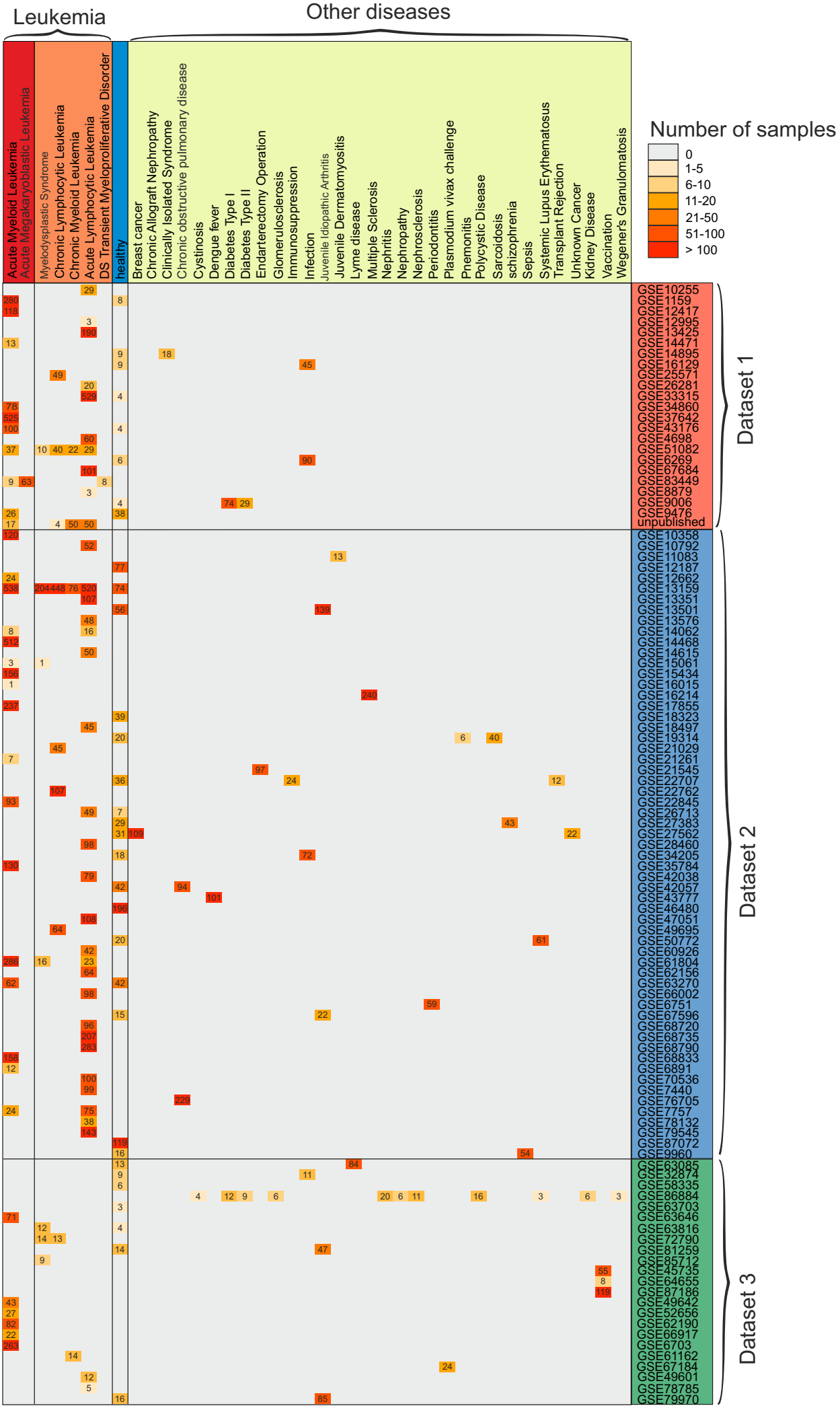
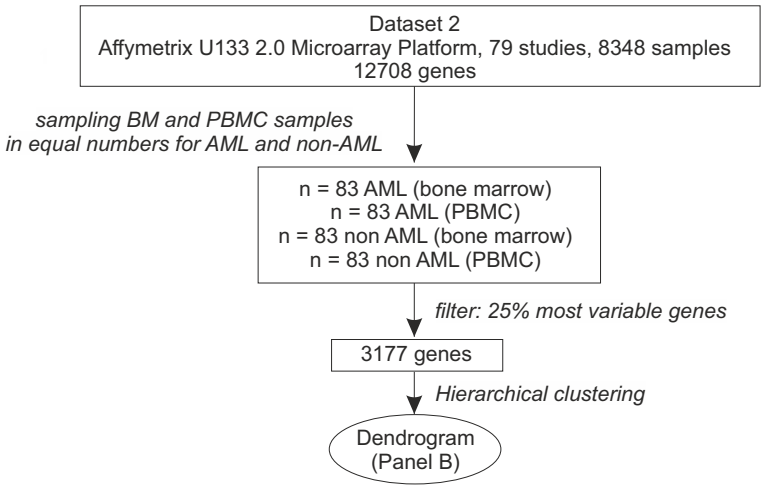

Leukemia
Other diseases
Acute Myeloid Leukemia
Acute Megakaryoblastic Leukemia
Myelodysplastic Syndrome
Chronic Lymphocytic Leukemia
Chronic Myeloid Leukemia
Acute Lymphocytic Leukemia
DS Transient Myeloproliferative Disorder
healthy
Breast cancer
Chronic Allograft Nephropathy
Clinically Isolated Syndrome
Chronic obstructive pulmonary disease
Cystinosis
Dengue fever
Diabetes Type I
Diabetes Type II
Endarterectomy Operation
Glomerulosclerosis
Immunosuppression
Infection
Juvenile Idiopathic Arthritis
Juvenile Dermatomyositis
Lyme disease
Multiple Sclerosis
Nephritis
Nephropathy
Nephrosclerosis
Periodontitis
Plasmodium vivax challenge
Pneumonitis
Polycystic Disease
Sarcoidosis
schizophrenia
Sepsis
Systemic Lupus Erythematosus
Transplant Rejection
Unknown Cancer
Kidney Disease
Vaccination
Wegener's Granulomatosis
Number of samples
0
1-5
6-10
11-20
21-50
51-100
> 100
GSE10255
GSE1159
GSE12417
GSE12995
GSE13425
GSE14471
GSE14895
GSE16129
GSE25571
GSE26281
GSE33315
GSE34860
GSE37642
GSE43176
GSE4698
GSE51082
GSE6269
GSE67684
GSE83449
GSE8879
GSE9006
GSE9476
unpublished
Dataset 1
GSE10358
GSE10792
GSE11083
GSE12187
GSE12662
GSE13159
GSE13351
GSE13501
GSE13576
GSE14062
GSE14468
GSE14615
GSE15061
GSE15434
GSE16015
GSE16214
GSE17855
GSE18323
GSE18497
GSE19314
GSE21029
GSE21261
GSE21545
GSE22707
GSE22762
GSE22845
GSE26713
GSE27383
GSE27562
GSE28460
GSE34205
GSE35784
GSE42038
GSE42057
GSE43777
GSE46480
GSE47051
GSE49695
GSE50772
GSE60926
GSE61804
GSE62156
GSE63270
GSE66002
GSE6751
GSE67596
GSE68720
GSE68735
GSE68790
GSE68833
GSE6891
GSE70536
GSE7440
GSE76705
GSE7757
GSE78132
GSE79545
GSE87072
GSE9960
Dataset 2
GSE63085
GSE32874
GSE58335
GSE86884
GSE63703
GSE63646
GSE63816
GSE72790
GSE81259
GSE85712
GSE45735
GSE64655
GSE87186
GSE49642
GSE52656
GSE62190
GSE66917
GSE6703
GSE61162
GSE67184
GSE49601
GSE78785
GSE79970
Dataset 3

Figure S2

A


Dataset 2
Affymetrix U133 2.0 Microarray Platform, 79 studies, 8348 samples
12708 genes
sampling BM and PBMC samples
in equal numbers for AML and non-AML
n = 83 AML (bone marrow)
n = 83 AML (PBMC)
n = 83 non AML (bone marrow)
n = 83 non AML (PBMC)
filter: 25% most variable genes
3177 genes
Hierarchical clustering
Dendrogram
(Panel B)

B

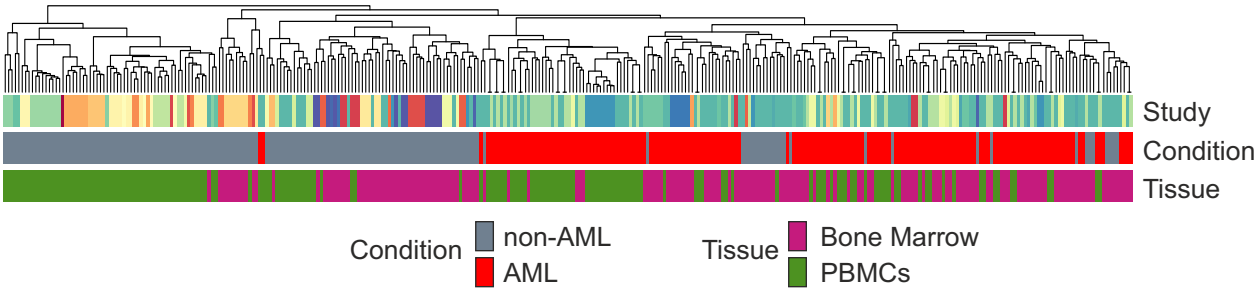
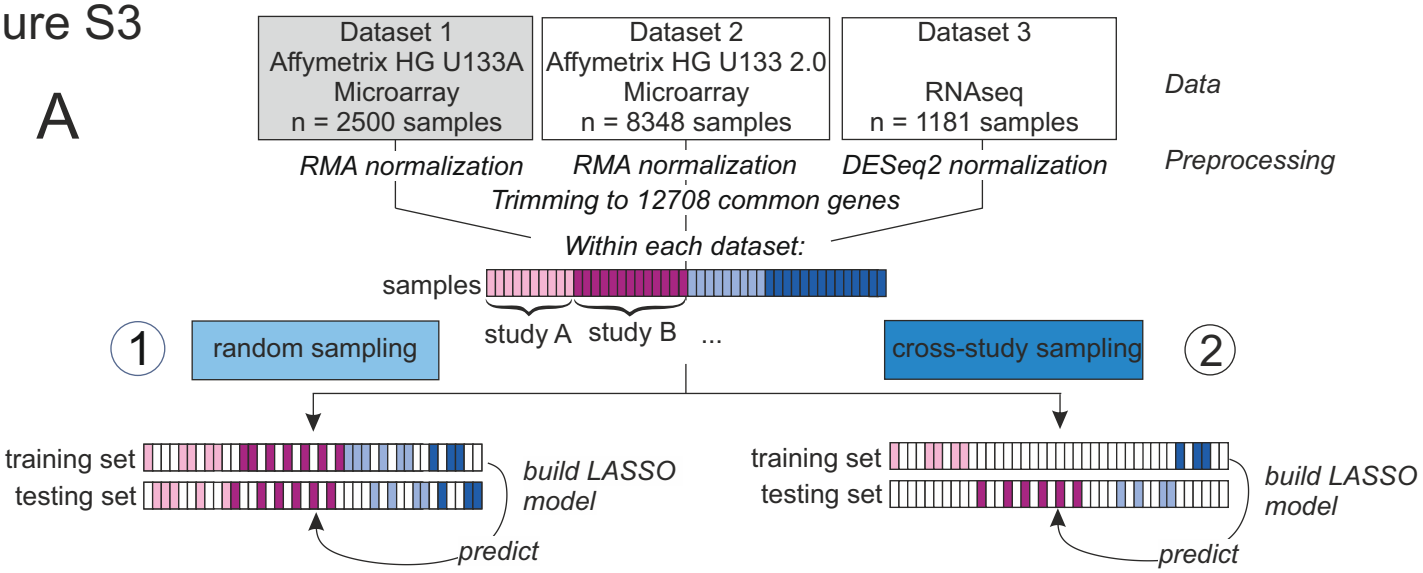
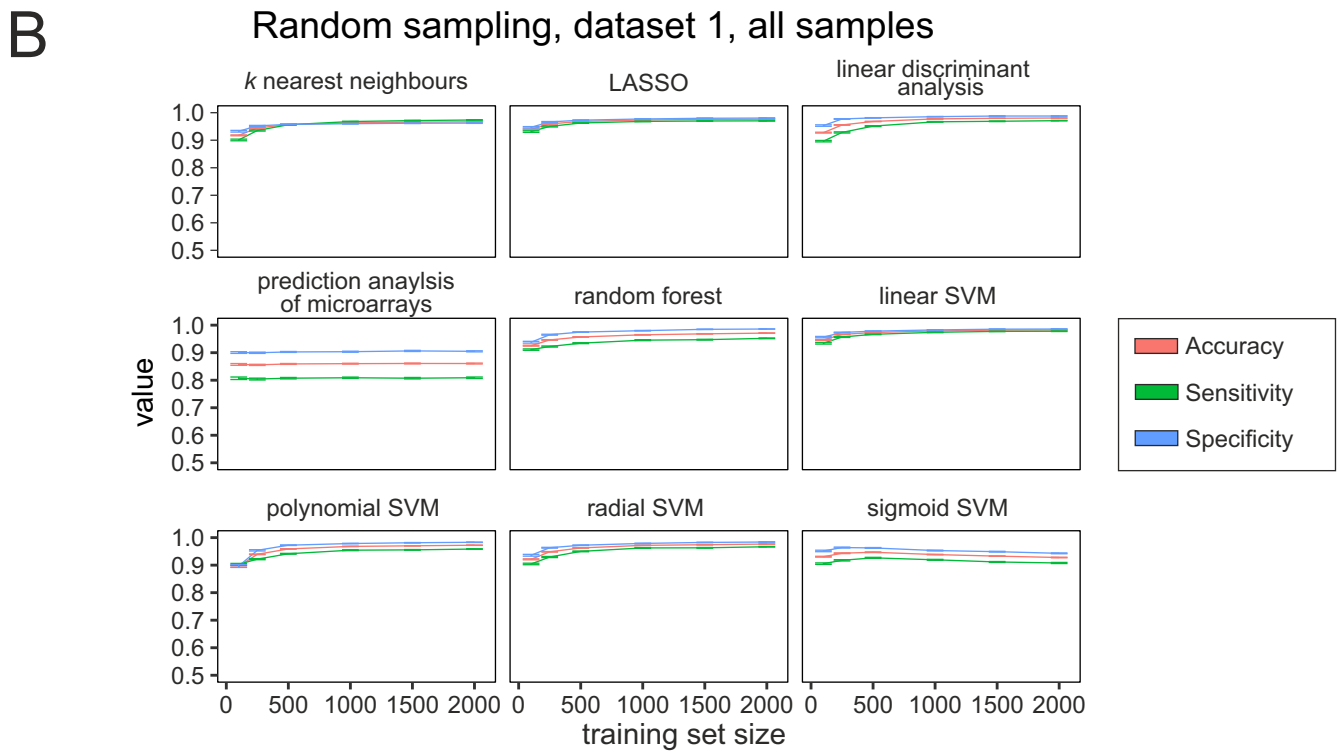

Study
Condition
Tissue
Condition
non-AML
AML
Tissue
Bone Marrow
PBMCs

Figure S3


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Random sampling, dataset 1, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

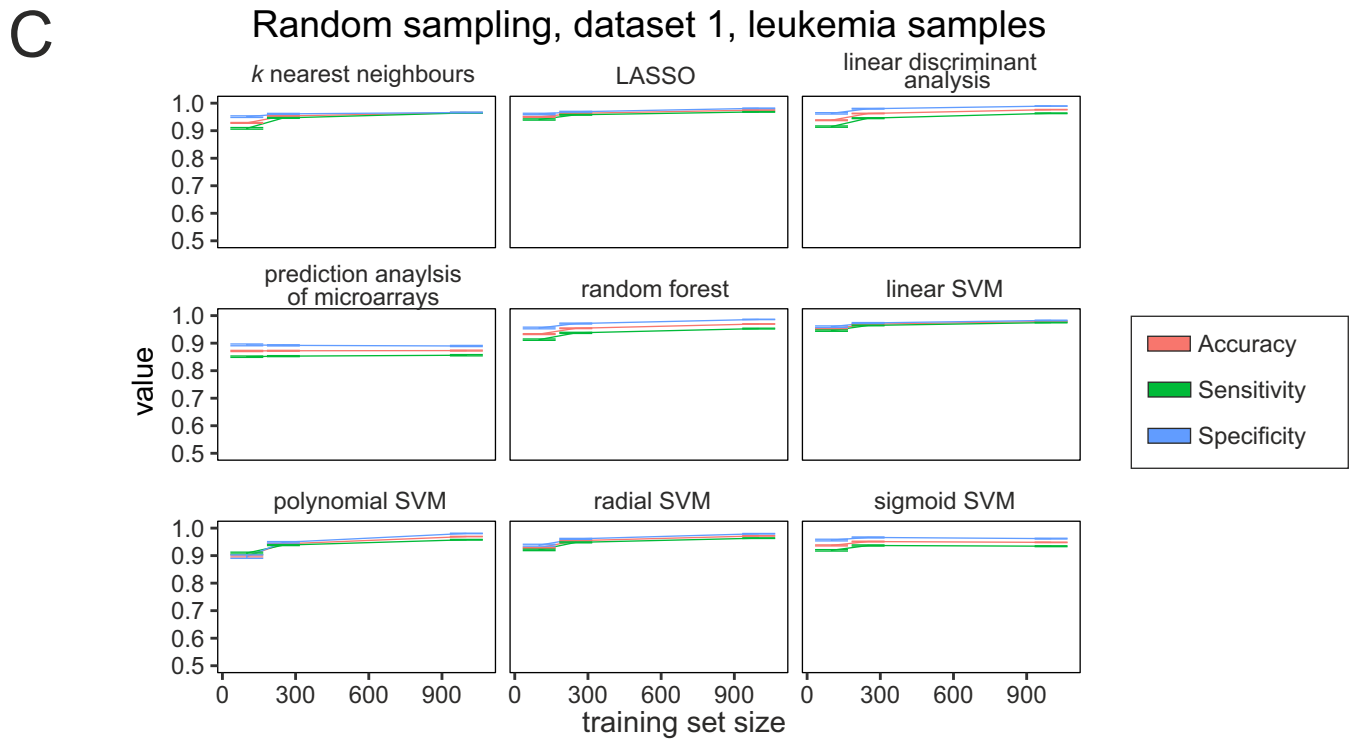

C
Random sampling, dataset 1, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S4

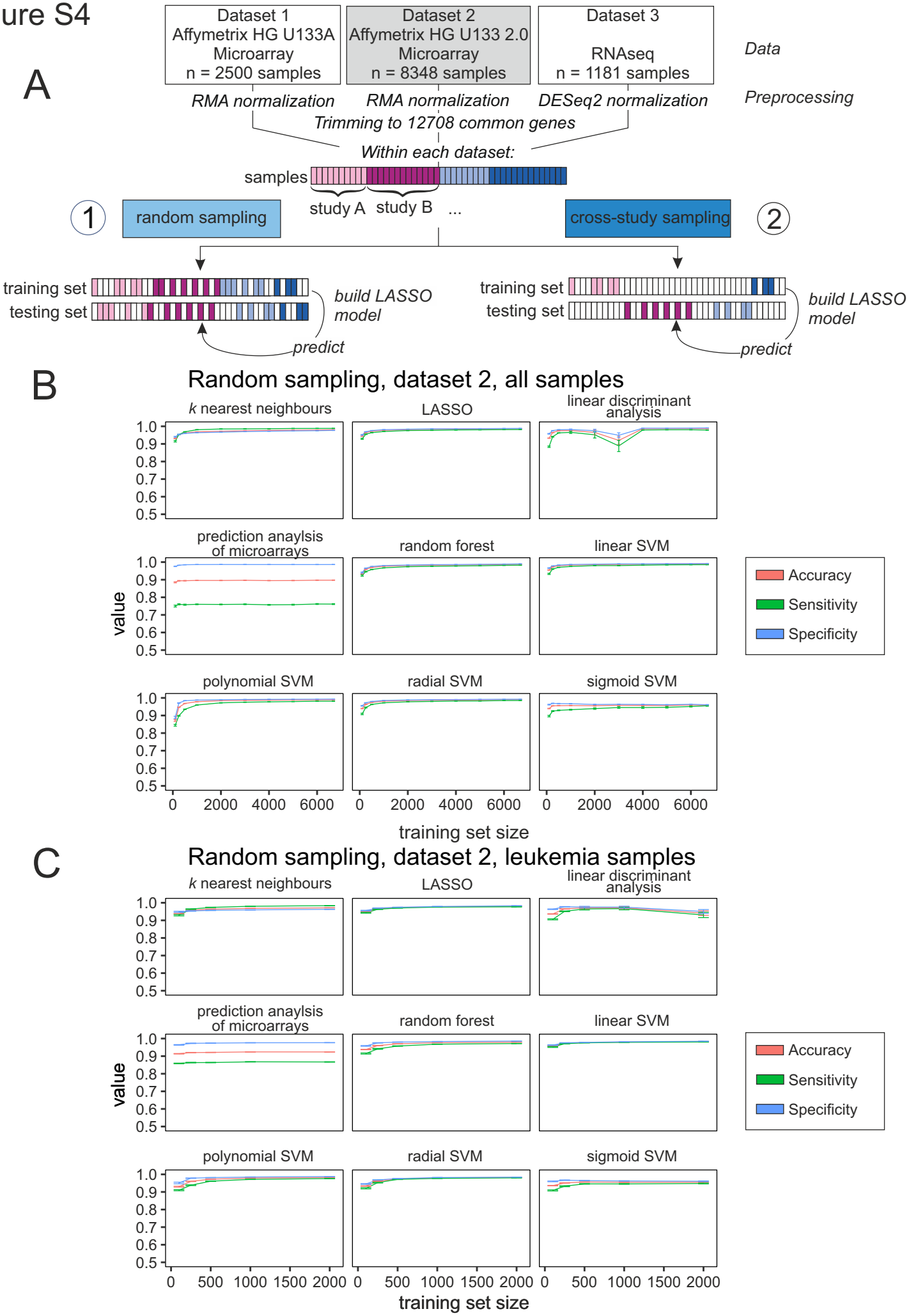
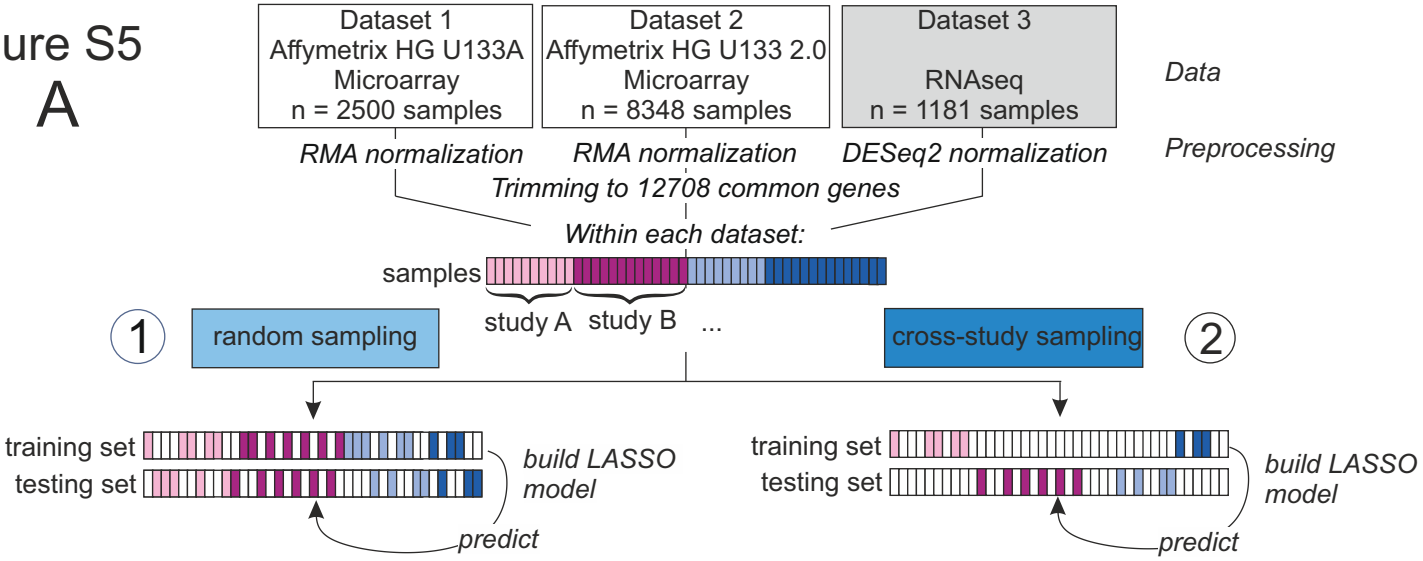

A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict
B
Random sampling, dataset 2, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
Accuracy
Sensitivity
Specificity
training set size
C
Random sampling, dataset 2, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
Accuracy
Sensitivity
Specificity
training set size

Figure S5

A


Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict

B

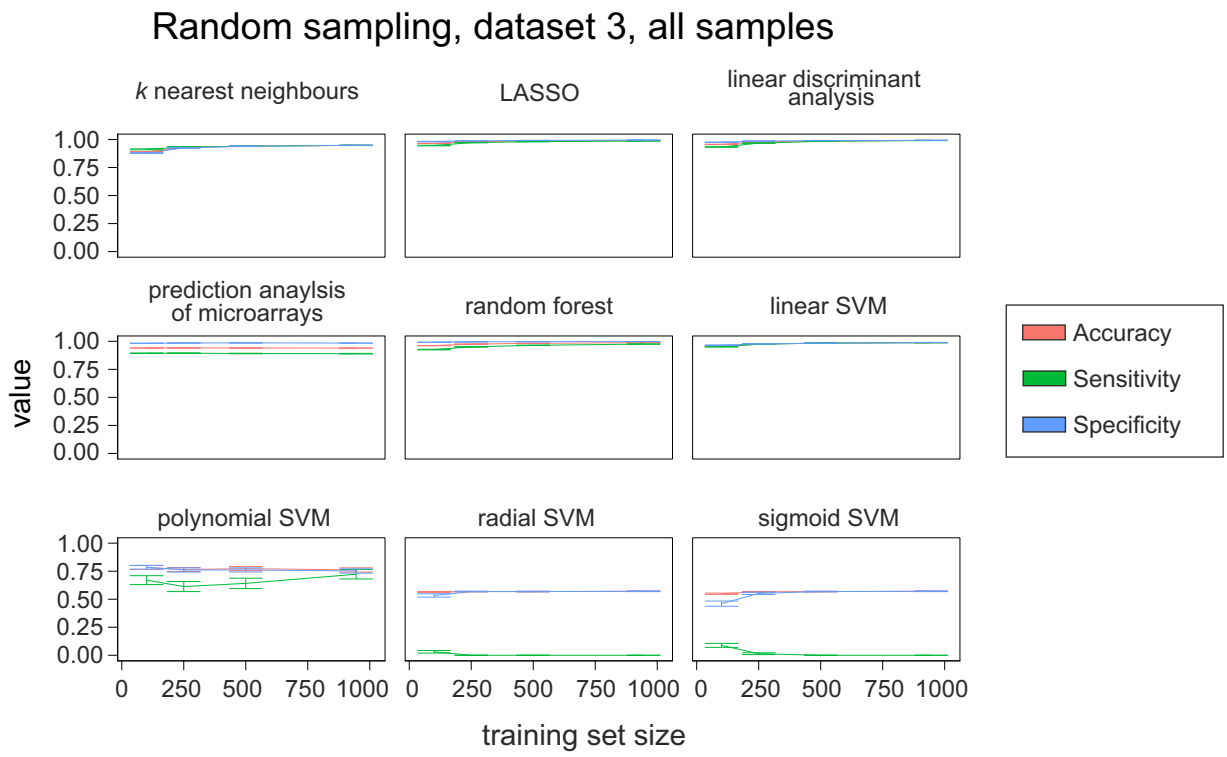
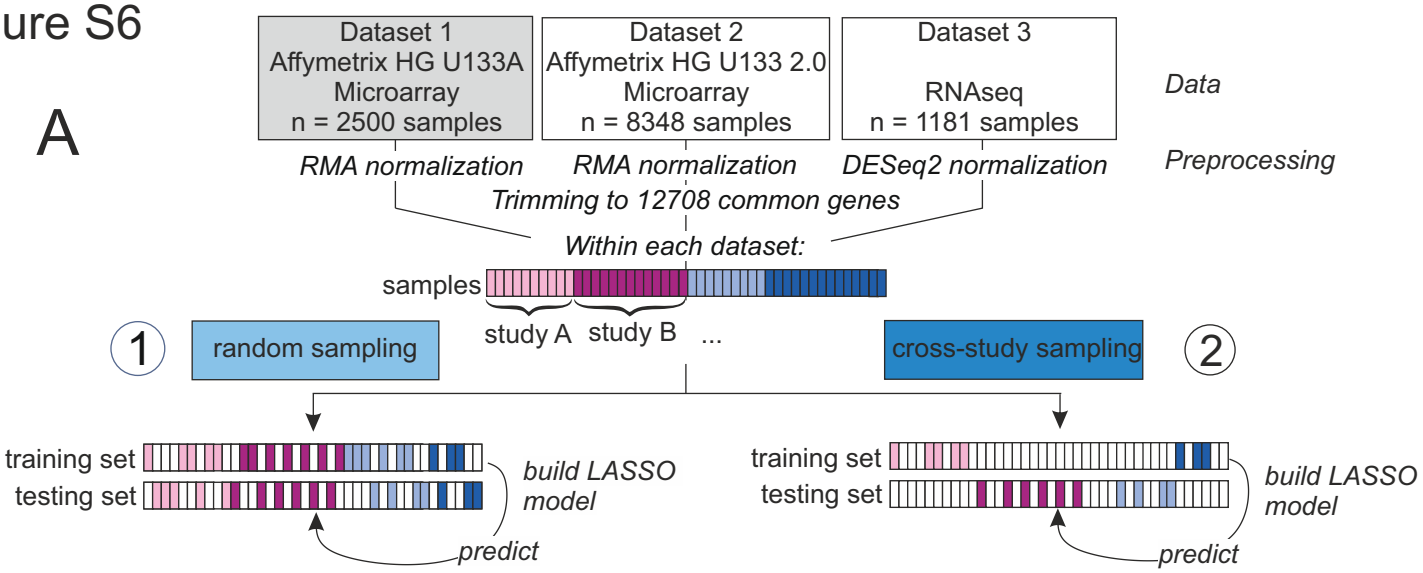
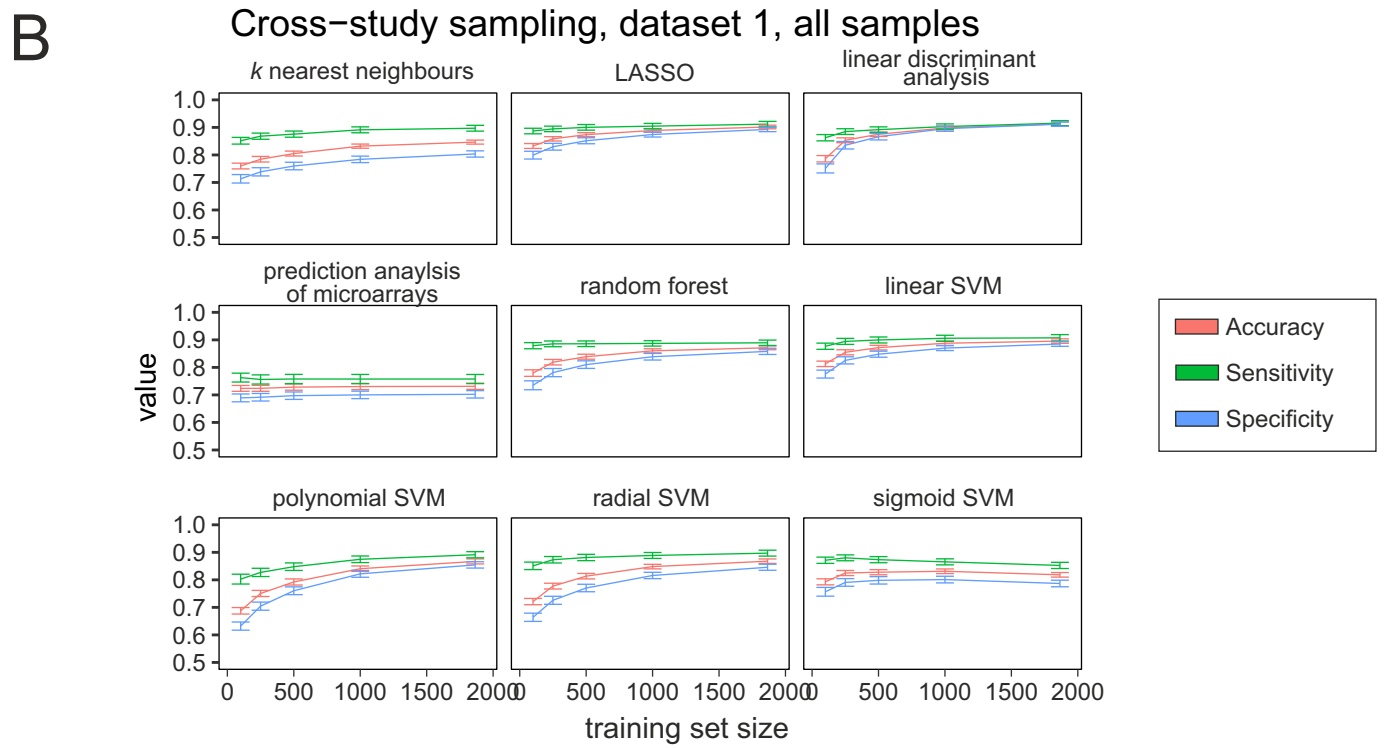

Random sampling, dataset 3, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
training set size
Accuracy
Sensitivity
Specificity

Figure S6


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Cross−study sampling, dataset 1, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

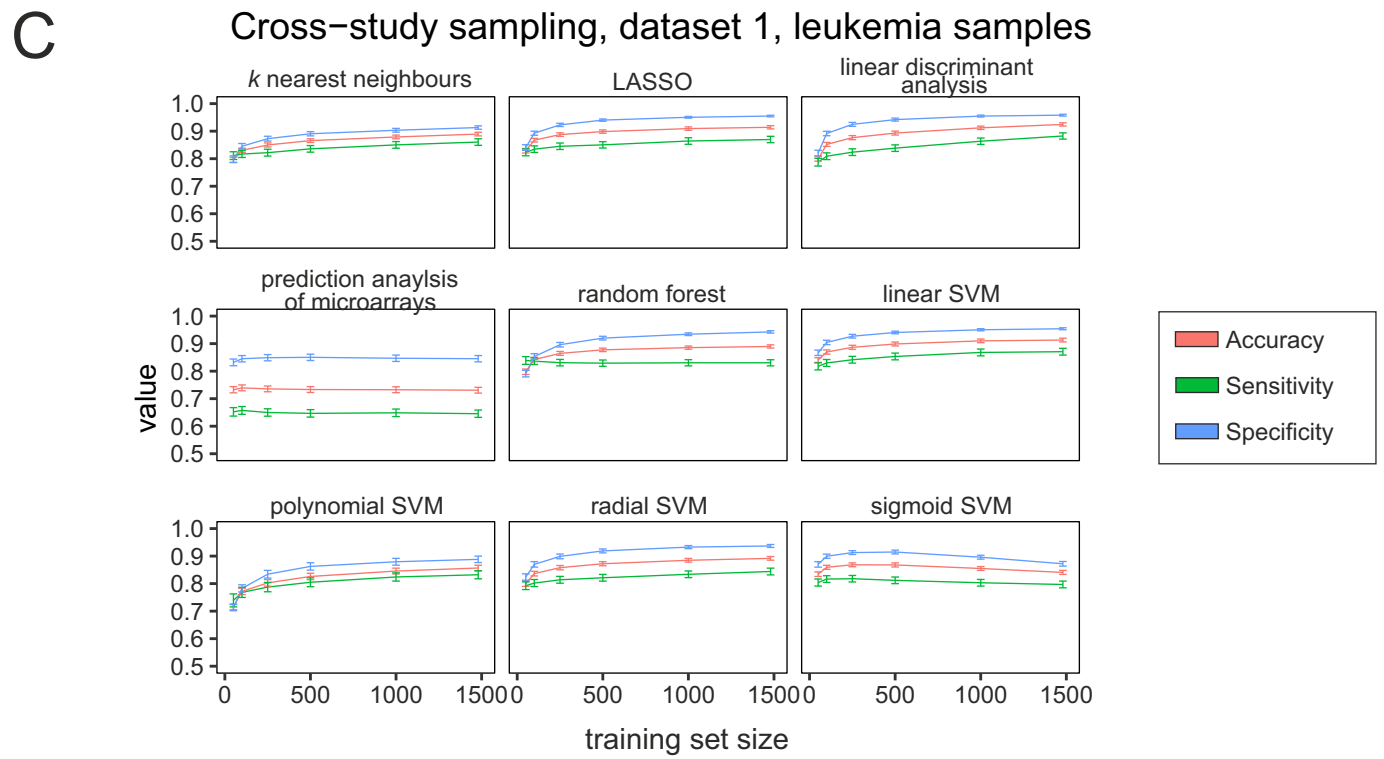
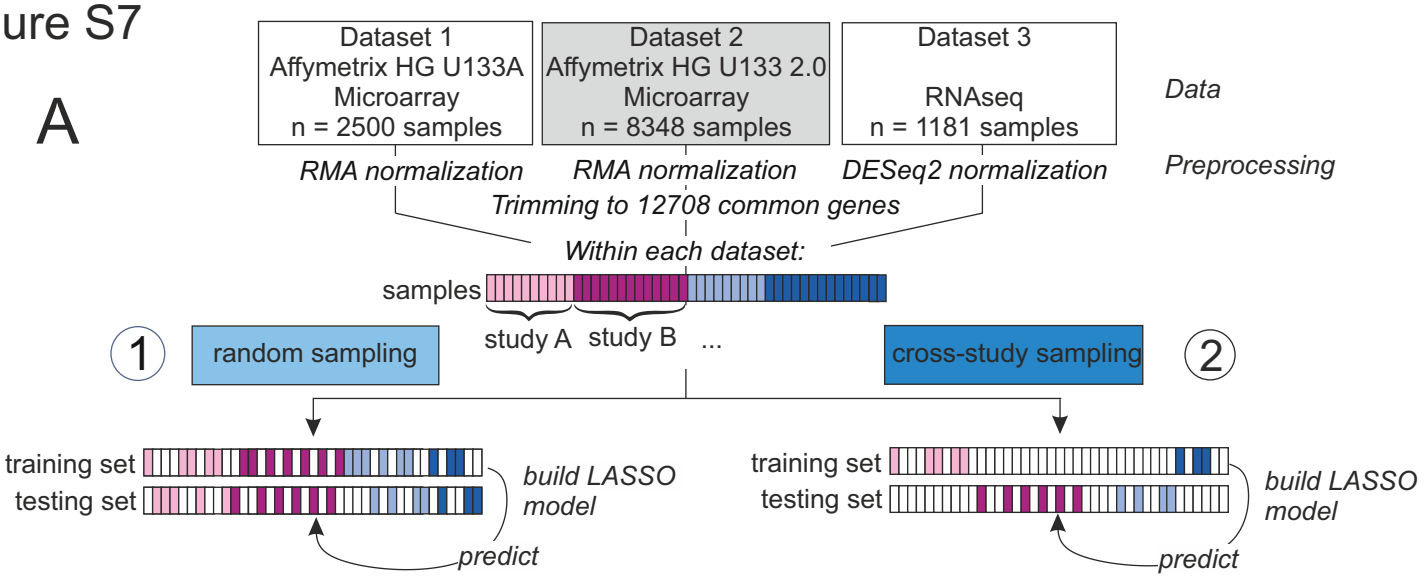
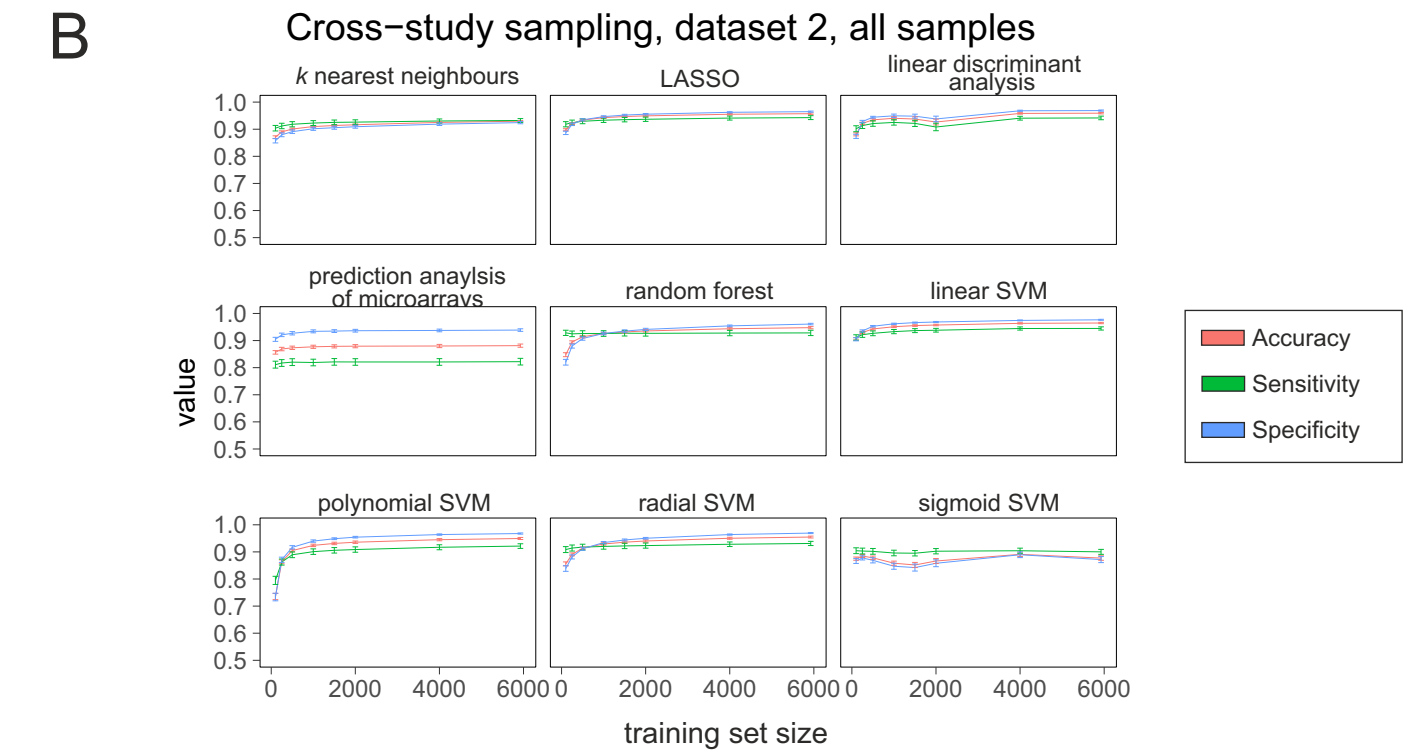

C
Cross−study sampling, dataset 1, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S7


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Cross−study sampling, dataset 2, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

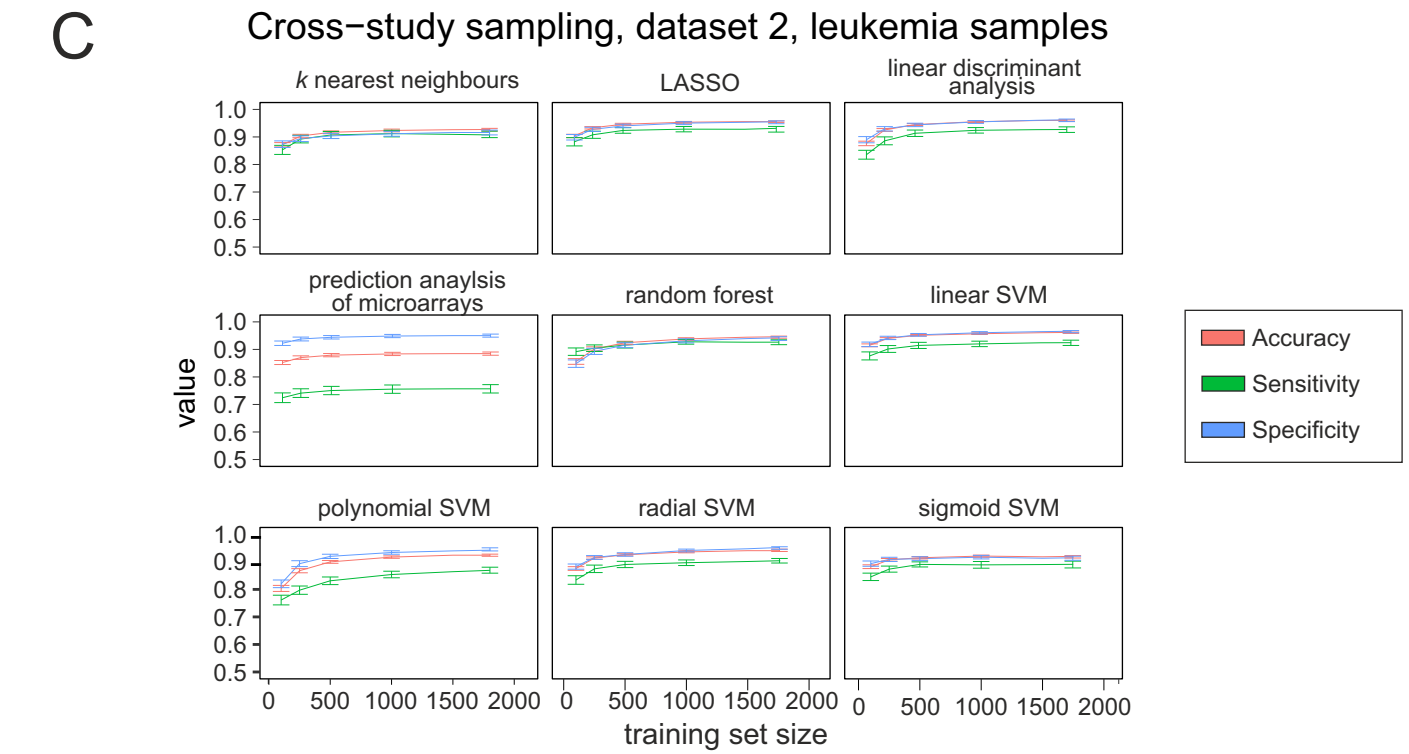
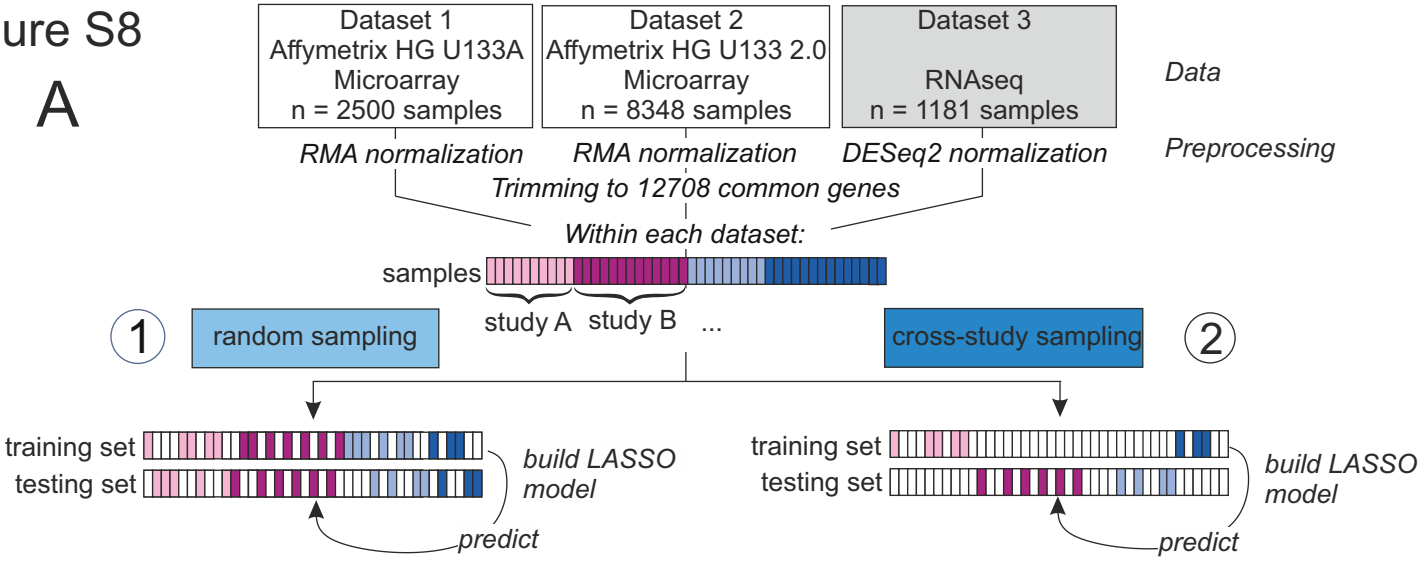

C
Cross−study sampling, dataset 2, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S8

A


Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict

B

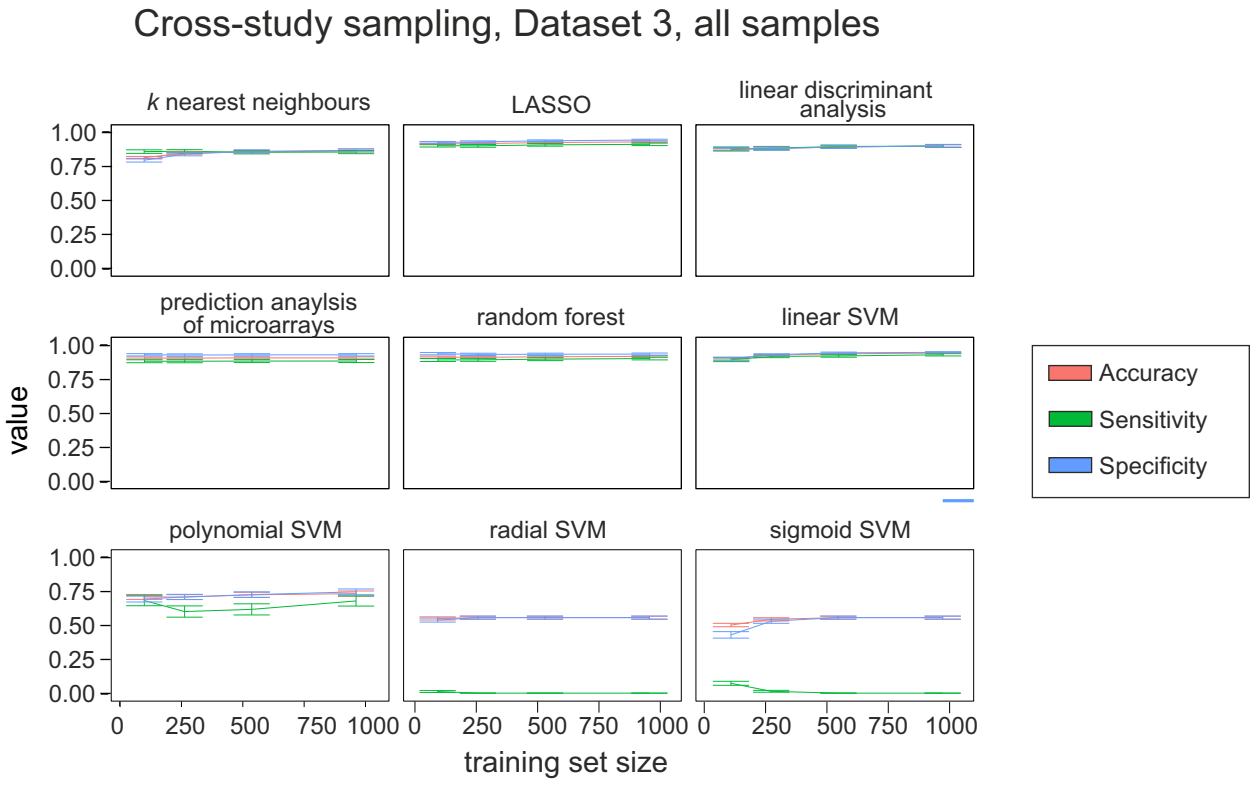
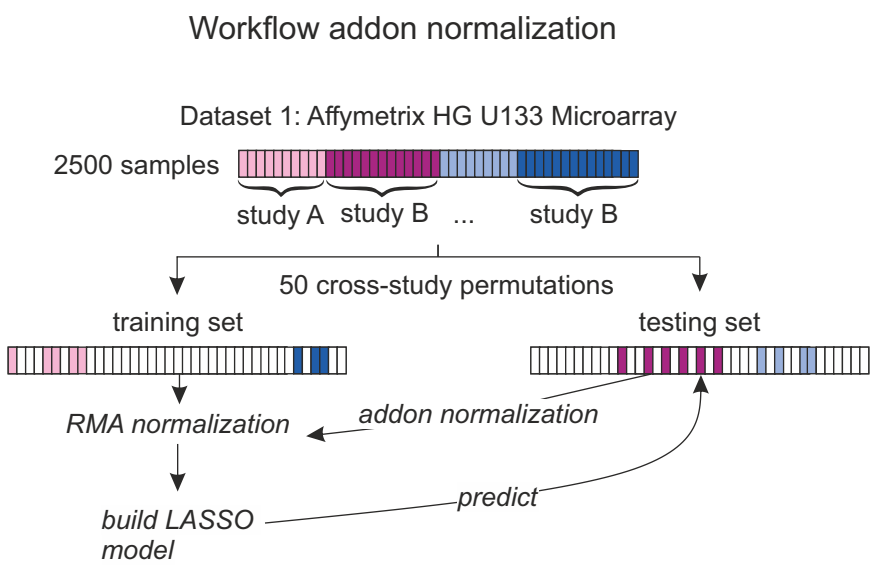

Cross-study sampling, Dataset 3, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
training set size
Accuracy
Sensitivity
Specificity

Figure S9

A


Workflow addon normalization
Dataset 1: Affymetrix HG U133 Microarray
2500 samples
study A
study B
...
study B
50 cross-study permutations
training set
testing set
RMA normalization
addon normalization
build LASSO model
predict

B

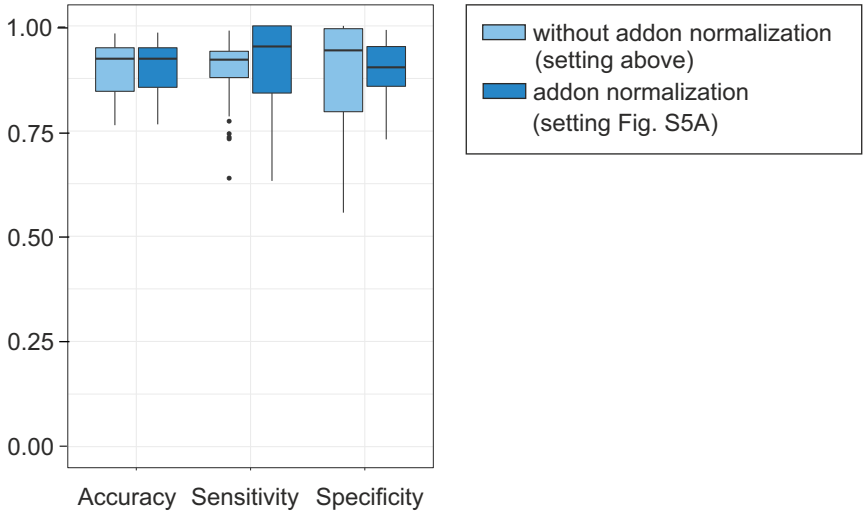
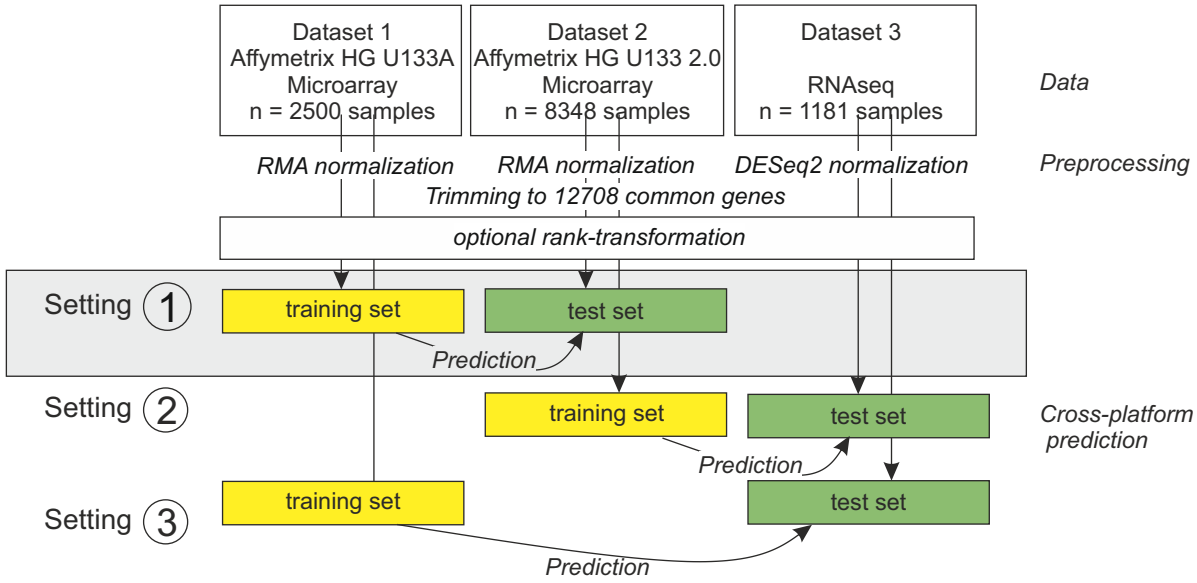
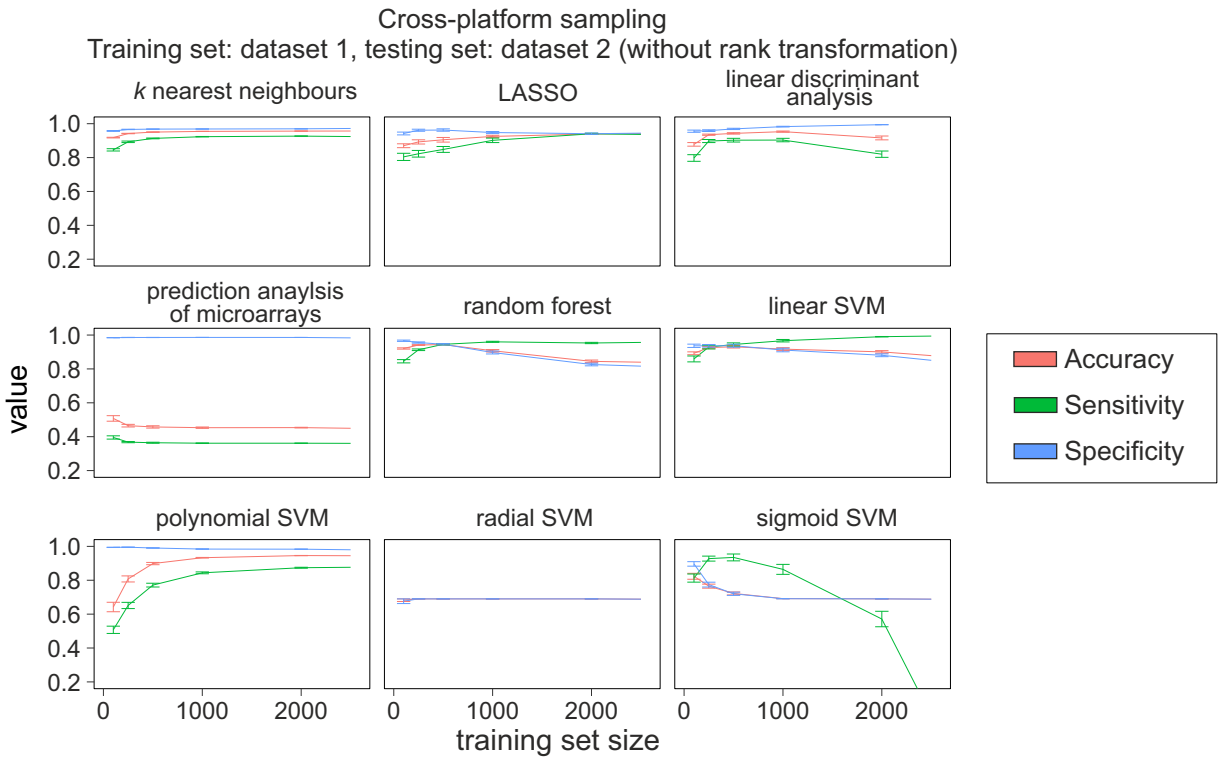

1.00
0.75
0.50
0.25
0.00
Accuracy
Sensitivity
Specificity
without addon normalization (setting above)
addon normalization (setting Fig. S5A)

Figure S10

A



B



C

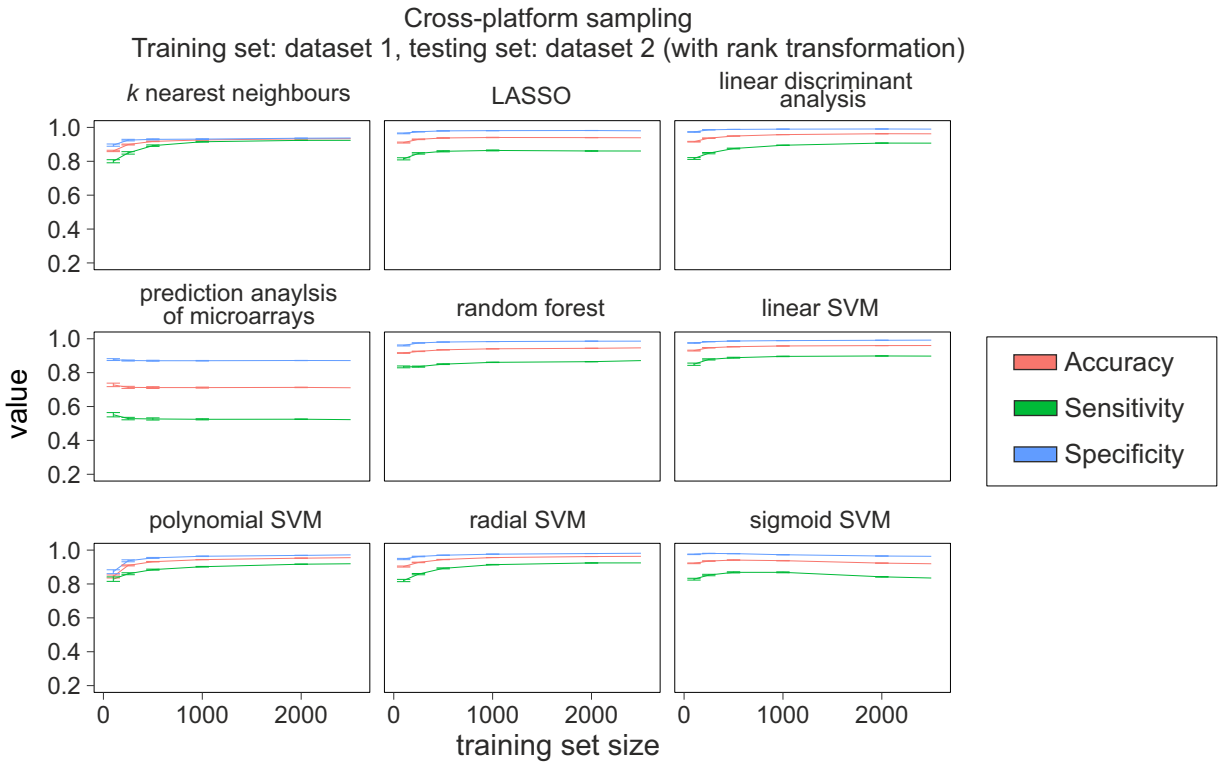
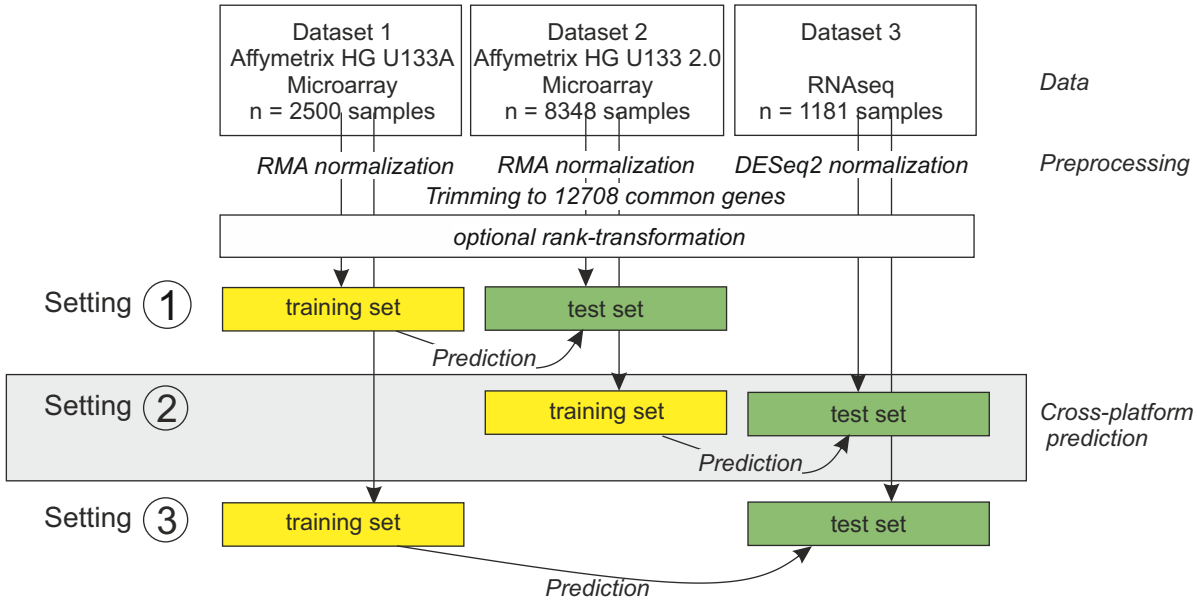
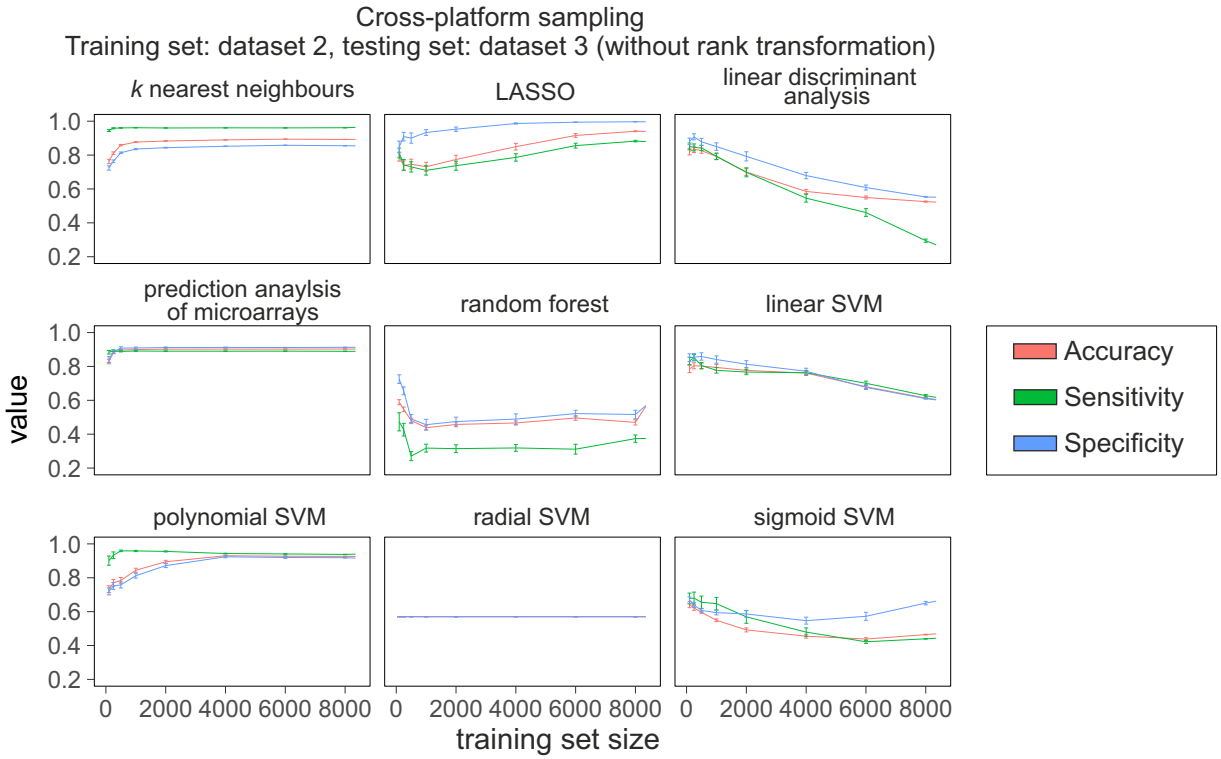


Figure S11

A



B



C

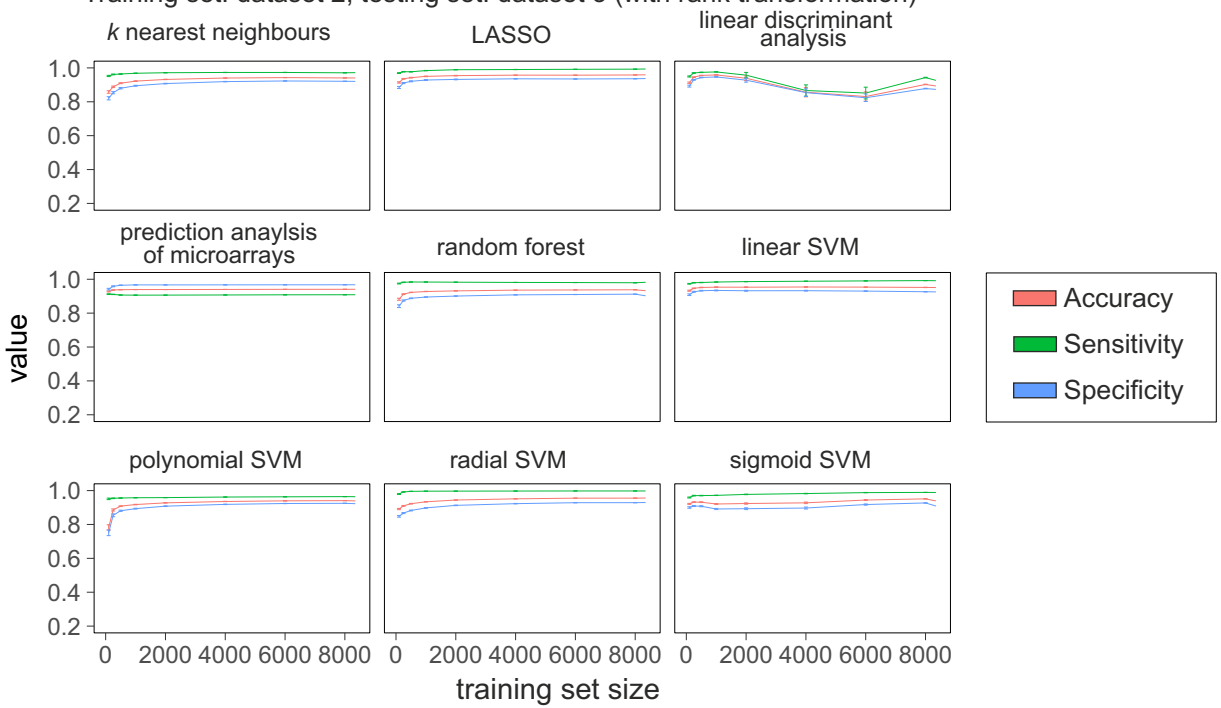
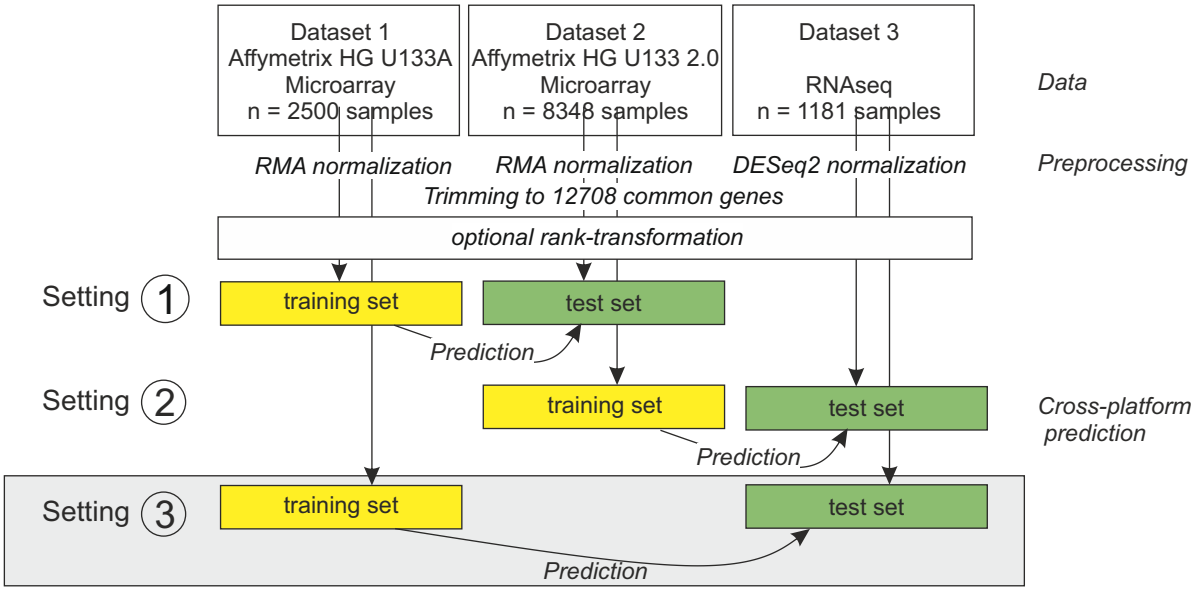
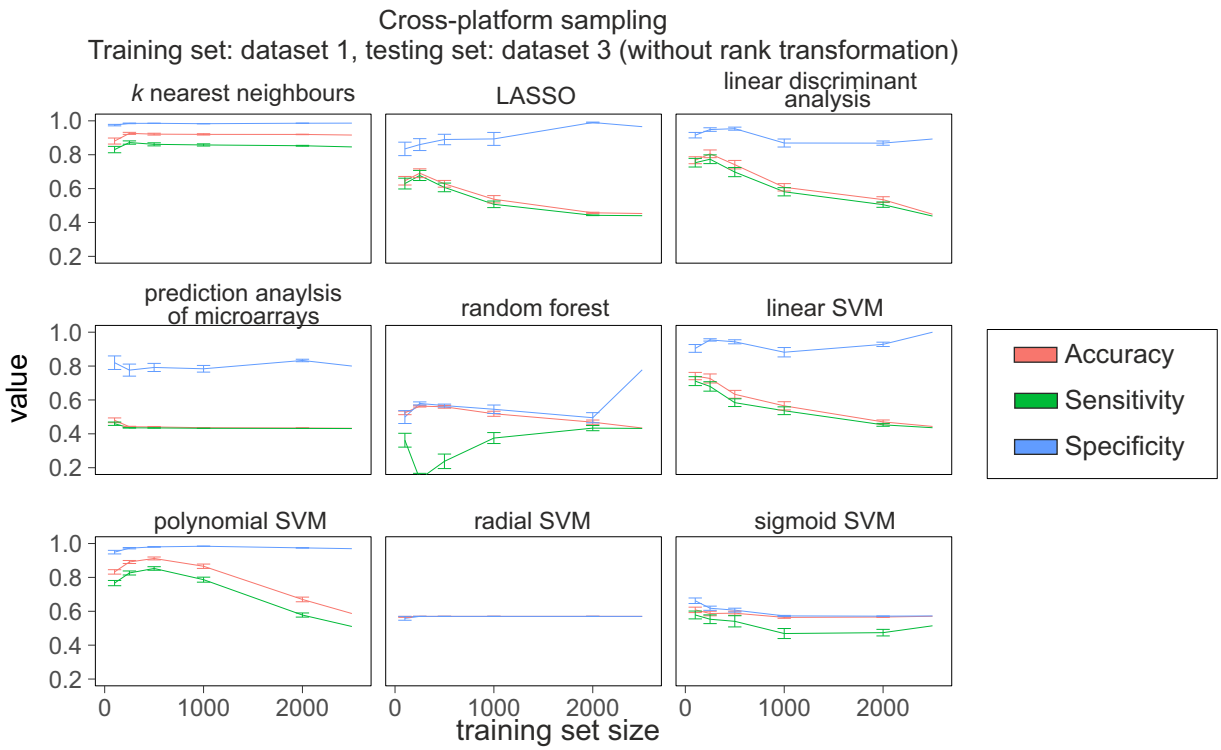


Figure S12

A



B



C

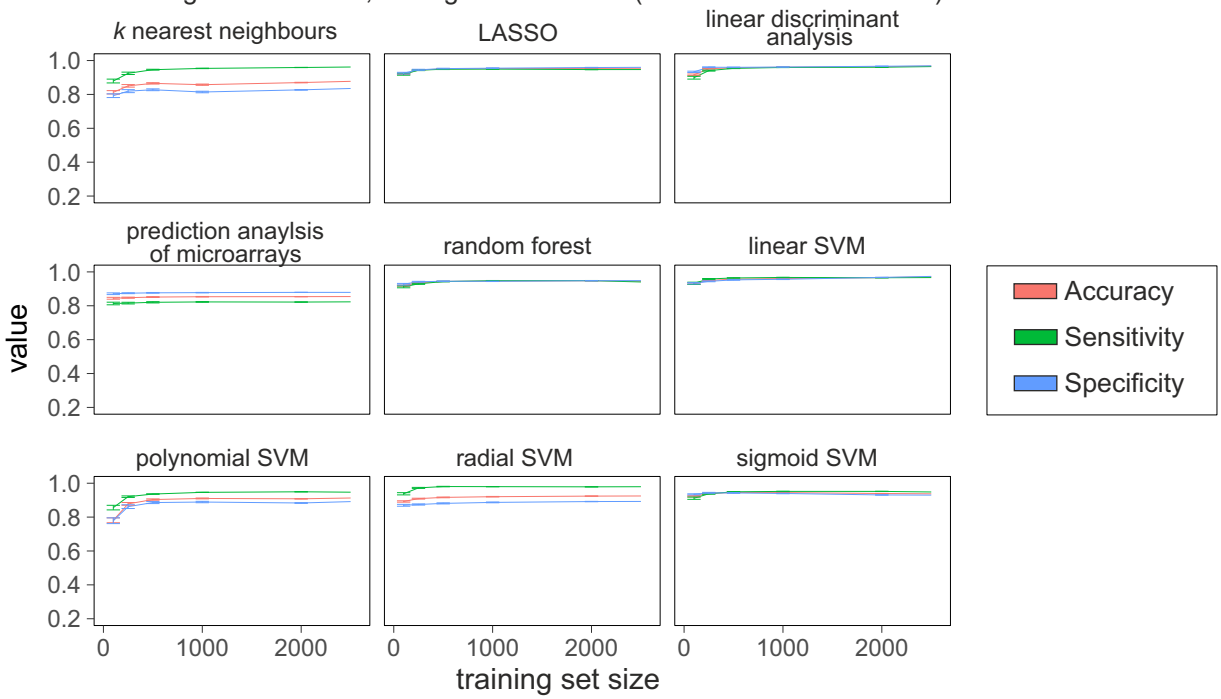
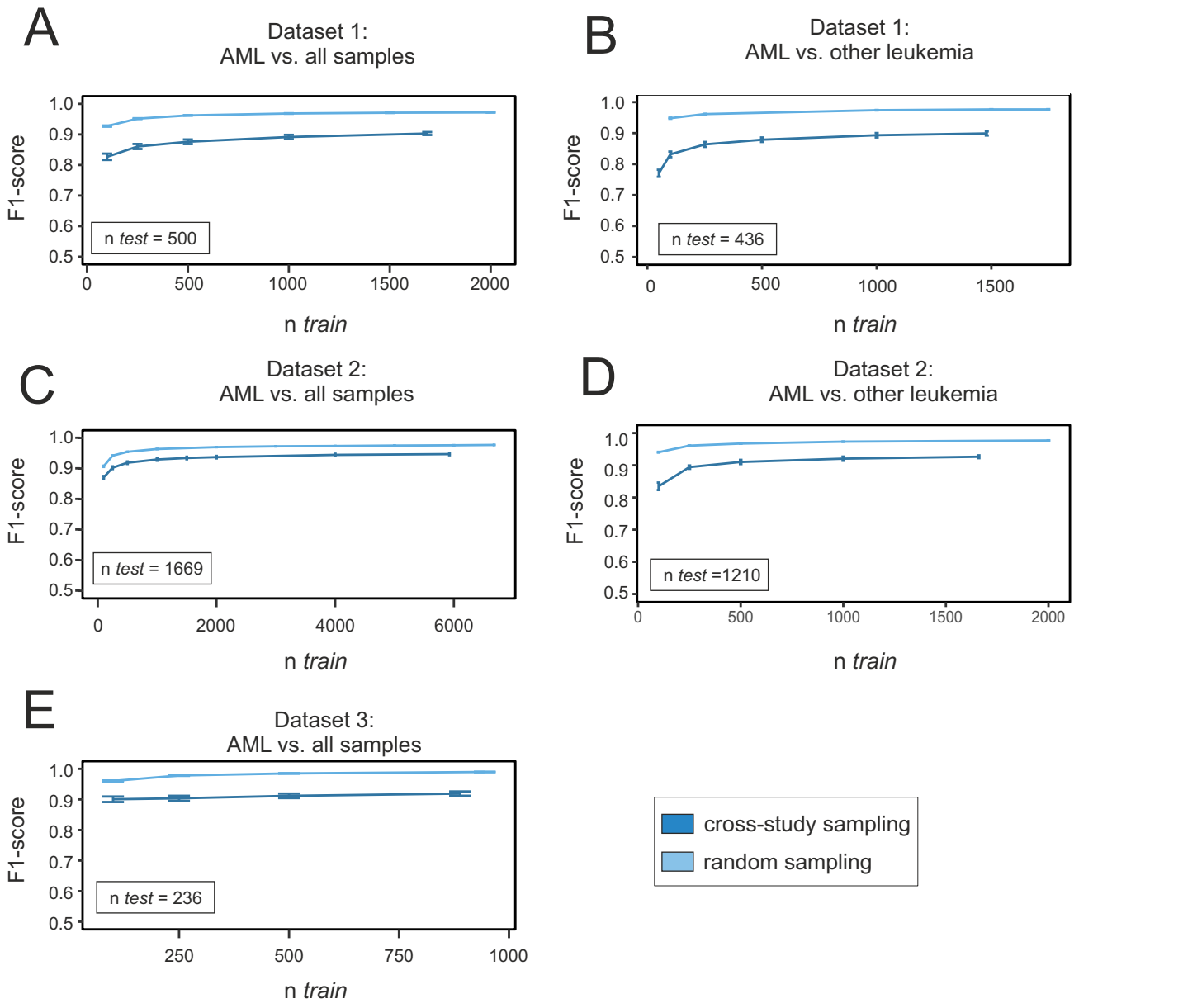


Figure S13


A
Dataset 1:
AML vs. all samples
F1-score
n test = 500
n train
B
Dataset 1:
AML vs. other leukemia
F1-score
n test = 436
n train
C
Dataset 2:
AML vs. all samples
F1-score
n test = 1669
n train
D
Dataset 2:
AML vs. other leukemia
F1-score
n test =1210
n train
E
Dataset 3:
AML vs. all samples
F1-score
n test = 236
n train
cross-study sampling
random sampling

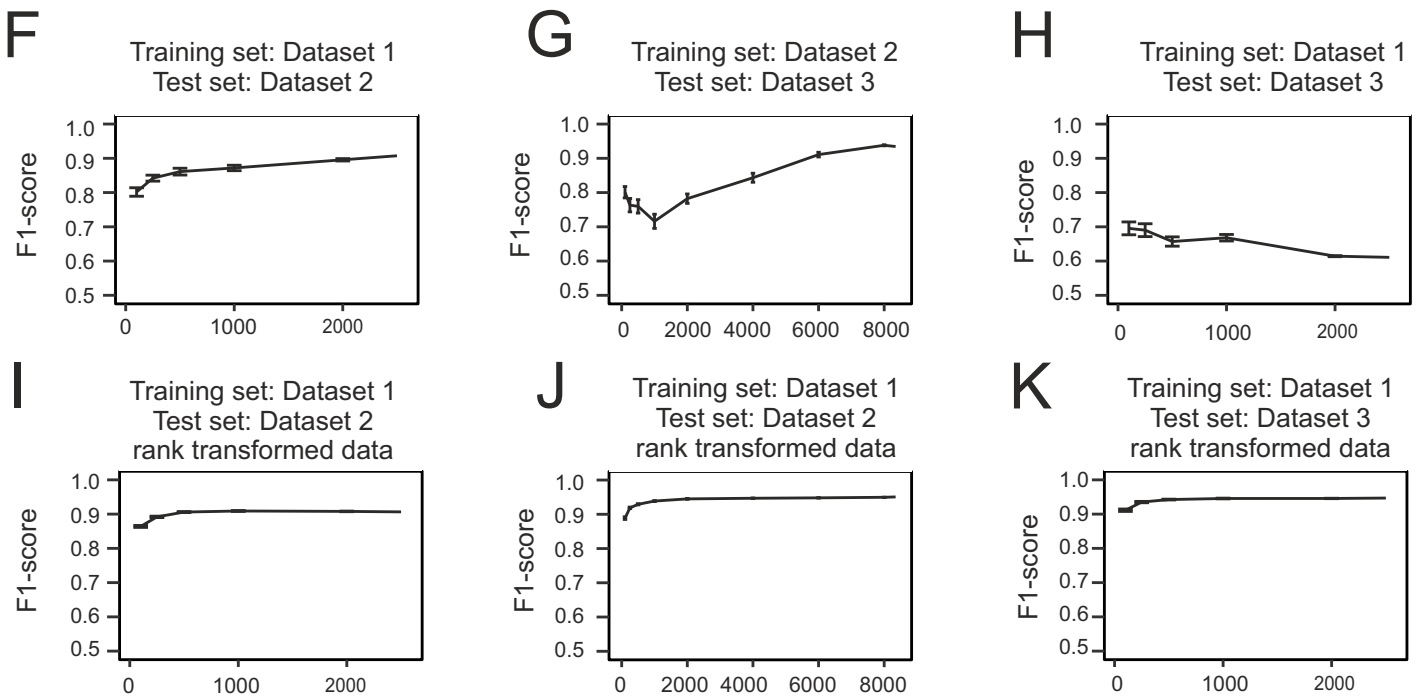
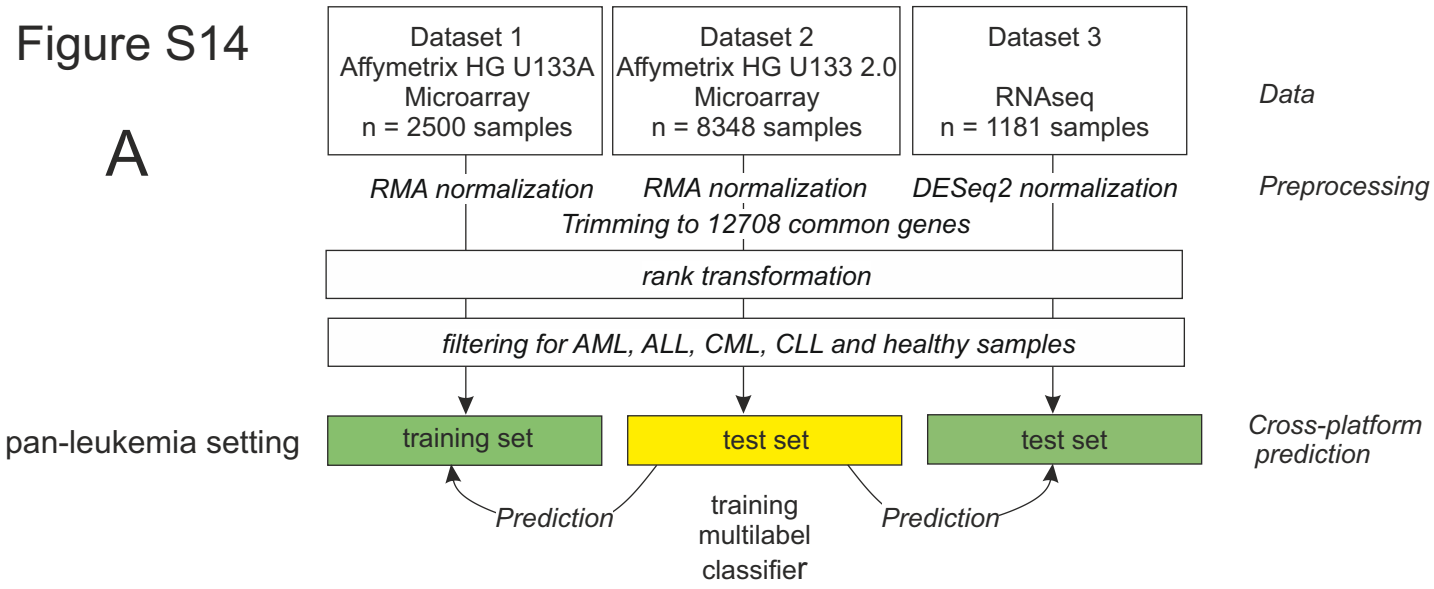
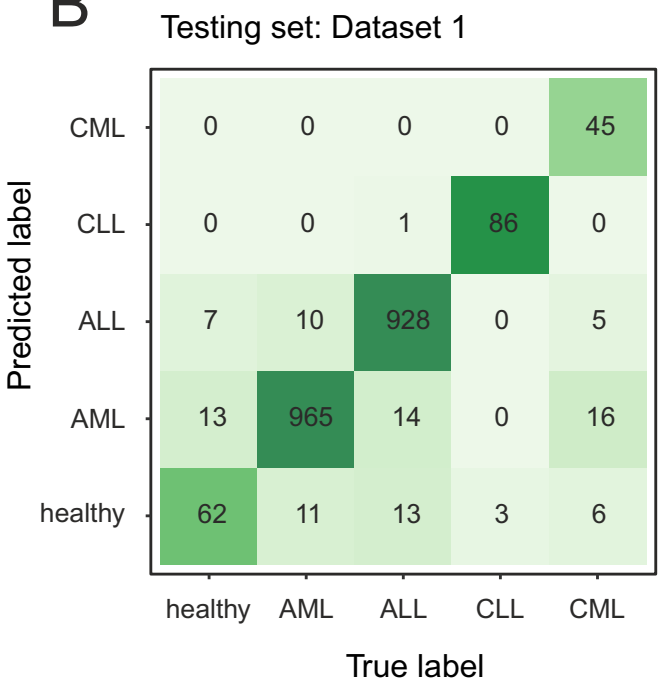

F
Training set: Dataset 1
Test set: Dataset 2
G
Training set: Dataset 2
Test set: Dataset 3
H
Training set: Dataset 1
Test set: Dataset 3
I
Training set: Dataset 1
Test set: Dataset 2
rank transformed data
J
Training set: Dataset 1
Test set: Dataset 2
rank transformed data
K
Training set: Dataset 1
Test set: Dataset 3
rank transformed data
F1-score

Figure S14

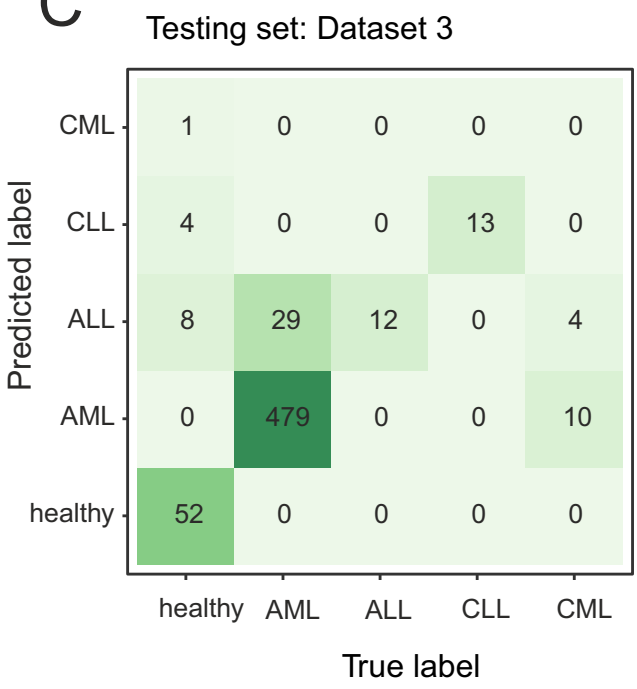
A



B



C



D

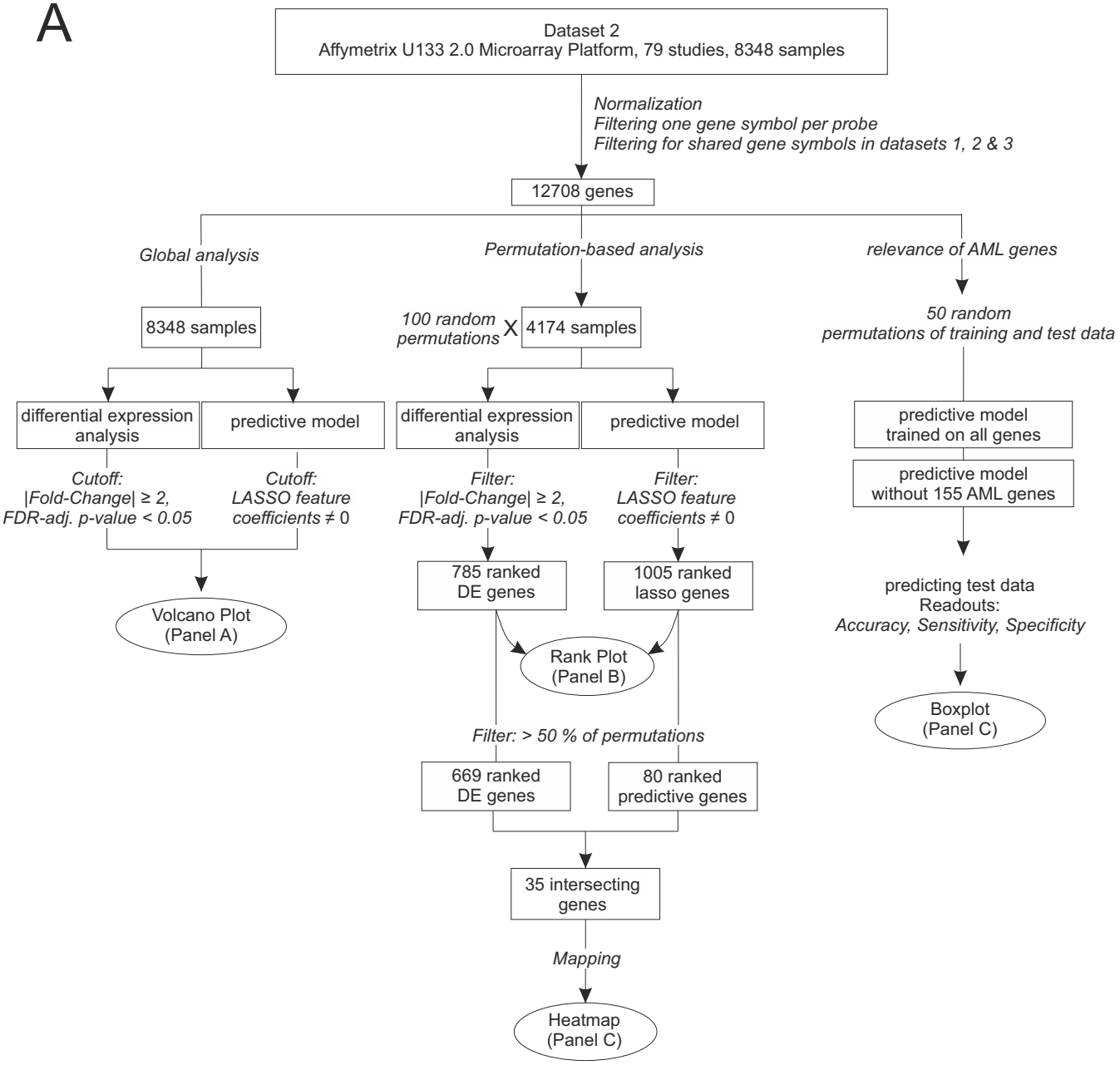
	healthy	AML	ALL	CLL	CML
bal. Accuracy	0.87	0.97	0.98	0.98	0.81
Sensitivity	0.76	0.98	0.97	0.97	0.63
Specificity	0.98	0.96	0.98	>0.99	0.63

E

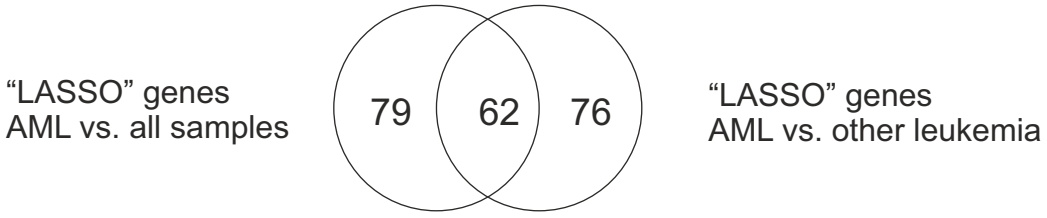
	healthy	AML	ALL	CLL	CML
bal. Accuracy	0.90	0.92	0.97	0.99	0.49
Sensitivity	0.80	0.94	1.00	1.00	0
Specificity	1.00	0.90	0.93	0.99	0.99

Figure S15

A



B



Supplemental Figure Legends

Figure S1: Sample overview, related to Figure 2

(**A**) Overview of the sample and study composition of all three datasets. The GSE number and the number of samples per disease are depicted for each study.

Figure S2: Comparison of bone marrow and PBMC samples, related to Figure 1

(**A**) Workflow: Dataset 2 was used to sample bone marrow and PBMC samples of AML patients and controls in equal numbers. **(B)** The resulting dataset of 332 samples was scaled and gene expression values of the top 25% variable genes were clustered and shown in a dendrogram.

Figure S3: Prediction of AML in random sampling scenarios (dataset 1), related to Figure 2

(**A**) Workflow: Dataset 1 (Affymetrix HG-U133 A) was RMA normalized and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 2000 samples and test data of n_{test} = 500 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 1. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS and down syndrome transient myeloproliferative disorder). Errorbars depict the standard deviation.

Figure S4: Prediction of AML in random sampling scenarios (dataset 2), related to Figure 2

(**A**) Workflow: Dataset 2 (Affymetrix HG-U133 2.0) was RMA normalized and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 6679 samples and test data of n_{test} = 1669 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 2. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS). Errorbars depict the standard deviation.

Figure S5: Prediction of AML in random sampling scenarios (dataset 3), related to Figure 2

(**A**) Workflow: Dataset 3 (RNA-seq) was normalized using DESeq2 and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 945 samples and test data of n_{test} = 236 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 3. Prediction of leukemia samples only was not possible due to small sample sizes (see Figure S1). Errorbars depict the standard deviation.

Figure S6: Prediction of AML in cross-study sampling scenarios (dataset 1), related to Figure 2

(**A**) Workflow: Dataset 1 (Affymetrix HG-U133 A) was RMA normalized and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 1, with training data samples from n_{train} = 100 to n_{train} = 1865 (mean) samples and test data of n_{test} = 500 samples. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling of leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS and down syndrome transient myeloproliferative disorder), with training data samples from n_{train} = 100 to n_{train} = 1480 (mean) samples and test data of n_{test} = 436 samples. Errorbars depict the standard deviation.

Figure S7: Effective prediction of AML in cross-study sampling scenarios (dataset 2), related to Figure 2

(**A**) Workflow: Dataset 2 (Affymetrix HG-U133 2.0) was RMA normalized and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 1, with training data samples from n_{train} = 100 to n_{train} = 5926 (mean) samples and test data of n_{test} = 1669 samples. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling of leukemia samples of dataset 1 (AML, ALL, CML, CLL and MDS), with training data samples from n_{train} = 100 to n_{train} = 1750 (mean) samples and test data of n_{test} = 1210 samples. Errorbars depict the standard deviation.

Figure S8: Effective prediction of AML in cross-study sampling scenarios (dataset 3), related to Figure 2

(**A**) Workflow: Dataset 3 (RNA-seq) was normalized using DESeq2 and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 3, with training data samples from n_{train} = 100 to n_{train} = 889 (mean) samples and test data of n_{test} = 236 samples. Prediction of leukemia samples only was not possible due to small sample sizes (see Figure S1). Errorbars depict the standard deviation.

Figure S9: Addon RMA normalization, related to Figure 2

(A) Schema for addon RMA normalization on dataset 1. The 2500 samples were subjected to 50 times cross-study sampling, which corresponds to the first 50 permutations in Figure 5SA. Different to the aforementioned approach, the data was not normalized beforehand, but after splitting the samples into training and test data. Training data was RMA-normalized and testing data was normalized “onto” the training data using addon normalization. **(B)** Accuracy, sensitivity and specificity of addon normalization as shown in **(A)** (light blue), compared to performance of the “standard” cross-study sampling approach as described in Figure 5SA.

Figure S10: Translating predictive signature across technological platforms (setting 1), related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12,708 common genes. The predictors were trained on subsamples of different sizes on dataset 1 and tested on all samples of dataset 2. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 2 (n_{test} = 8348). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 2 (n_{test} = 8348, rank transformed). Errorbars depict the standard deviation.

Figure S11: Translating predictive signature across technological platforms (setting 2), related to Figure 4

(A) Workflow: Datasets were normalized individually and trimmed to 12708 common genes. The predictors were trained on subsamples of different sizes on dataset 2 and tested on all samples of dataset 3. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 2 with training sample size from n_{train} = 100 to n_{train} = 8348 and tested on the full dataset 3 (n_{test} = 1181). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 2 with training sample size from n_{train} = 100 to n_{train} = 8348 and tested on the full dataset 3 (n_{test} = 1181, rank transformed). Errorbars depict the standard deviation.

Figure S12: Translating predictive signature across technological platforms (setting 3), related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12708 common genes. The predictors were trained on subsamples of different sizes on dataset 1 and tested on all samples of dataset 3. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 3 (n_{test} = 1181). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 3 (n_{test} = 1181, rank transformed). Errorbars depict the standard deviation.

Figure S13: F1 scores of AML prediction in random sampling, cross-study and cross-platform scenarios, related to Figures 2 and 5

F1 scores of prediction results in random and cross-study sampling scenarios in dataset 1, all samples (A), dataset 1, leukemia samples only (B), dataset 2, all samples (C), dataset 2, leukemia samples only (D), and dataset 3, all samples (E). F1 scores for cross-platform prediction results for the settings depicted in Figure 5. (F-K).

Figure S14: Pan-leukemia classification across platforms, related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12708 common genes and samples were filtered to include only AML, ALL, CML, CLL and healthy samples. A multilabel logistic regression model was fit on dataset 2 and then tested on the independently normalized datasets 1 and 3. **(B,C)** Confusion matrices comparing predicted labels to true labels for all tested leukemia types for testing on dataset 1 and 3, respectively. **(D,E)** Balanced accuracy, sensitivity and specificity of the multiclass prediction on dataset 1 and 3.

Figure S15: Workflow: Comparing differentially expressed and predictive genes, related to Figure 5

(**A**) Workflow to Figure 5: Dataset 2 was used to compare DE and the sparse predictive models. First, a global analysis of DE genes and lasso genes was performed and visualized in a heatmap. Second, dataset 2 was permuted and 35 genes that appeared at least 50 out of 100 times as “DE gene” or “lasso gene” were visualized in a heatmap. Third, predictive signatures were trained on all 12708 genes, with and without 155 known AML genes (genes included in DO and KEGG terms). Results were visualized in a boxplot. **(B)** Comparison of “lasso genes” of the prediction AML vs. all samples and AML vs. other leukemia samples of dataset 2 (same prediction setting as in Figures 2D, E).

Transparent Methods

Study search strategy

All data sets published in the National Center for Biotechnology Information Gene Expression Omnibus (GEO, (Edgar, 2002)) on 20 September 2017 were reviewed for inclusion in the present study. Basic criteria for inclusion were the cell type under study (human peripheral blood mononuclear cells (PMBCs) and/or bone marrow samples) as well as the species (*Homo sapiens*). Both tissues are considered equivalent in the diagnosis of AML. We compared bone marrow and PBMC samples of dataset 2 and did not identify overall differences in gene expression (Figure S2) and therefore did not differentiate between bone marrow and PBMC samples throughout the study. Furthermore, we excluded GEO SuperSeries to avoid duplicated samples (Table S1). We filtered the datasets for data generated with Affymetrix HG-U133 A microarrays, Affymetrix HG-U133 2.0 microarrays and high-throughput RNA sequencing (RNA-seq) and excluded studies with very small sample sizes (< 50 samples for microarray and < 10 samples for RNA-seq data). We then applied a disease-specific search, in which we filtered for acute myeloid leukemia, other leukemia and healthy or non-leukemia-related samples.

The results of this search strategy were then internally reviewed and data were excluded based on the following criteria: (i) exclusion of duplicated samples, (ii) exclusion of studies that sorted single cell types (e.g. T cells or B cells) prior to gene expression profiling, (iii) exclusion of studies with inaccessible data. Other than that, no studies were excluded from our analysis (see also Table S1). In addition, we included one unpublished dataset (in dataset 1). The above steps gave rise to the data referred to above as **dataset 1** (Affymetrix HG-U133 A microarrays), **dataset 2** (Affymetrix HG-U133 2.0 microarrays) and **dataset 3** (RNA-seq). The RNA-seq data contained was not filtered for any particular protocol and contained paired and well as single-end data of different sequencing depth. AML subtype annotations were taken from the respective metadata-files on GEO. Subgroups of FAB-classifications were combined to represent the major FAB class (e.g. AML M3 and AML M3v were combined to AML M3).

Pre-processing

All raw data files were downloaded from GEO. For normalization, we considered all platforms independently, meaning that normalization was performed separately for the samples in dataset 1, 2 and 3, respectively. Microarray data (datasets 1 and 2) were normalized using the robust multichip average (RMA) expression measures (Irizarry et al., 2003), as implemented in the R package affy (Gautier et al., 2004). RNA-seq data (dataset 3) was preprocessed using kallisto (Bray et al., 2016) and normalized with the R package DESeq2 using standard parameters (Love et al., 2014). In order to keep the datasets comparable, we filtered the data for genes annotated in all three datasets, which resulted in 12,708 genes. No filtering of low-expressed genes was performed. All scripts used in this study for pre-processing are provided as a docker container on Docker Hub (https://hub.docker.com/r/schultzelab/aml_classifier).

Prediction

Prior to classification, data sets were split into non-overlapping training and test data. For the comparisons of AML vs. all samples, all non-AML samples were used as controls, which would in clinical terms, reflect finding a diagnosis. For the prediction of AML vs. other leukemia, all non-AML leukemias, namely chronic myeloid leukemia (CML), acute lymphoblastic leukemia (ALL), chronic lymphoblastic leukemia (CLL), Myelodysplastic syndrome (MDS) and down syndrome transient myeloproliferative disorder were used as non-AML labels, which would be the equivalent of finding a differential diagnosis between different leukemias. All main classification tasks were performed in the programming language R (R Core Team, 2016). All main results were obtained using l_1-penalized logistic regression using the package glmnet (Friedman et al., 2010). Non-zero coefficients were extracted for feature ranking (Figure 4). The regularization parameter was set using 10-fold cross-validation (using training set data only). To assess predictive performance, accuracy, sensitivity, specificity and F1 score were calculated as well as positive predictive value (PPV) under several prevalence scenarios. For assessing the performance of support vector machines (SVMs), we used the R package e1071 for SVMs (linear, radial, polynomial and sigmoid kernels) (Meyer et al., 2015). The R package randomForest was used for random forest classification (Shi et al., 2004). K nearest neighbors classification was done using the knn function implemented in the class package in R (Venables and Ripley, 2002). Linear discriminant analysis was performed with the lda function implemented in the R package MASS (Venables and Ripley, 2002). For RNA-seq data, features with zero variance were excluded for LDA. Prediction analysis of microarrays was done with the pamr package

(Hastie et al., 2014). Neural networks were built using Keras (Chollet et al., 2017) with a Tensorflow backend (10 layers, ~7×10^6 parameters). Unless otherwise noted, default settings were used for tuning parameters as implemented in the respective packages.

Rank transformation to normality

As an example of a simple data transformation that would facilitate translation between gene expression platforms, we performed a rank transformation to normality. For this, gene expression values were transformed from microarray intensities (dataset 1 & 2) or RNAseq counts to their respective ranks. This was done gene-wise, meaning all gene expression values per gene were given a rank based on ordering them from lowest to highest value. The rankings were then turned into quantiles and transformed via the inverse cumulative distribution function of the Normal distribution. This leads to all genes following the exact same distribution (that is, a standard Normal with a mean of 0 and a standard deviation of 1) across all samples (Zwiener et al., 2014).

Differential expression analysis

For differential expression analysis of dataset 2 the R package limma was used (Ritchie et al., 2015). A linear model was fit on the data with inclusion of the study as a factor. Differentially expressed genes were called using an FDR-corrected p-value < 0.05 and a minimum fold change of +/- 2. For the permutation-based approach, 4174 samples were randomly drawn 100 times from the dataset. In each subset, DE genes were called as before, but without correcting for any batch in the model. The number of times each gene was called was summed up over all 100 permutations. Genes were ranked according to their overall DE count.

In addition to that a l_1-penalized logistic regression was performed using the package glmnet (Friedman et al., 2010) on the whole dataset and on each of the permutations. Genes were called to be of predictive importance if features had non-zero coefficients. The number of times each feature was of predictive importance was summed up, which resulted in a feature ranking of all “lasso genes”.

Hierarchical Clustering

35 genes which had a stability of > 50% over 100 permutations for lasso and DE genes were visualized using the R package pheatmap (Kolde, 2015) (Figure 6B). The data was z-scaled and columns clustered according to Euclidean distance. Rows were ordered according to diseases. Two gene clusters were visualized.

Exclusion of gene sets from prediction

In order to evaluate the robustness of our classification results (Figure 6C), we excluded 155 genes present in either the KEGG or the disease ontology term “Acute Myeloid Leukemia” and compared this to the results achieved when all 12078 genes of the dataset are included (random sampling, dataset 2).

Supplemental references

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Chollet, F., Allaire, J.J., and others (2017). R Interface to Keras.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2014). pamr: Pam: prediction analysis for microarrays.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Kolde, R. (2015). pheatmap: Pretty Heatmaps.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

R Core Team (2016). R: A Language and Environment for Statistical Computing.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., and Horvath, S. (2004). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol *18*, 547–557.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S (Springer).

Article

Swarm Learning for decentralized and confidential clinical machine learning

https://doi.org/10.1038/s41586-021-03583-3

Received: 3 July 2020

Accepted: 26 April 2021

Published online: 26 May 2021

Open access

Check for updates

Stefanie Warnat-Herresthal[1,2,127], Hartmut Schultze[3,127], Krishnaprasad Lingadahalli Shastry[3,127], Sathyanarayanan Manamohan[3,127], Saikat Mukherjee[3,127], Vishesh Garg[3,4,127], Ravi Sarveswara[3,127], Kristian Händler[1,5,127], Peter Pickkers[6,127], N. Ahmad Aziz[7,8,127], Sofia Ktena[9,127], Florian Tran[10,11], Michael Bitzer[12], Stephan Ossowski[13,14], Nicolas Casadei[13,14], Christian Herr[15], Daniel Petersheim[16], Uta Behrends[17], Fabian Kern[18], Tobias Fehlmann[18], Philipp Schommers[19], Clara Lehmann[19,20,21], Max Augustin[19,20,21], Jan Rybniker[19,20,21], Janine Altmüller[22], Neha Mishra[11], Joana P. Bernardes[11], Benjamin Krämer[23], Lorenzo Bonaguro[1,2], Jonas Schulte-Schrepping[1,2], Elena De Domenico[1,5], Christian Siever[3], Michael Kraut[1,5], Milind Desai[3], Bruno Monnet[3], Maria Saridaki[9], Charles Martin Siegel[3], Anna Drews[1,5], Melanie Nuesch-Germano[1,2], Heidi Theis[1,5], Jan Heyckendorf[23], Stefan Schreiber[10], Sarah Kim-Hellmuth[16], COVID-19 Aachen Study (COVAS)*, Jacob Nattermann[24,25], Dirk Skowasch[26], Ingo Kurth[27], Andreas Keller[18,28], Robert Bals[15], Peter Nürnberg[22], Olaf Rieß[13,14], Philip Rosenstiel[11], Mihai G. Netea[29,30], Fabian Theis[31], Sach Mukherjee[32], Michael Backes[33], Anna C. Aschenbrenner[1,2,5,29], Thomas Ulas[1,2], Deutsche COVID-19 Omics Initiative (DeCOI)*, Monique M. B. Breteler[7,34,128], Evangelos J. Giamarellos-Bourboulis[9,128], Matthijs Kox[6,128], Matthias Becker[1,5,128], Sorin Cheran[3,128], Michael S. Woodacre[3,128], Eng Lim Goh[3,128] & Joachim L. Schultze[1,2,5,128] ✉

Fast and reliable detection of patients with severe and heterogeneous illnesses is a major goal of precision medicine[1,2]. Patients with leukaemia can be identified using machine learning on the basis of their blood transcriptomes[3]. However, there is an increasing divide between what is technically possible and what is allowed, because of privacy legislation[4,5]. Here, to facilitate the integration of any medical data from any data owner worldwide without violating privacy laws, we introduce Swarm Learning—a decentralized machine-learning approach that unites edge computing, blockchain-based peer-to-peer networking and coordination while maintaining confidentiality without the need for a central coordinator, thereby going beyond federated learning. To illustrate the feasibility of using Swarm Learning to develop disease classifiers using distributed data, we chose four use cases of heterogeneous diseases (COVID-19, tuberculosis, leukaemia and lung pathologies). With more than 16,400 blood transcriptomes derived from 127 clinical studies with non-uniform distributions of cases and controls and substantial study biases, as well as more than 95,000 chest X-ray images, we show that Swarm Learning classifiers outperform those developed at individual sites. In addition, Swarm Learning completely fulfils local confidentiality regulations by design. We believe that this approach will notably accelerate the introduction of precision medicine.

Identification of patients with life-threatening diseases, such as leukaemias, tuberculosis or COVID-19[6,7], is an important goal of precision medicine[2]. The measurement of molecular phenotypes using 'omics' technologies[1] and the application of artificial intelligence (AI) approaches[4,8] will lead to the use of large-scale data for diagnostic purposes. Yet, there is an increasing divide between what is technically possible and what is allowed because of privacy legislation[5,9,10]. Particularly in a global crisis[6,7], reliable, fast, secure, confidentiality- and privacy-preserving AI solutions can facilitate answering important questions in the fight against such threats[11–13]. AI-based concepts range from drug target prediction[14] to diagnostic software[15,16]. At the same time, we need to consider important standards relating to data privacy and protection, such as Convention 108+ of the Council of Europe[17].

AI-based solutions rely intrinsically on appropriate algorithms[18], but even more so on large training datasets[19]. As medicine is inherently decentral, the volume of local data is often insufficient to train reliable classifiers[20,21]. As a consequence, centralization of data is one model that has been used to address the local limitations[22]. While beneficial from an AI perspective, centralized solutions have inherent disadvantages, including increased data traffic and concerns about data ownership, confidentiality, privacy, security and the creation of data monopolies that favour data aggregators[19]. Consequently, solutions

A list of affiliations appears at the end of the paper.

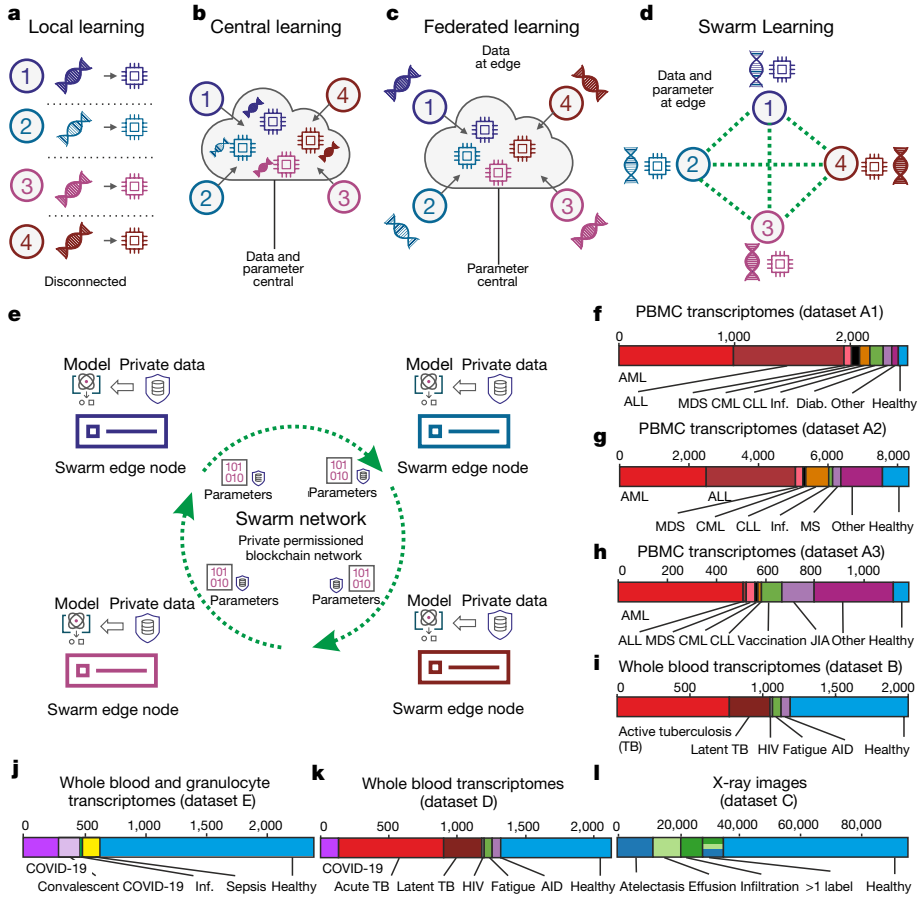


Fig. 1 | Concept of Swarm Learning. a, Illustration of the concept of local learning with data and computation at different, disconnected locations. **b**, Principle of cloud-based machine learning. **c**, Federated learning, with data being kept with the data contributor and computing performed at the site of local data storage and availability, but parameter settings orchestrated by a central parameter server. **d**, Principle of SL without the need for a central custodian. **e**, Schematic of the Swarm network, consisting of Swarm edge nodes that exchange parameters for learning, which is implemented using blockchain technology. Private data are used at each node together with the model provided by the Swarm network. **f**–**l**, Descriptions of the transcriptome datasets used. **f**, **g**, Datasets A1 (**f**; $n = 2{,}500$) and A2 (**g**; $n = 8{,}348$): two microarray-based transcriptome datasets of PBMCs. **h**, Dataset A3: 1,181 RNA-seq-based transcriptomes of PBMCs. **i**, Dataset B: 1,999 RNA-seq-based whole blood transcriptomes. **j**, Dataset E: 2,400 RNA-seq-based whole blood and granulocyte transcriptomes. **k**, Dataset D: 2,143 RNA-seq-based whole blood transcriptomes. **l**, Dataset C: 95,831 X-ray images. CML, chronic myeloid leukaemia; CLL, chronic lymphocytic leukaemia; Inf., infections; Diab., type II diabetes; MDS, myelodysplastic syndrome; MS, multiple sclerosis; JIA, juvenile idiopathic arthritis; TB, tuberculosis; HIV, human immunodeficiency virus; AID, autoimmune disease.

to the challenges of central AI models must be effective, accurate and efficient; must preserve confidentiality, privacy and ethics; and must be secure and fault-tolerant by design[23,24]. Federated AI addresses some of these aspects[19,25]. Data are kept locally and local confidentiality issues are addressed[26], but model parameters are still handled by central custodians, which concentrates power. Furthermore, such star-shaped architectures decrease fault tolerance.

We hypothesized that completely decentralized AI solutions would overcome current shortcomings, and accommodate inherently decentral data structures and data privacy and security regulations in medicine. The solution (1) keeps large medical data locally with the data owner; (2) requires no exchange of raw data, thereby also reducing data traffic; (3) provides high-level data security; (4) guarantees secure, transparent and fair onboarding of decentral members of the network without the need for a central custodian; (5) allows parameter merging with equal rights for all members; and (6) protects machine learning models from attacks. Here, we introduce Swarm Learning (SL), which combines decentralized hardware infrastructures, distributed machine learning based on standardized AI engines with a permissioned blockchain to securely onboard members, to dynamically elect the leader among members, and to merge model parameters. Computation is orchestrated by an SL library (SLL) and an iterative AI learning procedure that uses decentral data (Supplementary Information).

Concept of Swarm Learning

Conceptually, if sufficient data and computer infrastructure are available locally, machine learning can be performed locally (Fig. 1a). In cloud computing, data are moved centrally so that machine learning can be carried out by centralized computing (Fig. 1b), which can substantially increase the amount of data available for training and thereby improve machine learning results[19], but poses disadvantages such as data duplication and increased data traffic as well as challenges for data privacy and security[27]. Federated computing approaches[25] have been developed, wherein dedicated parameter servers are responsible for aggregating and distributing local learning (Fig. 1c); however, a remainder of a central structure is kept.

As an alternative, we introduce SL, which dispenses with a dedicated server (Fig. 1d), shares the parameters via the Swarm network and builds the models independently on private data at the individual sites (short 'nodes' called Swarm edge nodes) (Fig. 1e). SL provides security measures to support data sovereignty, security, and confidentiality

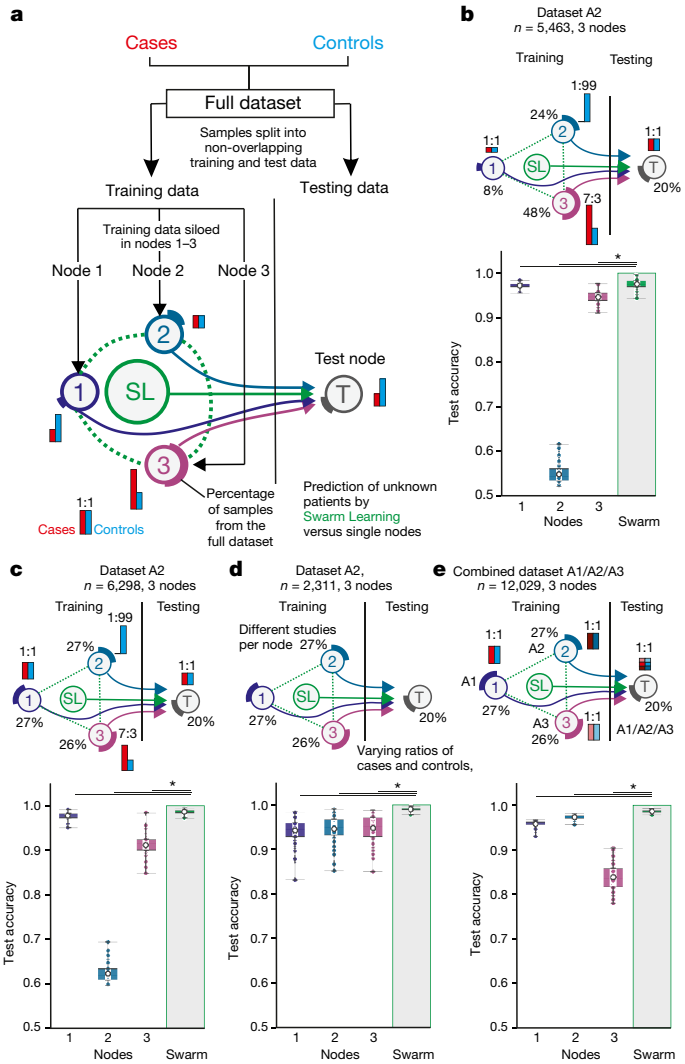


Fig. 2 | Swarm Learning to predict leukaemias from PBMC data. a, Overview of the experimental setup. Data consisting of biological replicates are split into non-overlapping training and test sets. Training data are siloed in Swarm edge nodes 1–3 and testing node T is used as independent test set. SL is achieved by integrating nodes 1–3 for training following the procedures described in the Supplementary Information. Red and blue bars illustrate the scenario-specific distribution of cases and controls among the nodes; percentages depict the percentage of samples from the full dataset. **b**, Scenario using dataset A2 with uneven distributions of cases and controls and of samples sizes among nodes. **c**, Scenario with uneven numbers of cases and controls at the different training nodes but similar numbers of samples at each node. **d**, Scenario with samples from independent studies from A2 sampled to different nodes, resulting in varying numbers of cases and controls per node. **e**, Scenario in which each node obtained samples from different transcriptomic technologies (nodes 1–3: datasets A1–A3). The test node obtained samples from each dataset A1–A3. **b**–**e**, Box plots show accuracy of 100 permutations performed for the 3 training nodes individually and for SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Accuracy is defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated using one-sided Wilcoxon signed-rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.

(Extended Data Fig. 1a) realized by private permissioned blockchain technology (Extended Data Fig. 1b). Each participant is well defined and only pre-authorized participants can execute transactions. Onboarding of new nodes is dynamic, with appropriate authorization measures to recognize network participants. A new node enrolls via a blockchain smart contract, obtains the model, and performs local model training until defined conditions for synchronization are met (Extended Data Fig. 1c). Next, model parameters are exchanged via a Swarm application programming interface (API) and merged to create an updated model with updated parameter settings before starting a new training round (Supplementary Information).

At each node, SL is divided into middleware and an application layer. The application environment contains the machine learning platform, the blockchain, and the SLL (including a containerized Swarm API to execute SL in heterogeneous hardware infrastructures), whereas the application layer contains the models (Extended Data Fig. 1d, Supplementary Information); for example, analysis of blood transcriptome data from patients with leukaemia, tuberculosis and COVID-19 (Fig. 1f–k) or radiograms (Fig. 1l). We selected both heterogeneous and life-threatening diseases to exemplify the immediate medical value of SL.

Swarm Learning predicts leukaemias

First, we used peripheral blood mononuclear cell (PBMC) transcriptomes from more than 12,000 individuals (Fig. 1f–h) in three datasets (A1–A3, comprising two types of microarray and RNA sequencing (RNA-seq))[3]. If not otherwise stated, we used sequential deep neural networks with default settings[28]. For each real-world scenario, samples were split into non-overlapping training datasets and a global test dataset[29] that was used for testing the models built at individual nodes and by SL (Fig. 2a). Within training data, samples were 'siloed' at each of the Swarm nodes in different distributions, thereby mimicking clinically relevant scenarios (Supplementary Table 1). As cases, we used samples from individuals with acute myeloid leukaemia (AML); all other samples were termed 'controls'. Each node within this simulation could stand for a medical centre, a network of hospitals, a country or any other independent organization that generates such medical data with local privacy requirements.

First, we distributed cases and controls unevenly at and between nodes (dataset A2) (Fig. 2b, Extended Data Fig. 2a, Supplementary Information), and found that SL outperformed each of the nodes (Fig. 2b). The central model performed only slightly better than SL in this scenario (Extended Data Fig. 2b). We obtained very similar results using datasets A1 and A3, which strongly supports the idea that the improvement in performance of SL is independent of data collection (clinical studies) or the technologies (microarray or RNA-seq) used for data generation (Extended Data Fig. 2c–e).

We tested five additional scenarios on datasets A1–A3: (1) using evenly distributed samples at the test nodes with case/control ratios similar to those in the first scenario (Fig. 2c, Extended Data Fig. 2f–j, Supplementary Information); (2) using evenly distributed samples, but siloing samples from particular clinical studies to dedicated training nodes and varying case/control ratios between nodes (Fig. 2d, Extended Data Fig. 3a–h, Supplementary Information); (3) increasing sample size for each training node (Extended Data Fig. 4a–f, Supplementary Information); (4) siloing samples generated with different technologies at dedicated training nodes (Fig. 2e, Extended Data Fig. 4g–i, Supplementary Information); and (5) using different RNA-seq protocols (Extended Data Fig. 4j–k, Supplementary Table 7, Supplementary Information). In all these scenarios, SL outperformed individual nodes and was either close to or equivalent to the central models.

We repeated several of the scenarios with samples from patients with acute lymphoblastic leukaemia (ALL) as cases, extended the prediction to a multi-class problem across four major types of leukaemia, extended the number of nodes to 32, tested onboarding of nodes at a later time point (Extended Data Fig. 5a–j) and replaced the deep neural network with LASSO (Extended Data Fig. 6a–c), and the results echoed the above findings (Supplementary Information).

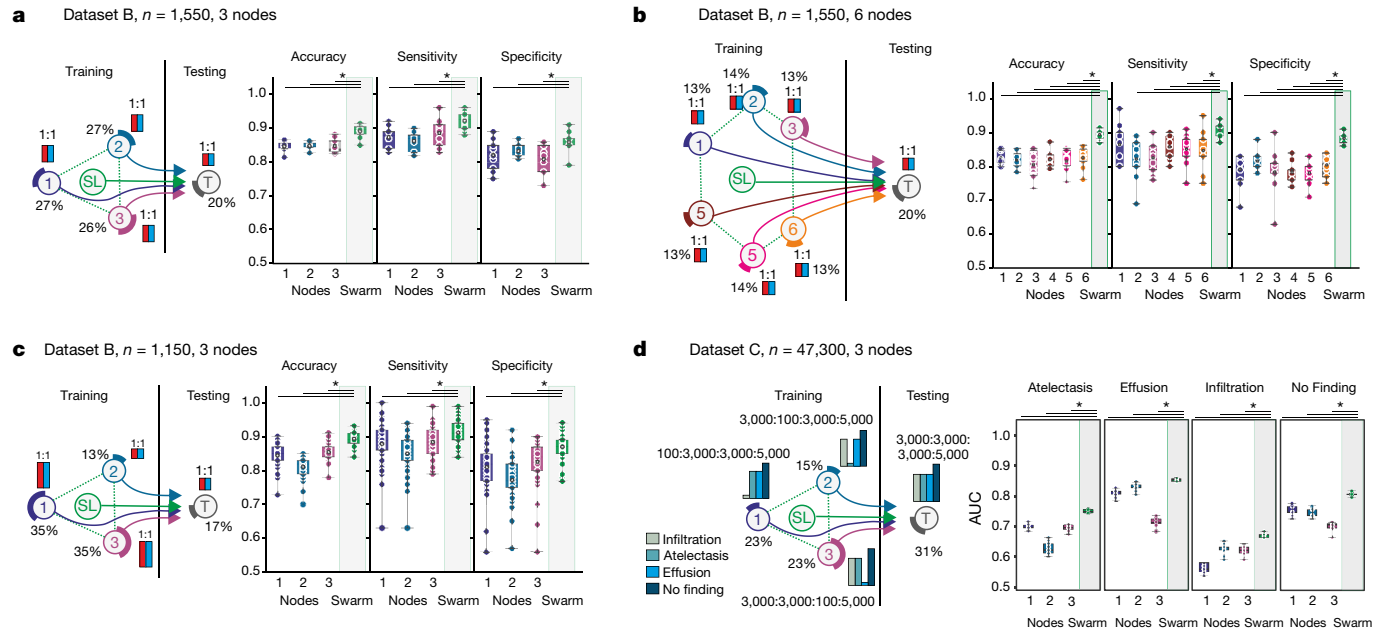


Fig. 3 | Swarm Learning to identify patients with TB or lung pathologies. **a**–**c**, Scenarios for the prediction of TB with experimental setup as in Fig. 2a. **a**, Scenario with even number of cases at each node; 10 permutations. **b**, Scenario similar to **a** but with six training nodes; 10 permutations. **c**, Scenario in which the training nodes have evenly distributed numbers of cases and controls at each training node, but node 2 has fewer samples; 50 permutations. **d**, Scenario for multilabel prediction of dataset C with uneven distribution of diseases at nodes; 10 permutations. **a**–**d**, Box plots show accuracy of all permutations for the training nodes individually and for SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Accuracy is defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.

Swarm Learning to identify tuberculosis

We built a second use case to identify patients with tuberculosis (TB) from blood transcriptomes[30,31] (Fig. 1i, Supplementary Information). First, we used all TB samples (latent and active) as cases and distributed TB cases and controls evenly among the nodes (Extended Data Fig. 7a). SL outperformed individual nodes and performed slightly better than a central model under these conditions (Extended Data Fig. 7b, Supplementary Information). Next, we predicted active TB only. Latently infected TB cases were treated as controls (Extended Data Fig. 7a) and cases and controls were kept even, but the number of training samples was reduced (Fig. 3a). Under these more challenging conditions, overall performance dropped, but SL still performed better than any of the individual nodes. When we further reduced training sample numbers by 50%, SL still outperformed the nodes, but all statistical readouts at nodes and SL showed lower performance; however, SL was still equivalent to a central model (Extended Data Fig. 7c, Supplementary Information), consistent with general observations that AI performs better when training data are increased[19]. Dividing up the training data at three nodes into six smaller nodes reduced the performance of each individual node, whereas the SL results did not deteriorate (Fig. 3b, Supplementary Information).

As TB has endemic characteristics, we used TB to simulate potential outbreak scenarios to identify the benefits and potential limitations of SL and determine how to address them (Fig. 3c, Extended Data Fig. 7d–f, Supplementary Information). The first scenario reflects a situation in which three independent regions (simulated by the nodes) would already have sufficient but different numbers of disease cases (Fig. 3c, Supplementary Information). In this scenario, the results for SL were almost comparable to those in Fig. 3a, whereas the results for node 2 (which had the smallest numbers of cases and controls) dropped noticeably. Reducing prevalence at the test node caused the node results to deteriorate, but the performance of SL was almost unaffected (Extended Data Fig. 7d, Supplementary Information).

We decreased case numbers at node 1 further, which reduced test performance for this node (Extended Data Fig. 7e), without substantially impairing SL performance. When we lowered prevalence at the test node, all performance parameters, including the F1 score (a measure of accuracy), were more resistant for SL than for individual nodes (Extended Data Fig. 7f–j).

We built a third use case for SL that addressed a multi-class prediction problem using a large publicly available dataset of chest X-rays[32] (Figs. 1l, 3d, Supplementary Information, Methods). SL outperformed each node in predicting all radiological findings included (atelectasis, effusion, infiltration and no finding), which suggests that SL is also applicable to non-transcriptomic data spaces.

Identification of COVID-19

In the fourth use case, we addressed whether SL could be used to detect individuals with COVID-19 (Fig. 1k, Supplementary Table 6). Although COVID-19 is usually detected by using PCR-based assays to detect viral RNA[33], assessing the specific host response in addition to disease prediction might be beneficial in situations for which the pathogen is unknown, specific pathogen tests are not yet possible, existing tests might produce false negative results, and blood transcriptomics can contribute to the understanding of the host's immune response[34–36].

In a first proof-of-principle study, we simulated an outbreak situation node with evenly distributed cases and controls at training nodes and test nodes (Extended Data Fig. 8a, b); this showed very high statistical performance parameters for SL and all nodes. Lowering the prevalence at test nodes reduced performance (Extended Data Fig. 8c), but F1 scores deteriorated only when we reduced prevalence further (1:44 ratio) (Extended Data Fig. 8d); even under these conditions, SL performed best. When we reduced cases at training nodes, all performance measures remained very high at the test node for SL and individual nodes (Extended Data Fig. 8e–j). When we tested outbreak scenarios

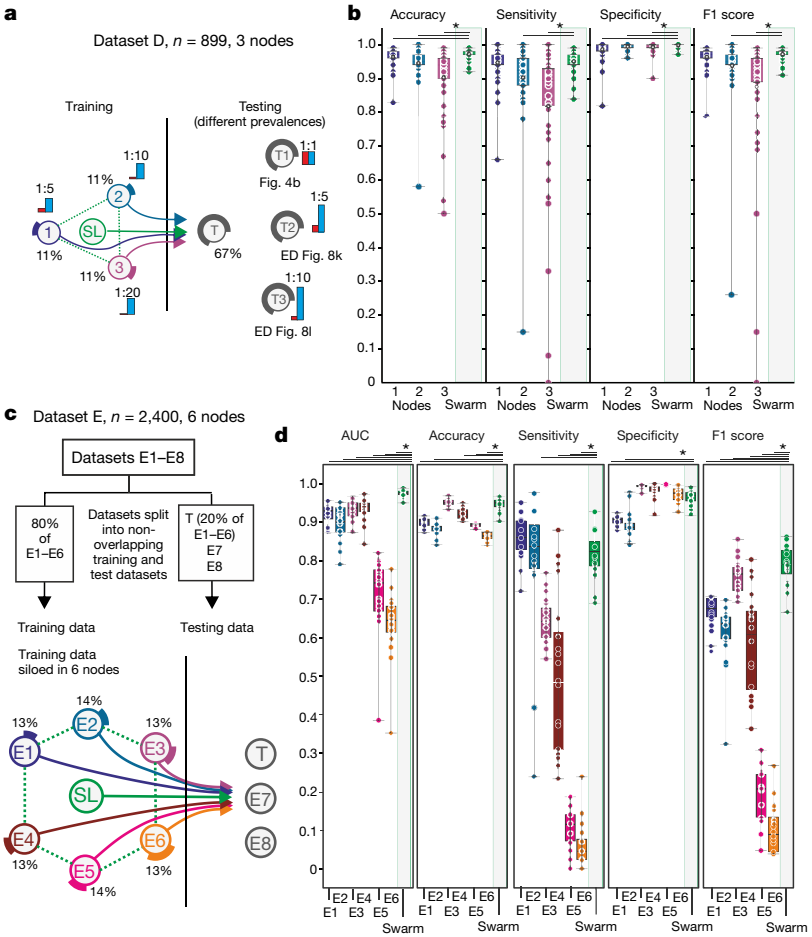


Fig. 4 | Identification of patients with COVID-19 in an outbreak scenario. a, An outbreak scenario for COVID-19 using dataset D with experimental setup as in Fig. 2a. **b**, Evaluation of **a** with even prevalence showing accuracy, sensitivity, specificity and F1 score of 50 permutations for each training node and SL, on the test node. **c**, An outbreak scenario with dataset E, particularly E1–6 with an 80:20 training:test split. Training data are distributed to six training nodes, independent test data are placed at the test node. **d**, Evaluation of **c** showing AUC, accuracy, sensitivity, specificity and F1 score of 20 permutations. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed-rank test with continuity correction; $^*P < 0.05$, all P values listed in Supplementary Table 5.

with very few cases at test nodes and varying prevalence at the independent test node (Fig. 4a), nodes 2 and 3 showed decreased performance; SL outperformed these nodes (Fig. 4b, Extended Data Fig. 8k, l) and was equivalent to the central model (Extended Data Fig. 8m). The model showed no sign of overfitting (Extended Data Fig. 8n) and comparable results were obtained when we increased the number of training nodes (Extended Data Fig. 9a–d).

We recruited further medical centres in Europe that differed in controls and distributions of age, sex, and disease severity (Supplementary Information), which yielded eight individual centre-specific sub-datasets (E1–8; Extended Data Fig. 9e).

In the first setting, centres E1–E6 teamed up and joined the Swarm network with 80% of their local data; 20% of each centre's dataset was distributed to a test node[29] (Fig. 4c) and the model was also tested on two external datasets, one with convalescent COVID-19 cases (E7) and one of granulocyte-enriched COVID-19 samples (E8). SL outperformed all nodes in terms of area under the curve (AUC) for the prediction of the global test datasets (Fig. 4d, Extended Data Fig. 9f, Supplementary Information). When looking at performance on testing samples split by centre of origin, it became clear that individual centre nodes could not have predicted samples from other centres (Extended Data Fig. 9g). By contrast, SL predicted samples from these nodes successfully. This was similarly true when we reduced the scenario, using E1, E2, and E3 as training nodes and E4 as an independent test node (Extended Data Fig. 9h).

In addition, SL can cope with biases such as sex distribution, age or co-infection bias (Extended Data Fig. 10a–c, Supplementary Information) and SL outperformed individual nodes when distinguishing mild from severe COVID-19 (Extended Data Fig. 10d, e). Collectively, we provide evidence that blood transcriptomes from COVID-19 patients represent a promising feature space for applying SL.

Discussion

With increasing efforts to enforce data privacy and security[5,9,10] and to reduce data traffic and duplication, a decentralized data model will become the preferred choice for handling, storing, managing, and analysing any kind of large medical dataset[19]. Particularly in oncology, success has been reported in machine-learning-based tumour detection[3,37], subtyping[38], and outcome prediction[39], but progress is hindered by the limited size of datasets[19], with current privacy regulations[5,9,10] making it less appealing to develop centralized AI systems. SL, as a decentralized learning system, replaces the current paradigm of centralized data sharing in cross-institutional medical research. SL's blockchain technology gives robust measures against dishonest participants or adversaries attempting to undermine a Swarm network. SL provides confidentiality-preserving machine learning by design and can inherit new developments in differential privacy algorithms[40], functional encryption[41], or encrypted transfer learning approaches[42] (Supplementary Information).

Global collaboration and data sharing are important quests[13] and both are inherent characteristics of SL, with the further advantage that data sharing is not even required and can be transformed into knowledge sharing, thereby enabling global collaboration with complete data confidentiality, particularly if using medical data. Indeed, statements by lawmakers have emphasized that privacy rules apply fully during a pandemic[43]. Particularly in such crises, AI systems need to comply with ethical principles and respect human rights[12]. Systems such as SL—allowing fair, transparent, and highly regulated shared data analytics while preserving data privacy—are to be favoured. SL should be explored for image-based diagnosis of COVID-19 from patterns in X-ray images or CT scans[15,16], structured health records[12], or data from wearables for disease tracking[12]. Collectively, SL and transcriptomics

(or other medical data) are a very promising approach to democratize the use of AI among the many stakeholders in the domain of medicine, while at the same time resulting in improved data confidentiality, privacy, and data protection, and a decrease in data traffic.

Online content

Any methods, additional references, Nature Research 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/s41586-021-03583-3.

1. Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature* **526**, 336–342 (2015).
2. Haendel, M. A., Chute, C. G. & Robinson, P. N. Classification, ontology, and precision medicine. *N. Engl. J. Med.* **379**, 1452–1462 (2018).
3. Warnat-Herresthal, S. et al. Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics. *iScience* **23**, 100780 (2020).
4. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
5. Price, W. N., II & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
6. Berlin, D. A., Gulick, R. M. & Martinez, F. J. Severe Covid-19. *N. Engl. J. Med.* **383**, 2451–2460 (2020).
7. Gandhi, R. T., Lynch, J. B. & Del Rio, C. Mild or moderate Covid-19. *N. Engl. J. Med.* **383**, 1757–1766 (2020).
8. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
9. Kels, C. G. HIPAA in the era of data sharing. *J. Am. Med. Assoc.* **323**, 476–477 (2020).
10. McCall, B. What does the GDPR mean for the medical community? *Lancet* **391**, 1249–1250 (2018).
11. Cho, A. AI systems aim to sniff out coronavirus outbreaks. *Science* **368**, 810–811 (2020).
12. Luengo-Oroz, M. et al. Artificial intelligence cooperation to support the global response to COVID-19. *Nat. Mach. Intell.* **2**, 295–297 (2020).
13. Peiffer-Smadja, N. et al. Machine learning for COVID-19 needs global collaboration and data-sharing. *Nat. Mach. Intell.* **2**, 293–294 (2020).
14. Ge, Y. et al. An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. *Signal Transduct. Target Ther.* **6**, 165 (2021).
15. Mei, X. et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**, 1224–1228 (2020).
16. Zhang, K. et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **182**, 1360 (2020).
17. Council of Europe: Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data. *Intl Legal Materials* **20**, 317–325 (1981).
18. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
19. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
20. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
21. Savage, N. Calculating disease. *Nature* **550**, S115–S117 (2017).
22. Ping, P., Hermjakob, H., Polson, J. S., Benos, P. V. & Wang, W. Biomedical informatics on the cloud: A treasure hunt for advancing cardiovascular medicine. *Circ. Res.* **122**, 1290–1301 (2018).
23. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
24. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
25. Konečný, J. et al. Federated learning: strategies for improving communication efficiency. Preprint at https://arxiv.org/abs/1610.05492 (2016).
26. Shokri, R. & Shmatikov, V. Privacy-preserving deep learning. *2015 53rd Annual Allerton Conf. Communication, Control, and Computing* 909–910 (IEEE, 2015).
27. Dove, E. S., Joly, Y., Tassé, A. M. & Knoppers, B. M. Genomic cloud computing: legal and ethical points to consider. *Eur. J. Hum. Genet.* **23**, 1271–1278 (2015).
28. Chollet, F. *Keras* https://github.com/keras-team/keras (2015).
29. Zhao, Y. et al. Federated learning with non-IID data. Preprint at https://arxiv.org/abs/1806.00582 (2018).
30. Leong, S. et al. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. *Tuberculosis* **109**, 41–51 (2018).
31. Zak, D. E. et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet* **387**, 2312–2322 (2016).
32. Wang, X. et al. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* 3462–3471 (IEEE, 2017).
33. Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* **25**, 2000045 (2020).
34. Aschenbrenner, A. C. et al. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med.* **13**, 7 (2021).
35. Chaussabel, D. Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin. Immunol.* **27**, 58–66 (2015).
36. Schulte-Schrepping, J. et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**, 1419–1440.e23 (2020).
37. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
38. Kaissis, G. et al. A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy. *PLoS One* **14**, e0218642 (2019).
39. Elshafeey, N. et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma. *Nat. Commun.* **10**, 3170 (2019).
40. Abadi, M. et al. Deep learning with differential privacy. *Proc. 2016 ACM SIGSAC Conf. Computer and Communications Security—CCS'16* 308–318 (ACM Press, 2016).
41. Ryffel, T., Dufour-Sans, E., Gay, R., Bach, F. & Pointcheval, D. Partially encrypted machine learning using functional encryption. Preprint at https://arxiv.org/abs/1905.10214 (2019).
42. Salem, M., Taheri, S. & Yuan, J.-S. Utilizing transfer learning and homomorphic encryption in a privacy preserving and secure biometric recognition system. *Computers* **8**, 3 (2018).
43. Kędzior, M. The right to data protection and the COVID-19 pandemic: the European approach. *ERA Forum* **21**, 533–543 (2021).

[1]Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany. [2]Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany. [3]Hewlett Packard Enterprise, Houston, TX, USA. [4]Mesh Dynamics, Bangalore, India. [5]PRECISE Platform for Single Cell Genomics and Epigenomics, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) and the University of Bonn, Bonn, Germany. [6]Department of Intensive Care Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands. [7]Population Health Sciences, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany. [8]Department of Neurology, Faculty of Medicine, University of Bonn, Bonn, Germany. [9]4th Department of Internal Medicine, National and Kapodistrian University of Athens, Medical School, Athens, Greece. [10]Department of Internal Medicine I, Christian-Albrechts-University and University Hospital Schleswig-Holstein, Kiel, Germany. [11]Institute of Clinical Molecular Biology, Christian-Albrechts-University and University Hospital Schleswig-Holstein, Kiel, Germany. [12]Department of Internal Medicine I, University Hospital, University of Tübingen, Tübingen, Germany. [13]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [14]NGS Competence Center Tübingen, Tübingen, Germany. [15]Department of Internal Medicine V, Saarland University Hospital, Homburg, Germany. [16]Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital LMU Munich, Munich, Germany. [17]Children's Hospital, Medical Faculty, Technical University Munich, Munich, Germany. [18]Clinical Bioinformatics, Saarland University, Saarbrücken, Germany. [19]Department I of Internal Medicine, Faculty of Medicine and University Hospital of Cologne, University of Cologne, Cologne, Germany. [20]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. [21]German Center for Infection Research (DZIF), Partner Site Bonn-Cologne, Cologne, Germany. [22]Cologne Center for Genomics, West German Genome Center, University of Cologne, Cologne, Germany. [23]Clinical Infectious Diseases, Research Center Borstel and German Center for Infection Research (DZIF), Partner Site Hamburg-Lübeck-Borstel-Riems, Borstel, Germany. [24]Department of Internal Medicine I, University Hospital Bonn, Bonn, Germany. [25]German Center for Infection Research (DZIF), Braunschweig, Germany. [26]Department of Internal Medicine II - Cardiology/Pneumology, University of Bonn, Bonn, Germany. [27]Institute of Human Genetics, Medical Faculty, RWTH Aachen University, Aachen, Germany. [28]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA. [29]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands. [30]Immunology & Metabolism, Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany. [31]Institute of Computational Biology, Helmholtz Center Munich (HMGU), Neuherberg, Germany. [32]Statistics and Machine Learning, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany. [33]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany. [34]Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), Faculty of Medicine, University of Bonn, Bonn, Germany. [127]These authors contributed equally: Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N. Ahmad Aziz, Sofia Ktena. [128]These authors jointly supervised this work: Monique M. B. Breteler, Evangelos J. Giamarellos-Bourboulis, Matthijs Kox, Matthias Becker, Sorin Cheran, Michael S. Woodacre, Eng Lim Goh, Joachim L. Schultze. *Lists of authors and their affiliations appear online. ✉e-mail: joachim.schultze@dzne.de

COVID-19 Aachen Study (COVAS)

Paul Balfanz[113], Thomas Eggermann[27], Peter Boor[114], Ralf Hausmann[115], Hannah Kuhn[116], Susanne Isfort[117], Julia Carolin Stingl[118], Günther Schmalzing[118], Christiane K. Kuhl[119], Rainer Röhrig[120], Gernot Marx[121], Stefan Uhlig[122], Edgar Dahl[123,124], Dirk Müller-Wieland[125], Michael Dreher[126] & Nikolaus Marx[125]

[113]Department of Cardiology, Angiology and Intensive Care Medicine, University Hospital RWTH Aachen, Aachen, Germany. [114]Institute of Pathology & Department of Nephrology, University Hospital RWTH Aachen, Aachen, Germany. [115]Institute of Clinical Pharmacology, University Hospital RWTH Aachen, Aachen, Germany. [116]Institute for Biology I, RWTH Aachen University, Aachen, Germany. [117]Department of Hematology, Oncology, Hemostaseology and Stem Cell Transplantation, Medical School, RWTH Aachen University, Aachen, Germany. [118]Institute of Clinical Pharmacology, University Hospital RWTH Aachen, Aachen, Germany. [119]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany. [120]Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, Germany. [121]Department of Intensive Care, University Hospital RWTH Aachen, Aachen, Germany. [122]Institute of Pharmacology and Toxicology, Medical Faculty Aachen, RWTH Aachen University, Aachen, Germany. [123]Molecular Oncology Group, Institute of Pathology, Medical Faculty, RWTH Aachen University, Aachen, Germany. [124]RWTH centralized Biomaterial Bank (RWTH cBMB) of the Medical Faculty, RWTH Aachen University, Aachen, Germany. [125]Department of Internal Medicine I, University Hospital RWTH Aachen, Aachen, Germany. [126]Department of Pneumology and Intensive Care Medicine, University Hospital RWTH Aachen, Aachen, Germany.

Deutsche COVID-19 Omics Initiative (DeCOI)

Janine Altmüller[22], Angel Angelov[14,35], Anna C. Aschenbrenner[1,2,5,29], Robert Bals[15], Alexander Bartholomäus[36], Anke Becker[37], Matthias Becker[1,5,128], Daniela Bezdan[13,14,38], Michael Bitzer[12], Conny Blumert[39], Ezio Bonifacio[40], Peer Bork[41], Bunk Boyke[42], Helmut Blum[43], Nicolas Casadei[13,14], Thomas Clavel[44], Maria Colome-Tatche[31,45,46], Markus Cornberg[47,48,49], Inti Alberto De La Rosa Velázquez[50], Andreas Diefenbach[51], Alexander Dilthey[52], Nicole Fischer[53], Konrad Förstner[54], Sören Franzenburg[11], Julia-Stefanie Frick[14,35], Gisela Gabernet[14,55], Julien Gagneur[56], Tina Ganzenmueller[38], Marie Gauder[14,55], Janina Geißert[14,35], Alexander Goesmann[57], Siri Göpel[12], Adam Grundhoff[23,58], Hajo Grundmann[59], Torsten Hain[60], Frank Hanses[61], Ute Hehr[62], André Heimbach[63], Marius Hoeper[64], Friedemann Horn[39], Daniel Hübschmann[65,66,67], Michael Hummel[68,69], Thomas Iftner[38], Angelika Iftner[38], Thomas Illig[70], Stefan Janssen[71], Jörn Kalinowski[72], René Kallies[73], Birte Kehr[74], Andreas Keller[18,28], Oliver T. Keppler[75,76], Sarah Kim-Hellmuth[16], Christoph Klein[16], Michael Knop[77,78], Oliver Kohlbacher[79,80], Karl Köhrer[81], Jan Korbel[41], Peter G. Kremsner[82], Denise Kühnert[83], Ingo Kurth[27], Markus Landthaler[84], Yang Li[85], Kerstin U. Ludwig[63], Oliwia Makarewicz[86], Manja Marz[87,88], Alice C. McHardy[89], Christian Mertes[56], Maximilian Münchhoff[75,76], Sven Nahnsen[14,55], Markus Nöthen[63], Francine Ntoumi[90], Peter Nürnberg[22], Stephan Ossowski[13,14], Jörg Overmann[42], Silke Peter[14,35], Klaus Pfeffer[52], Isabell Pink[47], Anna R. Poetsch[91], Ulrike Protzer[92], Alfred Pühler[72], Nikolaus Rajewsky[84], Markus Ralser[93], Kristin Reiche[39], Olaf Rieß[13,14], Stephan Ripke[94], Ulisses Nunes da Rocha[73], Philip Rosenstiel[11], Antoine-Emmanuel Saliba[95], Leif Erik Sander[96], Birgit Sawitzki[97], Simone Scheithauer[98], Philipp Schiffer[99], Jonathan Schmid-Burgk[100], Wulf Schneider[61], Eva-Christina Schulte[101], Joachim L. Schultze[1,2,5,128], Alexander Sczyrba[72], Mariam L. Sharaf[2], Yogesh Singh[13,14], Michael Sonnabend[14,35], Oliver Stegle[41,102], Jens Stoye[103], Fabian Theis[31], Thomas Ulas[1,2], Janne Vehreschild[21,104,105], Thirumalaisamy P. Velavan[90], Jörg Vogel[95], Sonja Volland[70], Max von Kleist[106,107], Andreas Walker[108], Jörn Walter[109], Dagmar Wieczorek[110], Sylke Winkler[111] & John Ziebuhr[112]

[35]Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany. [36]Geomicrobiology, German Research Centre for Geosciences (GFZ), Potsdam, Germany. [37]LOEWE Center for Synthetic Microbiology (SYNMIKRO), Philipps-Universität Marburg, Marburg, Germany. [38]Institute for Medical Virology and Epidemiology of Viral Diseases, University of Tübingen, Tübingen, Germany. [39]Fraunhofer Institute for Cell Therapy and Immunology (IZI), Leipzig, Germany. [40]Center for Regenerative Therapies Dresden (CRTD), Dresden, Germany. [41]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [42]DSMZ - German Collection of Microorganisms and Cell Cultures, Leibniz Institute, Braunschweig, Germany. [43]Gene Center - Functional Genomics Analysis, Ludwig-Maximilians-Universität München, München, Germany. [44]Institute for Medical Microbiology, University Hospital Aachen, RWTH Aachen, Germany. [45]European Research Institute for the Biology of Ageing, University of Groningen, Groningen, The Netherlands. [46]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [47]Klinik für Gastroenterologie, Hepatologie und Endokrinologie, Medizinische Hochschule Hannover (MHH), Hannover, Germany. [48]Centre for Individualised Infection Medicine (CiiM), Hannover, Germany. [49]German Center for Infection Research (DZIF), Hannover, Germany. [50]Genome Analysis Center, Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt, Neuherberg, Germany. [51]Institut für Mikrobiologie und Infektionsimmunologie, Charité – Universitätsmedizin Berlin, Berlin, Germany. [52]Institut für Medizinische Mikrobiologie und Krankenhaushygiene, Universitätsklinikum Düsseldorf, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany. [53]Institut für Medizinische Mikrobiologie, Virologie und Hygiene, Universitätsklinikum Hamburg- Eppendorf (UKE), Hamburg, Germany. [54]German Information Centre for Life Sciences (ZB MED), Cologne, Germany. [55]Quantitative Biology Center, University of Tübingen, Tübingen, Germany. [56]Informatik 29 - Computational Molecular Medicine, Technische Universität München, München, Germany. [57]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany. [58]Leibniz Institut für Experimentelle Virologie, Hamburg, Germany. [59]Institute for Infection Prevention and Hospital Hygiene, Universitätsklinikum Freiburg, Freiburg, Germany. [60]Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany. [61]Krankenhaushygiene und Infektiologie, Universitätsklinikum Regensburg, Regensburg, Germany. [62]Zentrum für Humangenetik Regensburg, Regensburg, Germany. [63]Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany. [64]Klinik für Pneumonologie, Medizinische Hochschule Hannover (MHH), Hannover, Germany. [65]Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany. [66]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg, Germany. [67]German Cancer Consortium (DKTK), Heidelberg, Germany. [68]Institute for Pathology, Molecular Pathology, Charité – Universitätsmedizin Berlin, Berlin, Germany. [69]German Biobank Node (bbmri.de), Berlin, Germany. [70]Medizinische Hochschule Hannover (MHH), Hannover Unified Biobank and Institute of Human Genetics, Hannover, Germany. [71]Algorithmic Bioinformatics, Justus Liebig University Giessen, Giessen, Germany. [72]Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany. [73]Department of Environmental Microbiology, Helmholtz-Zentrum für Umweltforschung (UFZ), Leipzig, Germany. [74]Algorithmische Bioinformatik, RCI Regensburger Centrum für Interventionelle Immunologie, Universitätsklinikum Regensburg, Regensburg, Germany. [75]Max von Pettenkofer Institute & Gene Center, Virology, National Reference Center for Retroviruses, LMU München, Munich, Germany. [76]German Center for Infection Research (DZIF), partner site Munich, München, Germany. [77]Center for Molecular Biology (ZMBH), Heidelberg University, Heidelberg, Germany. [78]Cell Morphogenesis and Signal Transduction, German Cancer Research Center (DKFZ), Heidelberg, Germany. [79]Applied Bioinformatics, University of Tübingen, Tübingen, Germany. [80]Translational Bioinformatics, University Hospital, University of Tübingen, Tübingen, Germany. [81]Genomics & Transcriptomics Labor (GTL), Universitätsklinikum Düsseldorf, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany. [82]Medical Clinic Internal Medicine VII, University Hospital, University of Tübingen, Tübingen, Germany. [83]Transmission, Infection, Diversification and Evolution Group, Max Planck Institute for the Science of Human History, Jena, Germany. [84]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [85]Centre for Individualized Infection Medicine (CiiM) & TWINCORE, joint ventures between the Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover, Germany. [86]Institute for Infection Medicine and Hospital Hygiene (IIMK), Uniklinikum Jena, Jena, Germany. [87]Michael Stifel Center Jena, Jena, Germany. [88]Bioinformatics/High-Throughput Analysis, Faculty of Mathematics and Computer Science, Friedrich-Schiller-Universität Jena, Jena, Germany. [89]Computational Biology for Infection Research, Helmholtz Centre for Infection Research (HZI), Brunswick, Germany. [90]Institute for Tropical Medicine, University Hospital, University of Tübingen, Tübingen, Germany. [91]Biotechnology Center (BIOTEC) TU Dresden, National Center for Tumor Diseases, Dresden, Germany. [92]Institute of Virology, Technical University of Munich, Munich, Germany. [93]Institute of Biochemistry, Charité – Universitätsmedizin Berlin, Berlin, Germany. [94]Department of Psychiatry and Neurosciences, Charité – Universitätsmedizin Berlin, Berlin, Germany. [95]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Center for Infection Research, Würzburg, Germany. [96]Department of Internal Medicine with emphasis on Infectiology, Respiratory-, and Critical-Care-Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany. [97]Institute of Medical Immunology, Charité – Universitätsmedizin Berlin, Berlin, Germany. [98]Institute of Infection Control and Infectious Diseases, University Medical Center, Georg August University, Göttingen, Germany. [99]Institute of Zoology, University of Cologne, Cologne, Germany. [100]Institute of Clinical Chemistry and Clinical Pharmacology, University Hospital, University of Bonn, Bonn, Germany. [101]Klinik für Psychiatrie und Psychotherapie und Institut für Psychiatrische Phänomik und Genomik, LMU München, Munich, Germany. [102]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [103]Genome Informatics, University of Bielefeld, Bielefeld, Germany. [104]Department I of Internal Medicine, University Hospital of Cologne, University of Cologne, Cologne, Germany. [105]University Hospital Frankfurt, Frankfurt am Main, Germany. [106]Institute for Bioinformatics, Freie Universität Berlin, Berlin, Germany. [107]Robert Koch Institute, Berlin, Germany. [108]Institut für Virologie, Universitätsklinikum Düsseldorf, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany. [109]Genetics and Epigenetics, Saarland University, Saarbrücken, Germany. [110]Institut für Humangenetik, Universitätsklinikum Düsseldorf, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany. [111]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany and DRESDEN concept Genome Center, TU Dresden, Dresden, Germany. [112]Institute of Medical Virology, Justus Liebig University Giessen, Giessen, Germany.

Methods

Pre-processing

PBMC transcriptome dataset (dataset A). We used a previously published dataset compiled for predicting AML in blood transcriptomes derived from PBMCs (Supplementary Information)[3]. In brief, all raw data files were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/) and the RNA-seq data were preprocessed using the kallisto v0.43.1 aligner against the human reference genome gencode v27 (GRCh38.p10). For normalization, we considered all platforms independently, meaning that normalization was performed separately for the samples in datasets A1, A2 and A3. Microarray data (datasets A1 and A2) were normalized using the robust multichip average (RMA) expression measures, as implemented in the R package affy v.1.60.0. The RNA-seq data (dataset A3) were normalized using the R package DESeq2 (v 1.22.2) with standard parameters. To keep the datasets comparable, data were filtered for genes annotated in all three datasets, which resulted in 12,708 genes. No filtering of low-expressed genes was performed. All scripts used in this study for pre-processing are provided as a docker container on Docker Hub (v 0.1, https://hub.docker.com/r/schultzelab/aml_classifier).

Whole-blood-derived transcriptome datasets (datasets B, D and E). As alignment of whole blood transcriptome data can be performed in many ways, we re-aligned all downloaded and collected datasets (Supplementary Information; these were 30.6 terabytes in size and comprised a total of 63.4 terabases) to the human reference genome gencode v33 (GRCh38.p13) and quantified transcript counts using STAR, an ultrafast universal RNA-seq aligner (v.2.7.3a). For all samples in datasets B, D, and E, raw counts were imported using DESeq (v.1.22.2, DESeqDataSetFromMatrix function) and size factors for normalization were calculated using the DESeq function with standard parameters. This was done separately for datasets B, D, and E. As some of the samples were prepared with poly-A selection to enrich for protein-coding mRNAs, we filtered the complete dataset for protein-coding genes to ensure greater comparability across library preparation protocols. Furthermore, we excluded all ribosomal protein-coding genes, as well as mitochondrial genes and genes coding for haemoglobins, which resulted in 18,135 transcripts as the feature space in dataset B, 19,358 in dataset D and 19,399 in dataset E. Furthermore, transcripts with overall expression <100 were excluded from further analysis. Other than that, no filtering of transcripts was performed. Before using the data in machine learning, we performed a rank transformation to normality on datasets B, D and E. In brief, transcript expression values were transformed from RNA-seq counts to their ranks. This was done transcript-wise, meaning that all transcript expression values per sample were given a rank based on ordering them from lowest to highest value. The rankings were then turned into quantiles and transformed using the inverse cumulative distribution function of the normal distribution. This leads to all transcripts following the exact same distribution (that is, a standard normal with a mean of 0 and a standard deviation of 1 across all samples). All scripts used in this study for pre-processing are provided on Github (https://github.com/schultzelab/swarm_learning) and normalized and rank-transformed count matrices used for predictions are provided via FASTGenomics at https://beta.fastgenomics.org/p/swarm-learning.

X-ray dataset (dataset C). The National Institutes of Health (NIH) chest X-Ray dataset (Supplementary Information) was downloaded from https://www.kaggle.com/nih-chest-xrays/data[32]. To preprocess the data, we used Keras (v.2.3.1) real-time data augmentation and generation APIs (keras.preprocessing.image.ImageDataGenerator and flow_from_dataframe). The following pre-processing arguments were used: height or width shift range (about 5%), random rotation range (about 5°), random zoom range (about 0.15), sample-wise centre and standard normalization. In addition, all images were resized to 128 × 128 pixels from their original size of 1,024 × 1,024 pixels and 32 images per batch were used for model training.

The Swarm Learning framework

SL builds on two proven technologies, distributed machine learning and blockchain (Supplementary Information). The SLL is a framework to enable decentralized training of machine learning models without sharing the data. It is designed to make it possible for a set of nodes—each node possessing some training data locally—to train a common machine learning model collaboratively without sharing the training data. This can be achieved by individual nodes sharing parameters (weights) derived from training the model on the local data. This allows local measures at the nodes to maintain the confidentiality and privacy of the raw data. Notably, in contrast to many existing federated learning models, a central parameter server is omitted in SL. Detailed descriptions of the SLL, the architecture principles, the SL process, implementation, and the environment can be found in the Supplementary Information.

Hardware architecture used for simulations

For all simulations provided in this project we used two HPE Apollo 6500 Gen10 servers, each with four Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz, a 3.2-terabyte hard disk drive, 256 GB RAM, eight Tesla P100 GPUs, a 1-GB network interface card for LAN access and an InfiniBand FDR for high speed interconnection and networked storage access. The Swarm network is created with a minimum of 3 up to a maximum of 32 training nodes, and each node is a docker container with access to GPU resources. Multiple experiments were run in parallel using this configuration.

Overall, we performed 16,694 analyses including 26 scenarios for AML, four scenarios for ALL, 13 scenarios for TB, one scenario for detection of atelectasis, effusion, and/or infiltration in chest X-rays, and 18 scenarios for COVID-19 (Supplementary Information). We performed 5–100 permutations per scenario and each permutation took approximately 30 min, which resulted in a total of 8,347 computer hours.

Computation and algorithms

Neural network algorithm. We leveraged a deep neural network with a sequential architecture as implemented in Keras (v 2.3.1)[28]. Keras is an open source software library that provides a Python interface to neural networks. The Keras API was developed with a focus on fast experimentation and is standard for deep learning researchers. The model, which was already available in Keras for R from the previous study[3], has been translated from R to Python to make it compatible with the SLL (Supplementary Information). In brief, the neural network consists of one input layer, eight hidden layers and one output layer. The input layer is densely connected and consists of 256 nodes, a rectified linear unit activation function and a dropout rate of 40%. From the first to the eighth hidden layer, nodes are reduced from 1,024 to 64 nodes, and all layers contain a rectified linear unit activation function, a kernel regularization with an L2 regularization factor of 0.005 and a dropout rate of 30%. The output layer is densely connected and consists of one node and a sigmoid activation function. The model is configured for training with Adam optimization and to compute the binary cross-entropy loss between true labels and predicted labels.

The model is used for training both the individual nodes and SL. The model is trained over 100 epochs, with varying batch sizes. Batch sizes of 8, 16, 32, 64 and 128 are used, depending on the number of training samples. The full code for the model is provided on Github (https://github.com/schultzelab/swarm_learning/)

Least absolute shrinkage and selection operator (LASSO). SL is not restricted to any particular classification algorithm. We therefore adapted the l1-penalized logistic regression[3] to be used with the SLL in the form of a Keras single dense layer with linear activation. The regularization

parameter lambda was set to 0.01. The full code for the model is provided on Github (https://github.com/schultzelab/swarm_learning/)

Parameter tuning. For most scenarios, default settings were used without parameter tuning. For some of the scenarios we tuned model hyperparameters. For some scenarios we also tuned SL parameters to get better performance (for example, higher sensitivity) (Supplementary Table 8). For example, for AML (Fig. 2e, f, Extended Data Fig. 2), the dropout rate was reduced to 10% to get better performance. For AML (Fig. 2b), the dropout rate was reduced to 10% and the epochs increased to 300 to get better performance. We also used the adaptive_rv parameter in the SL API to adjust the merge frequency dynamically on the basis of model convergence, to improve the training time. For TB and COVID-19, the test dropout rate was reduced to 10% for all scenarios. For the TB scenarios (Extended Data Fig. 7f, g), the node_weightage parameter of the SL callback API was used to give more weight to nodes that had more case samples. Supplementary Table 8 provides a complete overview of all tuning parameters used.

Parameter merging. Different functions are available for parameter merging as a configuration of the Swarm API, which are then applied by the leader at every synchronization interval. The parameters can be merged as average, weighted average, minimum, maximum, or median functions.

In this Article, we used the weighted average, which is defined as

$$P_{\mathrm{M}} = \frac{\sum_{k=1}^{n}(W_k \times P_k)}{n \times \sum_{k=1}^{n} W_k}$$

in which P_{M} is merged parameters, P_k is parameters from the kth node, W_k is the weight of the kth node, and n is the number of nodes participating in the merge process.

Unless stated otherwise, we used a simple average without weights to merge the parameter for neural networks and for the LASSO algorithm.

Quantification and statistical analysis

We evaluated binary classification model performance with sensitivity, specificity, accuracy, F1 score, and AUC metrics, which were determined for every test run. The 95% confidence intervals of all performance metrics were estimated using bootstrapping. For AML and ALL, 100 permutations per scenario were run for each scenario. For TB, the performance metrics were collected by running 10 to 50 permutations. For the X-ray images, 10 permutations were performed. For COVID-19 the performance metrics were collected by running 10 to 20 permutations for each scenario. All metrics are listed in Supplementary Tables 3, 4.

Differences in performance metrics were tested using the one-sided Wilcoxon signed rank test with continuity correction. All test results are provided in Supplementary Table 5.

To run the experiments, we used Python version 3.6.9 with Keras version 2.3.1 and TensorFlow version 2.2.0-rc2. We used scikit-learn library version 0.23.1 to calculate values for the metrics. Summary statistics and hypothesis tests were calculated using R version 3.5.2. Calculation of each metric was done as follows:

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\text{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$\text{F1score} = \frac{2\mathrm{TP}}{\mathrm{FP} + \mathrm{FN} + 2\mathrm{TP}}$$

where TP is true positive, FP is false positive, TN is true negative and FN is false negative. The area under the ROC curve was calculated using the R package ROCR version 1.0-11.

No statistical methods were used to predetermine sample size. The experiments were not randomized, but permutations were performed. Investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

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

Data availability

Processed data from datasets A1–A3 can be accessed from GEO via the superseries GSE122517 or the individual subseries GSE122505 (dataset A1), GSE122511 (dataset A2) and GSE122515 (dataset A3). Dataset B consists of the following series, which can be accessed at GEO: GSE101705, GSE107104, GSE112087, GSE128078, GSE66573, GSE79362, GSE84076, and GSE89403. Furthermore, it contains the data from the Rhineland Study. The Rhineland Study dataset falls under current General Data Protection Regulations (GDPR). Access to these data can be provided to scientists in accordance with the Rhineland Study's Data Use and Access Policy. Requests to access the Rhineland Study's dataset should be directed to RS-DUAC@dzne.de. New samples generated for datasets D and E have been deposited at the European Genome-Phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001004502. The healthy RNA-seq data included from Saarbrücken are available on application from PPMI through the LONI data archive at https://www.ppmi-info.org/data. Samples received from other public repositories are listed in Supplementary Table 2. Dataset C (NIH chest X-ray dataset) is available on Kaggle (https://www.kaggle.com/nih-chest-xrays/data). Normalized log-transformed and rank transformed expressions as used for the predictions are available via FASTGenomics at https://beta.fastgenomics.org/p/swarm-learning.

Code availability

The code for preprocessing and for predictions can be found at GitHub (https://github.com/schultzelab/swarm_learning). The Swarm Learning software can be downloaded from https://myenterpriselicense.hpe.com/.

Acknowledgements We thank the Michael J. Fox Foundation and the Parkinson's Progression Markers Initiative (PPMI) for contributing RNA-seq data; the CORSAAR study group for additional blood transcriptome samples; the collaborators who contributed to the collection of COVID-19 samples (B. Schlegelberger, I. Bernemann, J. C. Hellmuth, L. Jocham, F. Hanses, U. Hehr, Y. Khodamoradi, L. Kaldjob, R. Fendel, L. T. K. Linh, P. Rosenberger, H. Häberle and J. Böhne); and the NGS Competence Center Tübingen (NCCT), who contributed to the generation of data and the data sharing (in particular, J. Frick, M. Sonnabend, J. Geissert, A. Angelov, M. Pogoda, Y. Singh, S. Poths, S. Nahnsen and M. Gauder). This work was supported in part by the German Research Foundation (DFG) to J.L.S., O.R., P.R., P.N. (INST 37/1049-1, INST 216/981-1, INST 257/605-1, INST 269/768-1, INST 217/988-1, INST 217/577-1, INST 217/1011-1, INST 217/1017-1 and INST 217/1029-1); under Germany's Excellence Strategy (DFG – EXC2151 – 390873048); by the HGF Incubator grant sparse2big (ZT-I-0007); by EU projects SYSCID (grant 733100, P.R.) and ImmunoSep (grant 84722, J.L.S.); and by HPE to the DZNE for generating whole blood transcriptome data from patients with COVID-19. J.L.S. was further supported by the BMBF-funded excellence project Diet–Body–Brain (DietBB) (grant 01EA1809A), and J.L.S. and J.R. by NaFoUniMedCovid19 (FKZ: 01KX2021, project acronym COVIM). S.K. is supported by the Hellenic Institute for the Study of Sepsis. The clinical study in Greece was supported by the Hellenic Institute for the Study of Sepsis. E.J.G.-B. received funding from the FrameWork 7 programme HemoSpec (granted to the National and Kapodistrian University of Athens), the Horizon2020 Marie-Curie Project European Sepsis Academy (676129, granted to the National and Kapodistrian University of Athens), and the Horizon 2020 European Grant ImmunoSep (granted to the Hellenic Institute for the Study of Sepsis). P.R. was supported by DFG ExC2167, a stimulus fund from Schleswig-Holstein and the DFG NGS Centre CCGA. The clinical study in Munich was supported by the Care-for-Rare Foundation. S.K.-H. is a scholar of the Reinhard-Frank Stiftung. D.P. is funded by the Hector Fellow Academy. The work was additionally supported by the Michael J. Fox Foundation for Parkinson' Research under grant 14446. M.G.N. was supported by an ERC Advanced Grant (833247) and a Spinoza Grant of the Netherlands Organization for Scientific Research. R.B. and A.K. were

supported by Dr. Rolf M. Schwiete Stiftung, Staatskanzlei des Saarlandes and Saarland University. J.N. is supported by the DFG (SFB TR47, SPP1937) and the Hector Foundation (M88). M.A. is supported by COVIM: NaFoUniMedCovid19 (FKZ: 01KX2021). M. Becker is supported by the HGF Helmholtz AI grant Pro-Gene-Gen (ZT-I-PF-5-23).

Author contributions The idea was conceived by H.S., K.L.S., E.L.G., and J.L.S. Subprojects and clinical studies were directed by H.S., K.L.S., K.H., M. Bitzer, J.R., S.K.-H., J.N., I.K., A.K., R.B., P.N., O.R., P.R., M.M.B.B., M. Becker, and J.L.S. The conceptualization was performed by S.W.-H., H.S., K.L.S., M. Becker, S.C., M.S.W., E.L.G., and J.L.S. Direction of the clinical programs, collection of clinical information and patient diagnostics were done by P.P., N.A.A., S.K., F.T., M. Bitzer, C.H., D.P., U.B., F.K., T.F., P.S., C.L., M.A., J.R., B.K., M.S., J.H., S.S., S.K.-H., J.N., D.S., I.K., A.K., R.B., M.G.N., M.M.B.B., E.J.G.-B, and M.K. Patient samples were provided by P.P., N.A.A., S.K., F.T., M. Bitzer, S.O., N.C., C.H., D.P., U.B., F.K., T.F., P.S, C.L., M.A., J.R., B.K., M.S., J.H., S.S., S.K.-H., J.N., D.S., I.K., A.K., R.B., M.G.N., M.M.B.B., E.J.G.-B, and M.K. Laboratory experiments were performed by K.H., S.O., N.C., J.A., L.B., J.S.-S., E.D.D., M.K., and H.T. Primary data analysis and data QC were provided by S.W.-H., K.H., S.O., N.C., J.A., N.M., J.P.B., L.B., J.S.-S., E.D.D., M.N.-G., A.K., P.N., O.R., P.R., T.U., M. Becker, and J.L.S. Programming and coding for the current project were done by S.W.-H., Saikat Mukherjee, V.G., R.S., C.S., M.D., C.M.S., and M. Becker. The Swarm Learning environment was developed by S. Manamohan, Saikat Mukherjee, V.G., R.S., M.D., B.M., S.C., M.S.W., and E.L.G. Statistics and machine learning were done by S.W.-H., Saikat Mukherjee, V.G., R.S., M.D., F.T., Sach Mukherjee, S.C., E.L.G., and J.L.S. Data privacy and confidentiality concepts were developed by H.S., K.L.S., M. Backes, E.L.G., and J.L.S. Data interpretation was done by S.W.-H., H.S., Saikat Mukherjee, A.C.A., M. Becker, and J.L.S. Data were visualized by S.W.-H., H.S., M. Becker, and J.L.S. The original draft was written by S.W.-H., H.S., K.L.S., A.C.A., M. Becker, and J.L.S. Writing, reviewing and editing of revisions was done by S.W.-H., H.S., K.L.S., A.C.A., M.M.B.B., M. Becker, E.L.G., and J.L.S. Project management and administration were performed by H.S., K.L.S., A.D., A.C.A., M. Becker, and J.L.S. Funding was acquired by H.S., S.K., D.P., M.A., J.R., S.K.-H., J.N., A.K., R.B., P.N., O.R., P.R., M.G.N., F.T., E.J.G.-B, M.B., S.C., and J.L.S. All authors commented on the manuscript.

Funding Open access funding provided by Deutsches Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE) in der Helmholtz-Gemeinschaft.

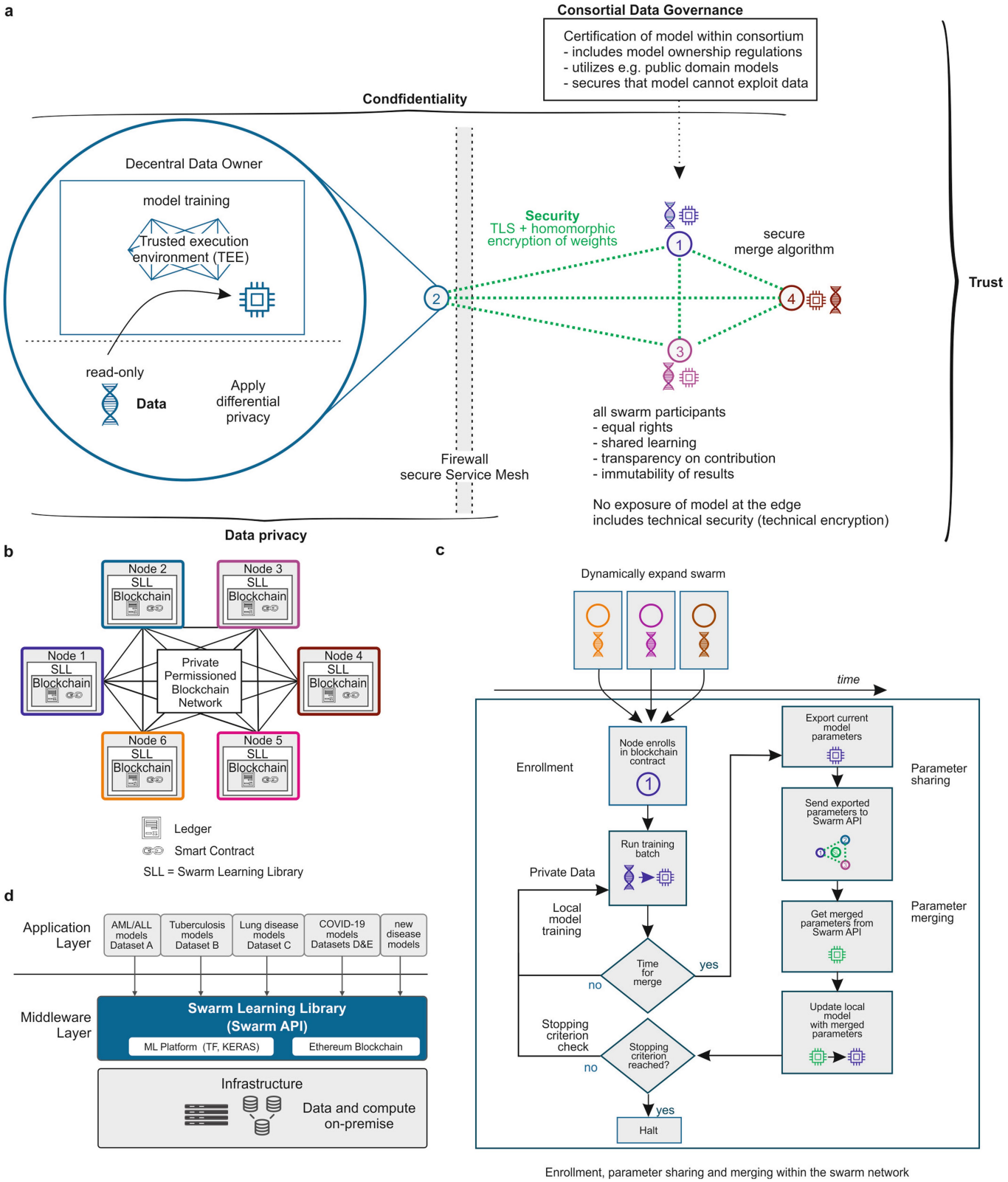
Competing interests H.S., K.L.S., S. Manamohan, Saikat Mukherjee, V.G., R.S., C.S., M.D., B.M, C.M.S., S.C., M.S.W. and E.L.G. are employees of Hewlett Packard Enterprise. Hewlett Packard Enterprise developed the SLL in its entirety as described in this work and has submitted multiple associated patent applications. E.J.G.-B. received honoraria from AbbVie USA, Abbott CH, InflaRx GmbH, MSD Greece, XBiotech Inc. and Angelini Italy and independent educational grants from AbbVie, Abbott, Astellas Pharma Europe, AxisShield, bioMérieux Inc, InflaRx GmbH, and XBiotech Inc. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03583-3.

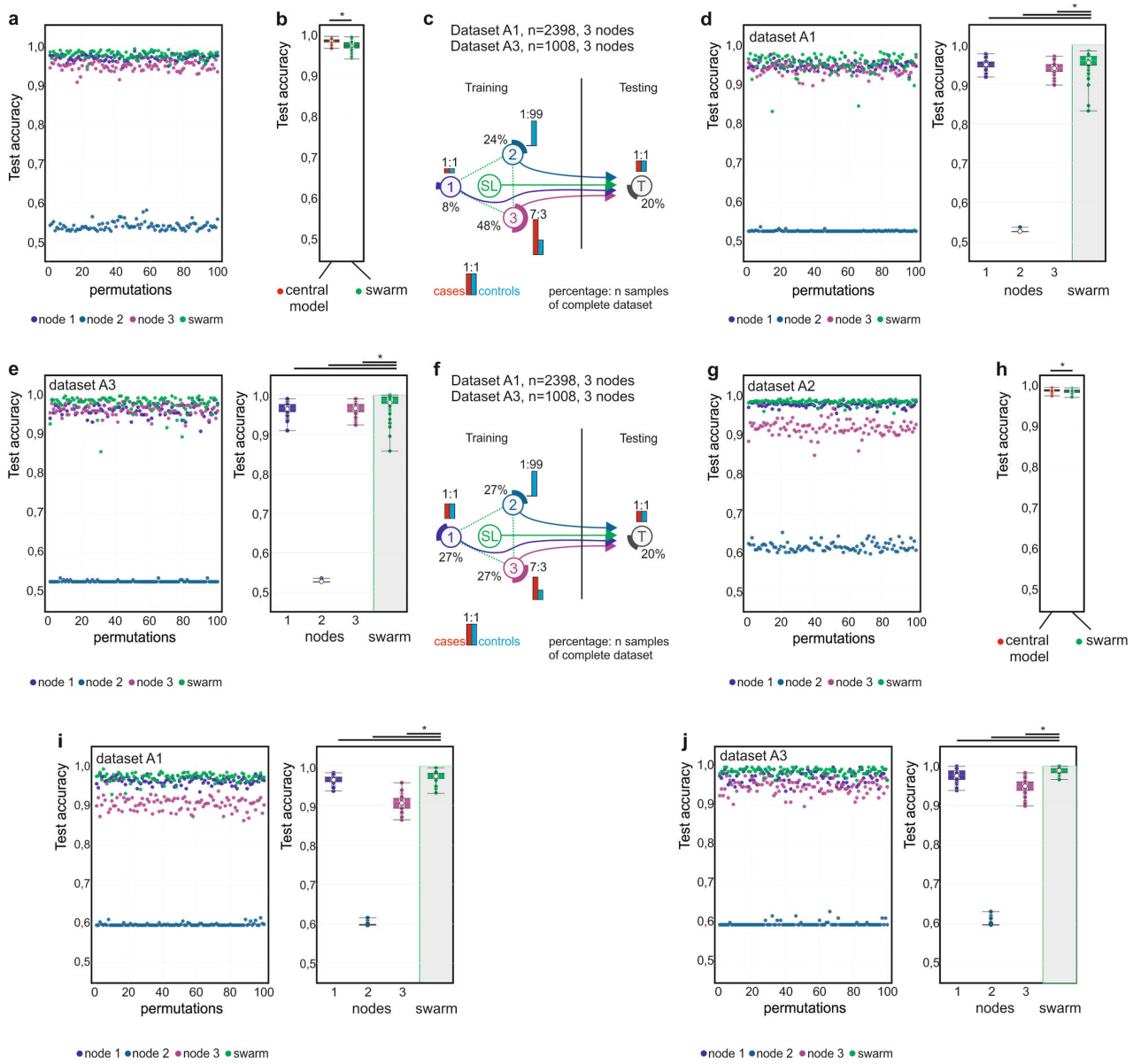
Correspondence and requests for materials should be addressed to J.L.S.

Peer review information *Nature* thanks Dianbo Liu, Christopher Mason and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

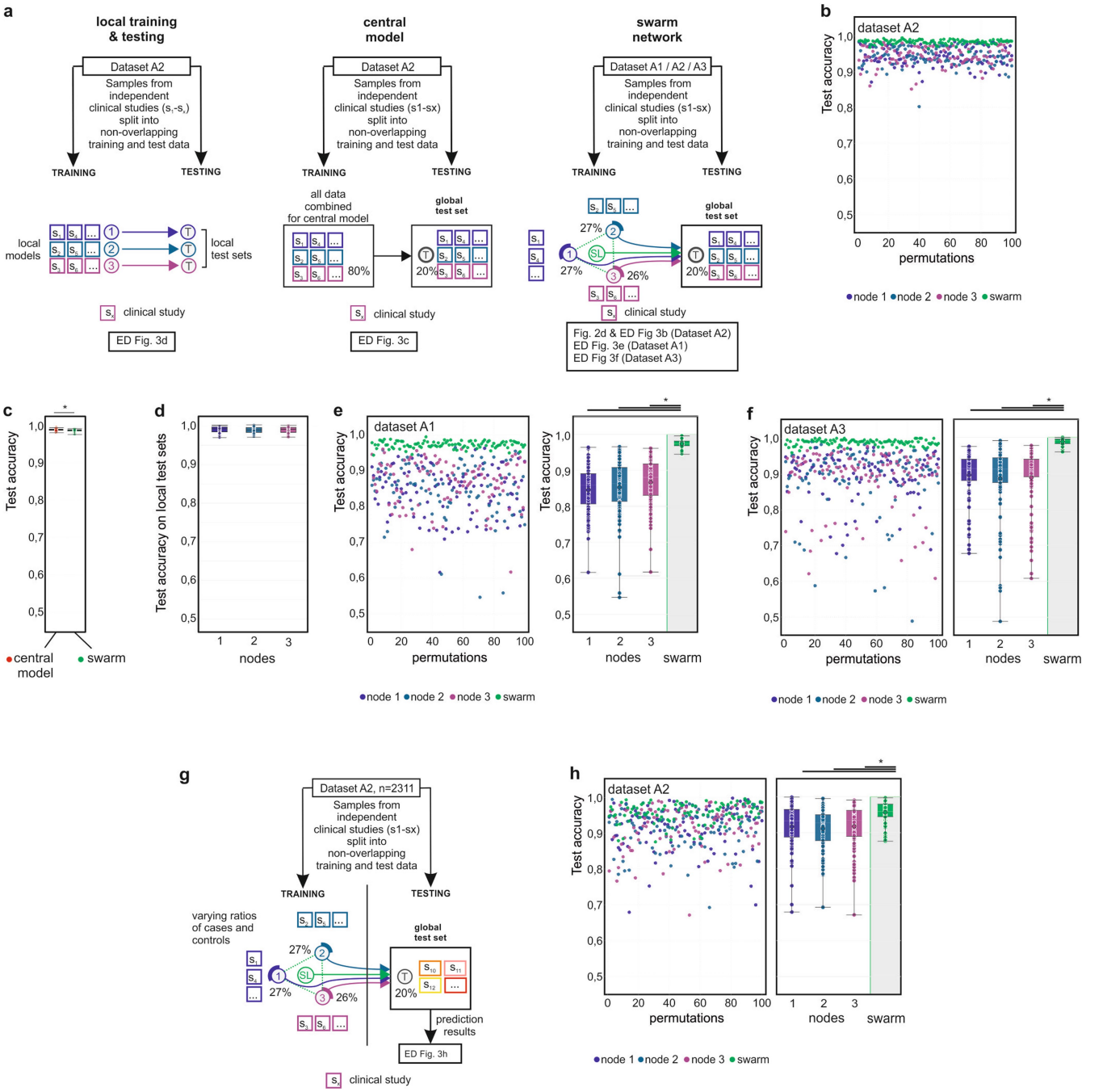


Enrollment, parameter sharing and merging within the swarm network

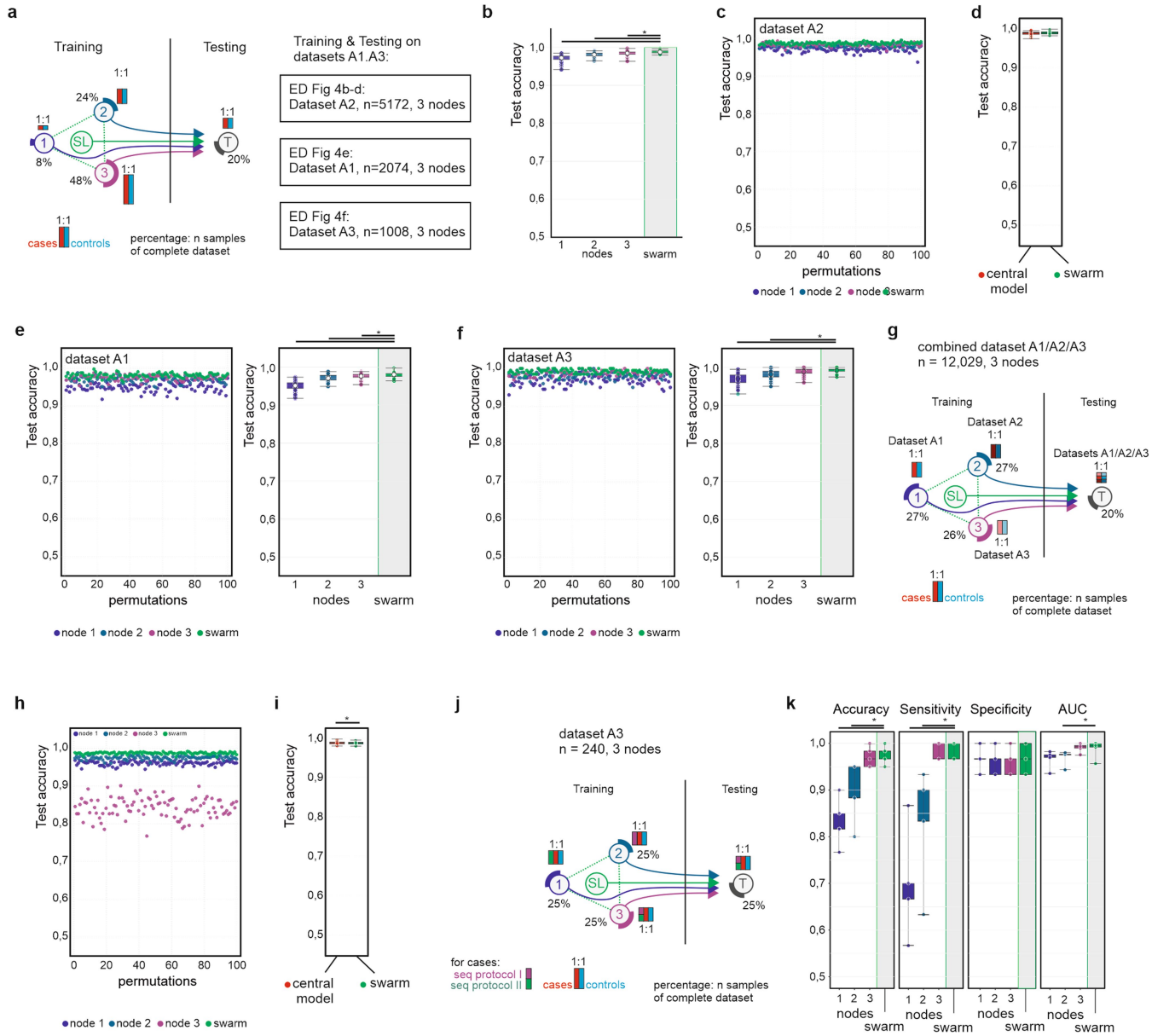
Extended Data Fig. 1 | Corresponding to Fig. 1. a, Overview of SL and the relationship to data privacy, confidentiality and trust. **b**, Concept and outline of the private permissioned blockchain network as a layer of the SL network. Each node consists of the blockchain, including the ledger and smart contract, as well as the SLL with the API to interact with other nodes within the network. **c**, The principles of the SL workflow once the nodes have been enrolled within the Swarm network via private permissioned blockchain contract and dynamic onboarding of new Swarm nodes. **d**, Application and middleware layer as part of the SL concept.



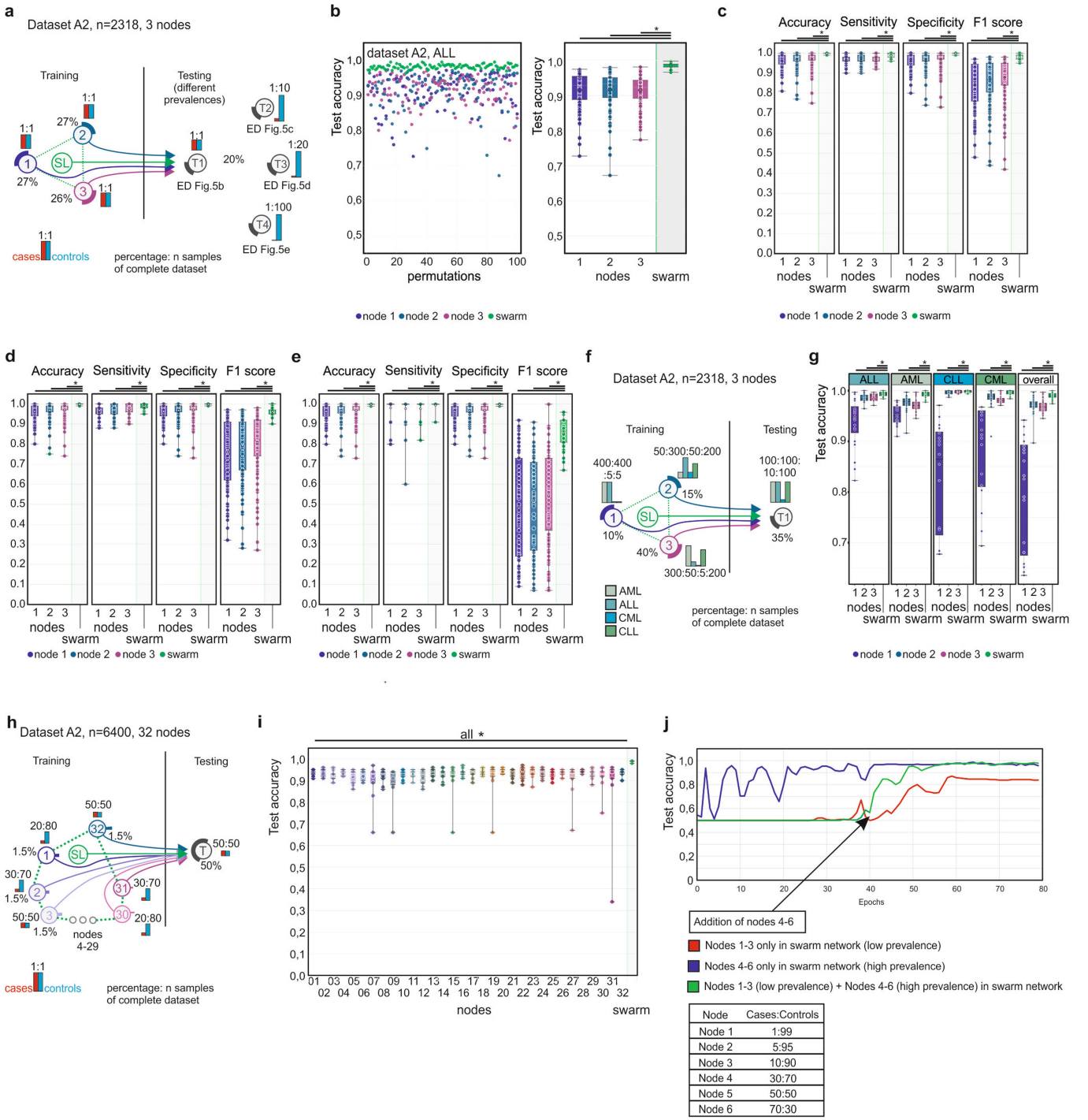
Extended Data Fig. 2 | Scenario corresponding to Fig. 2b, c in datasets A1 and A3. Main settings and representation of schema and data visualization as described in Fig. 2a. **a**, Evaluation of test accuracy for 100 permutations of the scenario shown in Fig. 2b. **b**, Evaluation of SL versus central model for the scenario shown in Fig. 2b for 100 permutations. **c**, Scenario with different prevalences of AML and numbers of samples at each training node. The test dataset has an even distribution. **d**, Evaluation of test accuracy for 100 permutations of dataset A1 per node and SL. **e**, Evaluation using dataset A3 for 100 permutations. **f**, Scenario with similar training set sizes per node but decreasing prevalence. The test dataset ratio is 1:1. **g**, Evaluation of test accuracy for 100 permutations of the scenario shown in Fig. 2c. **h**, Evaluation of SL versus central model of the scenario shown in Fig. 2c for 100 permutations. **i**, Evaluation of test accuracy over 100 permutations for dataset A1 with the scenario shown in **f**. **j**, Evaluation of test accuracy over 100 permutations for dataset A3 with the scenario shown in **f**. **b**, **d**, **e**, **h**–**j**, Box plots show representation of accuracy of 100 permutations performed for the 3 training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Accuracy is defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.



Extended Data Fig. 3 | Scenario to test for batch effects of siloed studies in datasets A1–A3 and scenario with multiple consortia. Main settings and representation of schema and data visualization are as in Fig. 2a. **a**, Scenario with training nodes coming from independent clinical studies for local models (left), central model (middle) and the Swarm network (right) and testing on a non-overlapping global test with samples from the same studies. **b**, Evaluation of test accuracy over 100 permutations for dataset A2 with the scenario shown in **a** (right) and Fig. 2d. **c**, Comparison of test accuracy between central model (**a**, middle) and SL (**a**, right). **d**, Comparison of test accuracy on the local test datasets (**a**, left) for 100 permutations. **e**, Evaluation of test accuracy of individual nodes versus SL over 100 permutations for dataset A1 when training nodes have data from independent clinical studies. **f**, Evaluation of test accuracy of individual nodes versus SL over 100 permutations for dataset A3 when training nodes have data from independent clinical studies. **g**, Scenario with three consortia contributing training nodes and a fourth one providing the testing node. **h**, Evaluation of test accuracy for scenario shown in **g** over 100 permutations for dataset A2. **d**–**f**, **h**, Box plots show representation of accuracy of all permutations performed for the 3 training nodes individually as well as the results obtained by SL (**d** only for local models). All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values are listed in Supplementary Table 5.

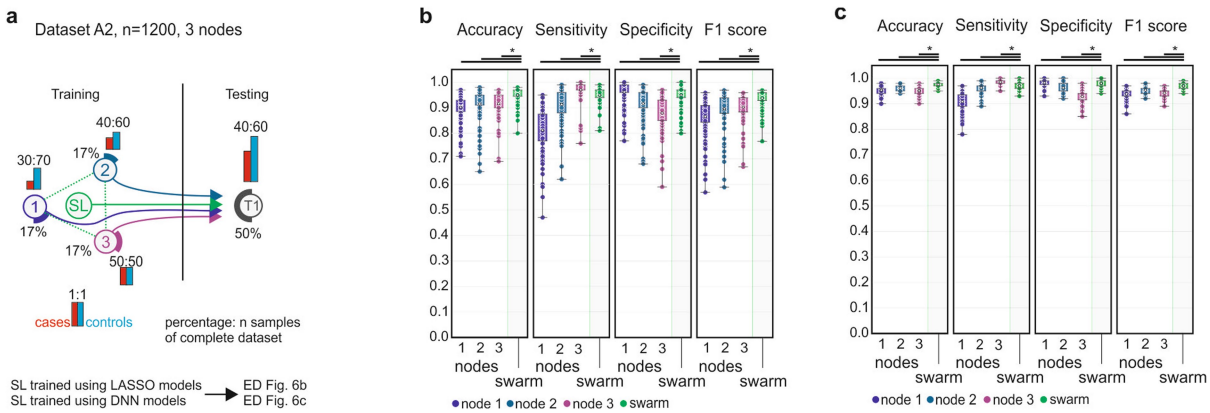


Extended Data Fig. 4 | Scenario corresponding to Fig. 2e in datasets A1 and A3 and scenario using different data generation methods in each training node. Main settings and representation of schema and data visualization are as in Fig. 2a. **a**, Scenario with even distribution of cases and controls at each training node and the test node, but different numbers of samples at each node and overall increase in numbers of samples. **b**, **c**, Test accuracy for evaluation of dataset A2 over 100 permutations. **d**, Comparison of central model with SL over 100 permutations. **e**, Test accuracy for evaluation of dataset A1 over 99 permutations. **f**, Test accuracy for evaluation of dataset A3 over 100 permutations. **g**, Scenario where datasets A1, A2, and A3 are assigned to a single training node each. **h**, Evaluation of test accuracy over 100 permutations. **i**, Comparison of the test accuracy of central model and SL over 98 permutations. **j**, Scenario similar to **g** but where the nodes use datasets from different RNA-seq protocols. **k**, Evaluation of results for accuracy, AUC, sensitivity, and specificity over five permutations. **d**–**f**, **i**, **k**, Box plots show predictive performance over all permutations performed for the three training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.

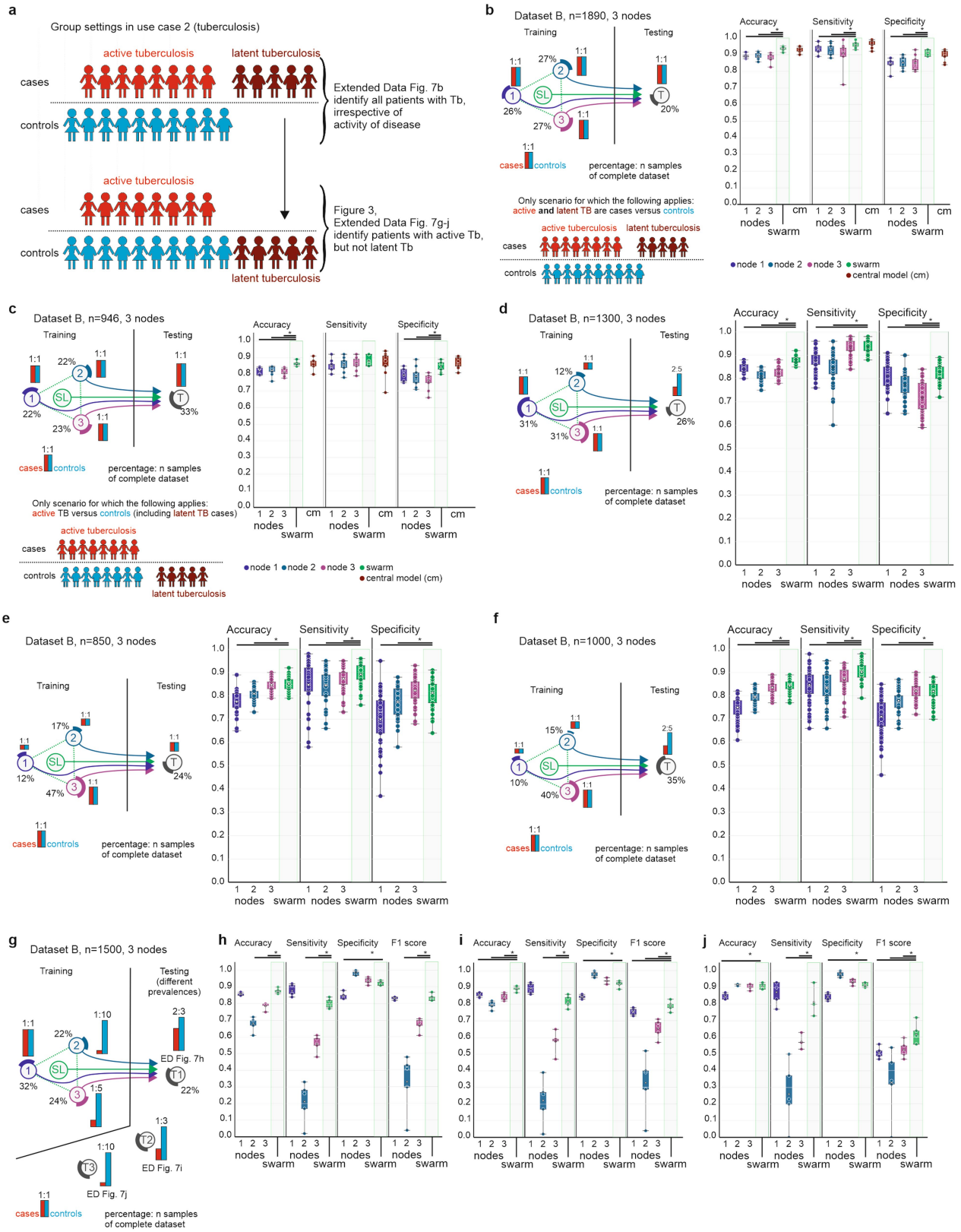


Node	Cases:Controls
Node 1	1:99
Node 2	5:95
Node 3	10:90
Node 4	30:70
Node 5	50:50
Node 6	70:30

Extended Data Fig. 5 | Scenario for ALL in dataset 2 and multi-class prediction and expansion of SL. Main settings are identical to what is described in Fig. 2a. Here cases are samples derived from patients with ALL, while all other samples are controls (including AML). **a**, Scenario for the detection of ALL in dataset A2. The training sets are evenly distributed among the nodes with varying prevalence at the testing node. Data from independent clinical studies are samples to each node, as described for AML in Fig. 2d. **b**, Evaluation of scenario in **a** for test accuracy over 100 permutations with a prevalence ratio of 1:1. **c**, Evaluation using a test dataset with prevalence ratio of 10:100 over 100 permutations. **d**, Evaluation using a test dataset with prevalence ratio of 5:100 over 100 permutations. **e**, Evaluation using a test dataset with prevalence ratio of 1:100. **f**, Scenario for multi-class prediction of different types of leukaemia in dataset A2. Each node has a different prevalence. **g**, Test accuracy for the different types of leukaemia over 20 permutations. **h**, Scenario that simulates 32 small Swarm nodes. **i**, Evaluation of test accuracy for the 32 nodes and the Swarm over 10 permutations. **j**, Development of accuracy over training epochs with addition of new nodes. **b**–**e**, **g**, **i**, Box plots show performance of all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent test node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.

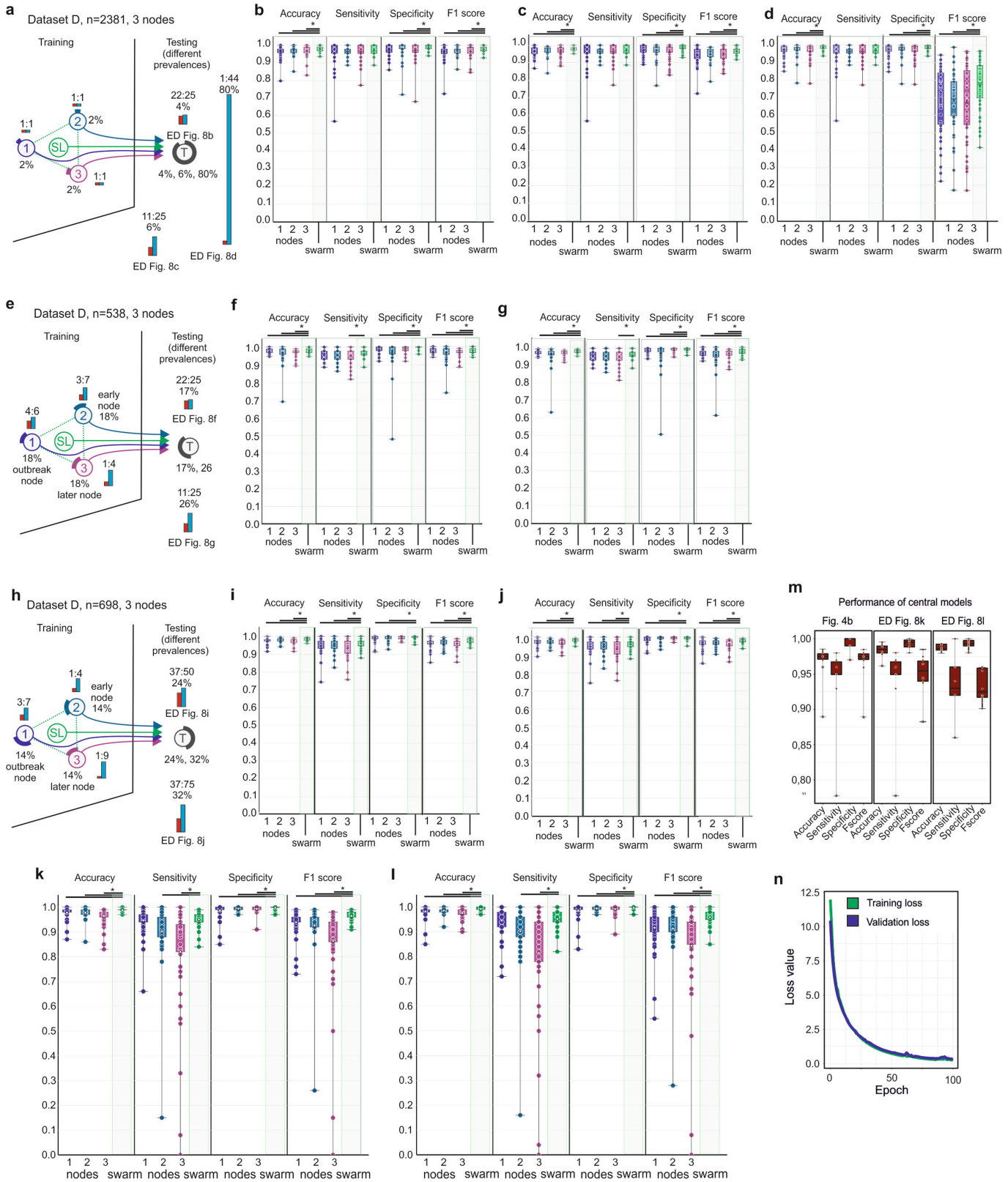


Extended Data Fig. 6 | Comparison of LASSO and neural networks. **a**, Scenario for training different models in the Swarm. **b**, Evaluation of a LASSO model for accuracy, sensitivity, specificity and F1 score over 100 permutations. **c**, Evaluation of a Neural Network model for accuracy, sensitivity, specificity and F1 score over 100 permutations. **b**, **c**, Box plots show performance of all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.



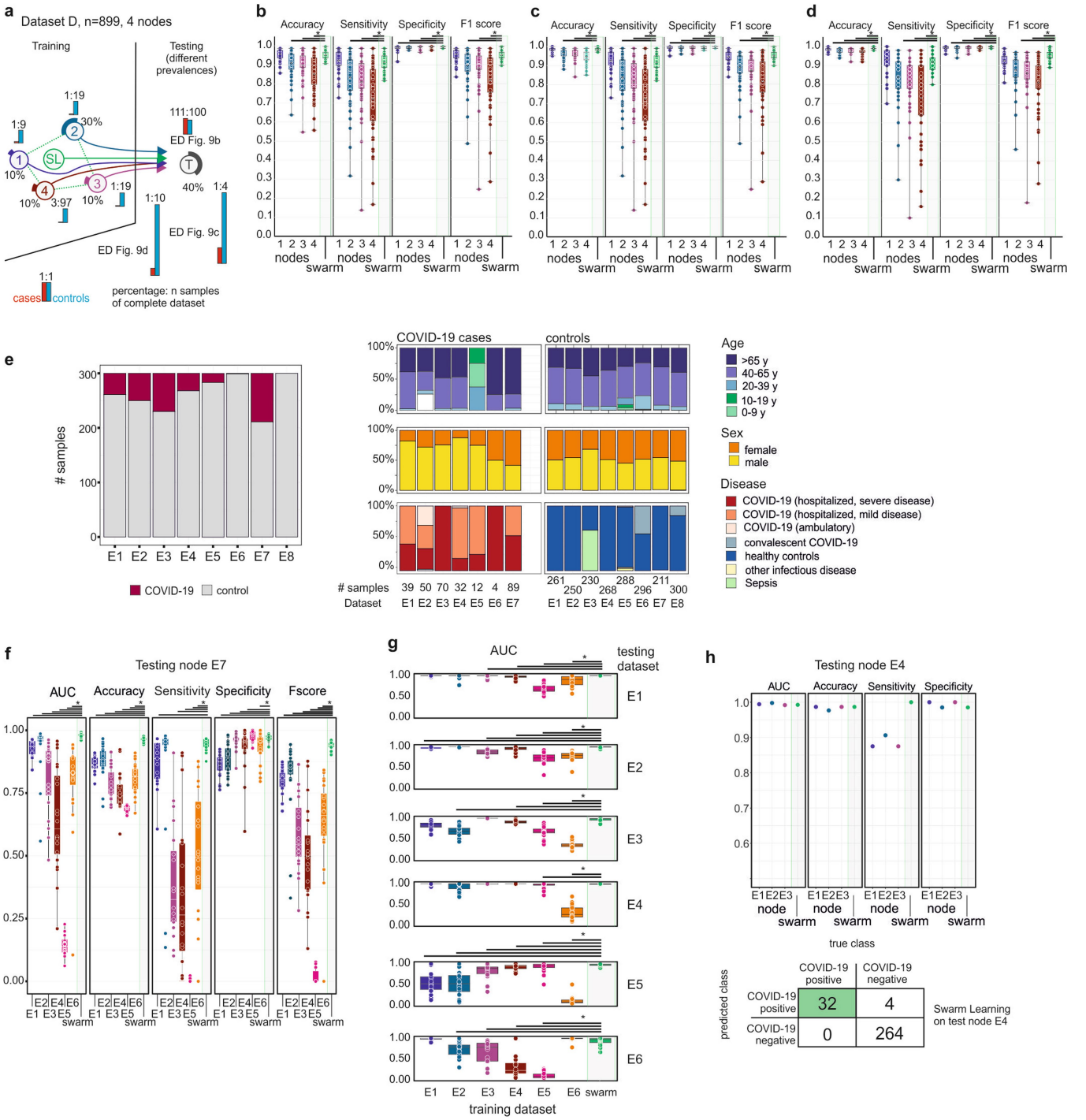
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Scenarios for detecting all TB versus controls and for detecting active TB with low prevalence at training nodes. Main settings are as in Fig. 2a. **a**, Different group settings used with assignment of latent TB to control or case. **b**, Left, evaluation of a scenario where active and latent TB are cases. The data are evenly distributed among the training nodes. Right, test accuracy, sensitivity and specificity for nodes, Swarm and a central model over 10 permutations. **c**, Left, scenario similar to **b** but with latent TB as control. Right, test accuracy, sensitivity and specificity for nodes, Swarm and a central model over 10 permutations. **d**, Left, scenario with reduced prevalence at the test node. Right, test accuracy, sensitivity and specificity for nodes and Swarm over 10 permutations. **e**, Scenario with even distribution of cases and controls at each training node, where node 1 has a very small training set. The test dataset is evenly distributed. Right, test accuracy, sensitivity and specificity over 50 permutations. **f**, Left, scenario similar to **e** but with uneven distribution in the test node. Right, test accuracy, sensitivity and specificity over 50 permutations. **g**, Scenario with each training node having a different prevalence. Three prevalence scenarios were used in the test dataset. **h**, Accuracy, sensitivity, specificity and F1 score over five permutations for testing set T1 as shown in **g**. **i**, As in **h** but with prevalence changed to 1:3 cases:controls in the training set. **j**, As in **h** but with prevalence changed to 1:10 cases:controls in the training set. **b**–**f**, **h**–**j**, Box plots show performance of all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^{*}P < 0.05$, exact P values listed in Supplementary Table 5.

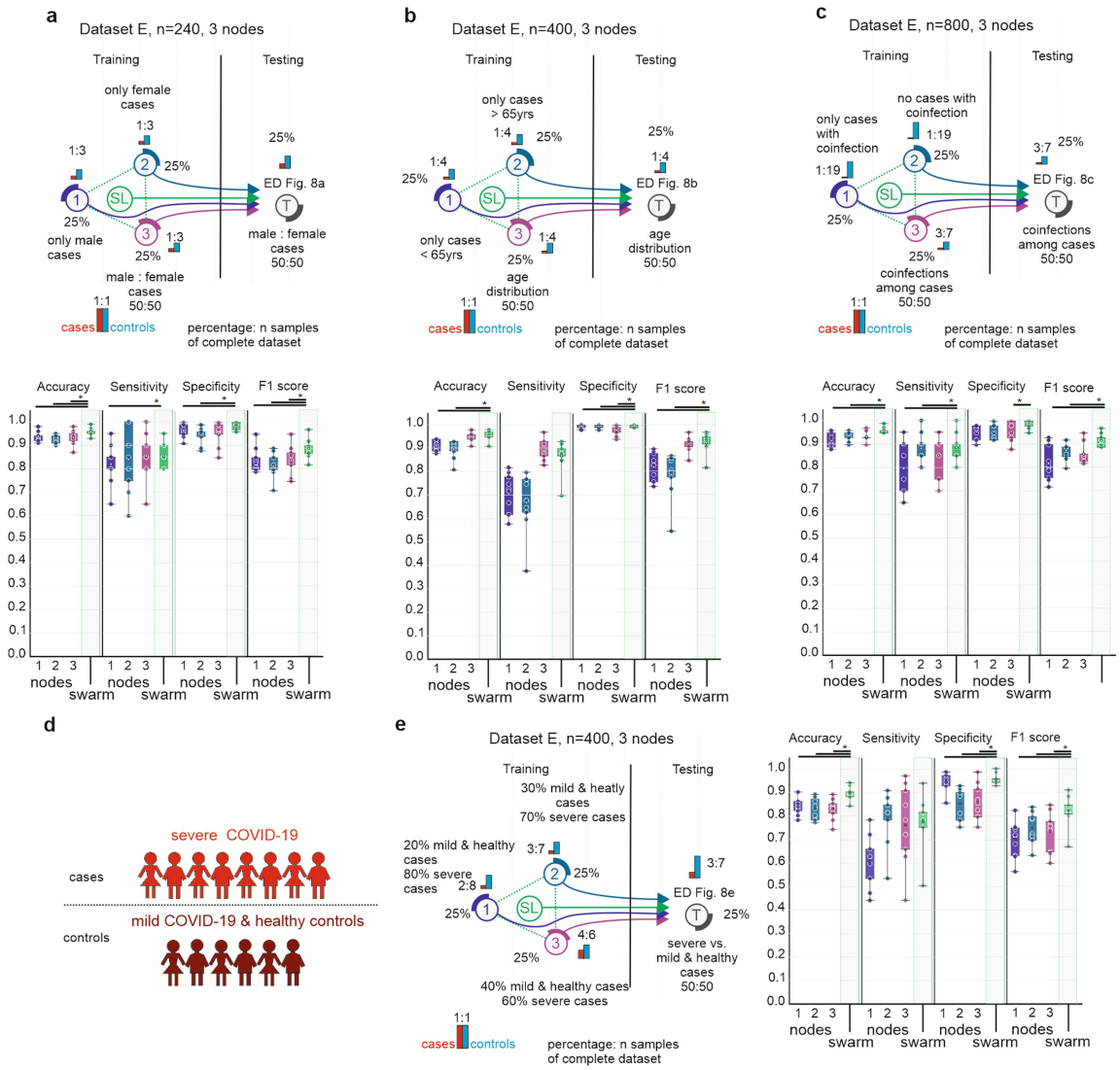


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Baseline scenario for detecting patients with COVID-19 and scenario with reduced prevalence at training nodes. Main settings are as in Fig. 2a. **a**, Scenario for detecting COVID-19 with even training set distribution among nodes 1–3. Three testing sets with different prevalences were simulated. **b**, Accuracy, sensitivity, specificity and F1 score over 50 permutations for scenario in **a** with a 22:25 case:control ratio. **c**, As in **b** for an 11:25 ratio. **d**, As in **b** for a 1:44 ratio. **e**, Scenario with the same sample size at each training node, but prevalence decreasing from node 1 to node 3. There are two test datasets (**f**, **g**). **f**, Evaluation of scenario in **e** with 22:25 ratio at the test node over 50 permutations. **g**, Evaluation of scenario in **e** with reduced prevalence over 50 permutations. **h**, Scenario similar to **e** but with a steeper decrease in prevalence between nodes 1 and 3. **i**, Evaluation of scenario in **h** with a ratio of 37:50 at the test node over 50 permutations. **j**, Evaluation of scenario in **h** with a reduced prevalence compared to **i** over 50 permutations. **k**, Scenario as in Fig. 4a using a 1:5 ratio for cases and controls in the test dataset evaluated over 50 permutations. **l**, Scenario as in Fig. 4a using a 1:10 ratio in the test dataset to simulate detection in regions with new infections, evaluated over 50 permutations. **m**, Performance of central models for **k**, **l** and Fig. 4b. **n**, Loss function of training and validation loss over 100 training epochs. **b**–**d**, **f**, **g**, **i**–**m**, Box plots show performance of all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^{*}P < 0.05$, exact P values listed in Supplementary Table 5.



Extended Data Fig. 9 | Scenario with reduced prevalence in training and test datasets and multi-centre scenario at a four-node setting. Main settings as in Fig. 2a. **a**, Scenario with prevalences from 10% at node 1 to 3% at node 4. There are three test datasets (**b**–**d**) with decreasing prevalence and increasing total sample size. **b**, Evaluation of scenario in **a** with 111:100 ratio over 50 permutations. **c**, Evaluation of scenario in **a** with 1:4 ratio and increased sample number of the test dataset over 50 permutations. **d**, Evaluation of scenario in **a** with 1:10 prevalence and increased sample number of the test dataset over 50 permutations. **e**, Dataset properties for the participating cities E1–E8, indicating case:control ratio and demographic properties. **f**, AUC, accuracy, sensitivity, specificity and F1 score over 20 permutations for scenario that uses E1–E6 as training nodes and E7 as external test node. **g**, Evaluation of a multi-city scenario where a medical centre (in each row) serves as a test node. The AUC for each training node and the SL is shown for 20 permutations. **h**, Multi-city scenario. Only three nodes (E1–E3) are used for training and the external test node E4 uses data from a different sequencing facility. AUC, accuracy, sensitivity and specificity as well as the confusion matrix for one prediction. **b**–**d**, **f**, **g**, Box plots show performance of all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^{*}P < 0.05$, exact P values listed in Supplementary Table 5.



Extended Data Fig. 10 | Scenarios for testing different factors and scenario for testing disease severity. Main settings as in Fig. 2a. **a**, Top, scenario to test influence of sex with three training nodes. Training node 1 has only male cases, node 2 has only female cases. Training node 3 and the test node have a 50%/50% split. Bottom, accuracy, sensitivity, specificity and F1 score for each training node and the Swarm in 10 permutations. **b**, Top, scenario to test influence of age with three training nodes. Training node 1 only has cases younger than 65 years, node 2 only has cases older than 65 years. Training node 3 and the test node have a 50%/50% split of cases above and below 65 years. Bottom, accuracy, sensitivity, specificity and F1 score for each training node and the Swarm in 10 permutations. **c**, Top, scenario to test influence of co-infections with three training nodes. Training node 1 has only cases with co-infections, node 2 has no cases with co-infections. Training node 3 and the test node have a 50%/50% split. Bottom, accuracy, sensitivity, specificity and F1 score for each training node and the Swarm in 10 permutations. **d**, Prediction setting. Severe cases of COVID-19 are cases, mild cases of COVID-19 and healthy donors are controls. **e**, Left, scenario to test influence of disease severity with three training nodes. Training node 1 has 20% mild or healthy and 80% severe cases, node 3 has 40% mild or healthy and 60% severe cases. Training node 2 and the test node have 30% mild or healthy and 70% severe cases. Right, accuracy, sensitivity, specificity and F1 score for each training node and the Swarm for 10 permutations. **a**–**c**, **e**, Box plots show performance all permutations performed for the training nodes individually as well as the results obtained by SL. All samples are biological replicates. Centre dot, mean; box limits, 1st and 3rd quartiles; whiskers, minimum and maximum values. Performance measures are defined for the independent fourth node used for testing only. Statistical differences between results derived by SL and all individual nodes including all permutations performed were calculated with one-sided Wilcoxon signed rank test with continuity correction; $^*P < 0.05$, exact P values listed in Supplementary Table 5.

nature research

Corresponding author(s): Joachim L. Schultze

Last updated by author(s): Apr 7, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Dataset A: All raw data files were downloaded from GEO and the RNA-seq data was preprocessed using the kallisto aligner v.0.43.1 against the human reference genome gencode v27 (GRCh38.p10). For normalization, we considered all platforms independently, meaning that normalization was performed separately for the samples in Dataset A1, A2 and A3, respectively. Microarray data (Datasets A1 and A2) was normalized using the robust multichip average (RMA) expression measures, as implemented in the R package affy (version 1.60.0). RNA-seq data (Dataset A3) was normalized with the R package DESeq2 (version 1.22.2) using standard parameters. In order to keep the datasets comparable, data was filtered for genes annotated in all three datasets, which resulted in 12,708 genes. No filtering of low-expressed genes was performed. All scripts used in this study for pre-processing are provided as a docker container on Docker Hub (version 0.1, https://hub.docker.com/r/schultzelab /aml_classifier) and GitHub (https://github.com/schultzelab/swarm_learning). Dataset B,D,E: All raw data file were downloaded from GEO or collected at the partner hospitals and aligned to the human reference genome gencode v33 (GRCh38.p13) and quantified transcript counts using STAR v 2.7.3a. For all samples in Datasets B and D,E, raw counts were imported using the R package DESeq2 (version 1.22.2, DESeqDataSetFromMatrix function) and size factors for normalization were calculated using the DESeq function using standard parameters. Dataset C: The NIH Chest X-Ray dataset was downloaded from https://www.kaggle.com/nih-chest-xrays/data. In order to preprocess the data, we used Python (version 3.6.9) and Keras (version 2.3.1) real-time data augmentation and generation APIs (keras.preprocessing.image.ImageDataGenerator and flow_from_dataframe). The following pre-processing arguments were used: height or width shift range (~ 5%), random rotation range (~ 5 degree), random zoom range (~ 0.15), sample-wise center and standard normalization. Additionally, all images are resized to (128 * 128) from their original size of (1024 * 1024).
Data analysis	All models for the experiments have been implemented using Python (version 3.6.9), Keras (version 2.3.1), Tensorflow (2.2.0-rc2) and scikit-learn (version 0.23.1). The LASSO algorithm has been implemented using Keras (version 2.3.1). All code is available on GitHub (https://github.com/schultzelab/swarm_learning). Measurements of sensitivity, specificity, accuracy and F1 score of each permutation run was read into a table in Excel (Microsoft Excel for Microsoft 365 MSO: Version: 2008 13127.21348 (16.0.13127_21336 64-bit)) using Power Query (Microsoft Excel for Microsoft 365 MSO: Version: 2008 13127.21348 (16.0.13127_21336 64-bit)) and used for visualization for the different scenarios in Power BI [Version:

2.81.5831.821 64-bit (Mai 2020)] with Box and Whisker chart by MAQ Software (https://appsource.microsoft.com/en-us/product/power-bi-visuals/WA104381351, version 3.2.1).
AUC, positive predictive value, all confidence intervals and statistical tests were calculated using R (version 3.5.2) and the R packages MKmisc (version 1.6) and ROCR (version 1.0.7).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Processed data can be accessed via the SuperSeries GSE122517 or via the individual SubSeries GSE122505 (dataset A1), GSE122511 (dataset A2) and GSE122515 (dataset A3). Dataset B consists of the following series which can be accessed at GEO: GSE101705, GSE107104, GSE112087, GSE128078, GSE66573, GSE79362, GSE84076, and GSE89403. Furthermore, it contains the Rhineland study. This dataset is not publicly available because of data protection regulations. Access to data can be provided to scientists in accordance with the Rhineland Study's Data Use and Access Policy. Requests for further information or to access the Rhineland Study's dataset should be directed to RS-DUAC@dzne.de. Dataset D and E contain dataset B and additional samples for COVID-19. These datasets are made available at the European Genome-Phenome Archive (EGA) under accession number EGAS00001004502 , which is hosted by the EBI and the CRG. The healthy RNA-seq data included from Saarbrücken is available from PPMI through the LONI data archive, https://www.ppmi-info.org/data. The NIH CC Chest X-Ray (Dataset C) can be downloaded from https://www.kaggle.com/nih-chest-xrays/data. Normalized log transformed expression matrices of datasets A1, A2, A3, B, D and E as used for the predictions are made available via FASTGenomics at https://beta.fastgenomics.org/p/swarm-learning.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the 12029 samples from data set A (AML), we followed work of Warnat-Herresthal et al, 2020, (doi: 10.1016/j.isci.2019.100780). Dataset B (Tb, 1999 samples) is a collection of all available PAX-based high-quality Tb datasets and controls on GEO. For COVID-19 in dataset D, the collection of 134 samples and 9 controls was driven by availability of consenting patients. For dataset E, the collection of 2400 samples was driven by availability of consenting patients. Dataset C has been compiled and published by the NIH CC and contains 112120 X-ray images. It is one of the largest community data sets and has been used in many studies.
Data exclusions	We used a minimum of five million aligned reads per samples to exclude low-quality samples from the Covid samples. This number is recommended as a minimum for bulk RNA sequencing, as e.g. stated by Illumina (https://support.illumina.com/bulletins/2017/04/considerations-for-rna-seq-read-length-and-coverage-.html)
Replication	The swarm learning approach has been successfully replicated in five data sets (A,B,C,D,E) with multiple permutations.
Randomization	The allocation into experimental group was determined by disease/condition and no other covariates were used. An additional experiment tested the impact of age, sex and COVID-19 diseases severity.
Blinding	Blinding was not applicable, since we collected pre-existing data sets. Additionally to guarantee independent sampling, we performed random permutations of training and test data sets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The Rhineland Study participants stem from an ongoing community-based cohort study in which all inhabitants of two geographically defined areas in the city of Bonn, Germany aged 30–100 years are being invited to participate. Persons living in these areas are predominantly German with Caucasian ethnicity. Participation in the study is possible by invitation only. The only exclusion criterion is insufficient German language skills to give informed consent. The COVID-19 samples are described in Supplementary Table 6.
Recruitment	The Rhineland Study is an ongoing community-based cohort study in which all inhabitants of two geographically defined areas in the city of Bonn, Germany, aged 30 years and above are being invited to participate. Persons living in these areas are predominantly German from Caucasian descent. Participation in the study is possible by invitation only. The only exclusion criterion is insufficient command of the German language to give informed consent. Therefore, given that participation in the Rhineland Study does not depend on any health-related outcome (e.g. the presence or absence of any particular lifestyle, disease or therapy), the potential risk of any selection bias impacting our results is, in all likelihood, very low. COVID-19 samples were collected based on availability. For all COVID-19 patients, the study was carried out in accordance with the applicable rules concerning the review of research ethics committees and informed consent. All patients or legal representatives were informed about the study details and could decline to participate. COVID-19 was diagnosed by a positive SARS-CoV-2 RT-PCR test in nasopharyngeal or throat swabs and/or by typical chest CT-scan finding.
Ethics oversight	Approval to undertake the Rhineland Study was obtained from the ethics committee of the University of Bonn, Medical Faculty. Collection of Covid19 samples was overseen by the research ethics committees at Radboud University Medical Centre in Nijmegen, the Netherlands (local ethics committee CMO Arnhem-Nijmegen, registration no. 2016-2923), and the Sotiria Athens General Hospital (Ethics Committee of Sotiria Athens General Hospital, IRB 23/12.08.2019) or the ATTIKON University General Hospital ((Ethics Committee of ATTIKON University General Hospital, IRB 26.02.2019) in Athens, Greece as well as the respective committees at the other sites: Kiel, Germany (COVIDOM, Ethics Committee of the University of Kiel, IRB D466/20), Saarbrücken, Germany (CORSAAR, Ethics Committee Medical Association of the Saarland, IRB 62/20, IRB 20200597), Munich, Germany (Ethics Committee of the LMU Munich, IRB 286/2020B01), Tübingen Germany (DeCOI Host Genomes, Ethics Committee of the Medical Faculty of the University of Tübingen, IRB 286/2020B01), Aachen, Germany (COVAS, Ethics Committee of the Medical Faculty of the Technical University Aachen, IRB 20-085), Cologne, Germany (Ethics Committee of the University of Cologne, IRB 20-1187_1) and Bonn, Germany (Ethics Committee of the Medical Faculty of the University of Bonn, IRB 073/19, 134/20). Dataset C is IRB approved (personal communication by Dr. Summers Senior Investigator, Clinical Image Processing Service, NIH CC).

Note that full information on the approval of the study protocol must also be provided in the manuscript.