

Discovering Latent Structure in High-Dimensional Healthcare Data:
Toward Improved Interpretability

I n a u g u r a l d i s s e r t a t i o n

zur

Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Ernst-Moritz-Arndt-Universität Greifswald

vorgelegt von
Ann-Kristin Becker

Greifswald, August 2021

Dekan: Prof. Dr. Gerald Kerth

1. Gutachter: Prof. Dr. Mario Stanke

2. Gutachter: Prof. Dr. Lars Kaderali

3. Gutachter: Prof. Dr. Holger Fröhlich

Tag der Promotion: 28.02.2022

OUTLINE AND SCOPE

This cumulative thesis describes contributions to the field of interpretable machine learning in the healthcare domain. In this thesis, three research articles are presented that lie at the intersection of biomedical and machine learning research. They illustrate how incorporating latent structure can provide a valuable compression of the information hidden in complex healthcare data.

The included articles provide insights into the entire process: from data processing, via method and workflow development, to analysis and interpretation of the results. Articles I and II include models learned from epidemiological data of different cohorts of the population-based Study of Health in Pomerania (SHIP). These data are extremely heterogeneous and multicollinear. In contrast, Article III is based on a study of proteomics, where data are rather homogeneous but the number of biological replicates is usually very low.

Article I: From heterogeneous healthcare data to disease-specific biomarker networks: a hierarchical Bayesian network approach.

The core result of this thesis builds on Bayesian networks, a type of probabilistic graphical model. In Article I, an approach to learn Bayesian networks in a group-based fashion is developed. The approach overcomes problems in Bayesian network structure learning that are raised by multicollinearity, heterogeneity, and dimensionality of the data. In a first step, heterogeneous features are hierarchically clustered around latent factors. The hierarchical structure is used as a basis throughout the learning process. It guides the optimization of the grouping with regard to an outcome of interest. An implementation of the method is available online at the Comprehensive R Archive Network (CRAN) from <https://CRAN.R-project.org/package=GroupBN>. The approach is tested on synthetic and example data and is finally used to analyze data from the SHIP cohort SHIP-Trend with regard to non-alcoholic fatty liver disease (NAFLD) and hypertension. The advantage of the approach is that the initially purely unsupervised model can be used with a varying resolution to account for an outcome of interest. The method helps to create concise yet flexible multivariate and disease-specific models of biomarker and risk factor interactions.

Article II: Discovering association patterns of individual serum thyrotropin concentrations using machine learning: An example from the Study of Health in Pomerania (SHIP).

Article II focuses on the prediction of individual serum thyrotropin concentrations, a key biomarker of thyroid function. Methodologically, we combine random forests with a post-hoc Bayesian network analysis for improved interpretation. As an ensemble method, random forests train highly flexible, nonparametric predictors. However, their interpretability is limited, as multiple decision paths of various trees are combined for prediction. That is why the interactions of those predictors that are identified as relevant from the random forest model are additionally analyzed in detail in a Bayesian network approach. They are finally discussed in the context of recent thyroid research.

Article III: **Metabolic cross-talk between human bronchial epithelial cells and internalized Staphylococcus aureus as a driver for infection.**

In Article III, temporal proteomic profiles from Staphylococcus aureus (*S. aureus*) and human bronchial epithelial cells (HBE) are measured to reveal insight into the metabolic cross-talk between bacteria and host. In this case, data analysis is complicated because proteomes may vary highly between cells and over time and they can be measured only indirectly via peptide abundances. For that reason, data are analyzed using protein-centered regression splines. Furthermore, interpretation is supported by time-series clustering, which reveals groups of proteins reacting similarly over time. The clustering also allows the identification of time points at which significant changes occur.

Methodologically, this thesis gives an overview of interpretable machine learning and the discovery of latent structure, including clusters, latent factors, graph structure, and hierarchical structure. Different methods are developed and applied to two main types of complex healthcare data (cohort study data and time-resolved molecular data). On the application side, we provide accurate predictive or discriminative models focusing on relevant medical conditions, related biomarkers, and their interactions. The presented models focus on non-alcoholic fatty liver disease (NAFLD) (Article I), hypertension (Article I), thyroid function (Article II), and host-pathogen interaction for *S. aureus* and host cells (Article III).

The thesis is structured as follows: Part [i](#) contains background information and summarizes the accomplishments. Here, chapter [1](#) outlines the opportunities and obstacles of machine learning in the healthcare domain. Chapter [2](#) introduces relevant methodological foundations and puts them into context. First, all included supervised models are defined and discussed in the context of interpretability. Afterward, different approaches to latent structure discovery are discussed. Chapter [3](#) shortly summarizes the overall results achieved within the included articles. Chapter [4](#) contains general conclusions. In Part [ii](#), the three research articles are presented. Parts [iii](#) and [iv](#) provide references and supplementary information.

CONTENTS

Outline and Scope iii

I RESEARCH SUMMARY

1	BACKGROUND AND MOTIVATION	1
1	Machine Learning in Healthcare	1
2	Major Healthcare Data Challenges	1
3	Benefits of Incorporating Latent Structure	2
4	The Role of Interpretability for Interdisciplinary Collaboration	3
2	METHODOLOGICAL FOUNDATIONS	5
1	Model Interpretability	5
2	Interpreting Supervised Models	6
2.1	Linear Models	8
2.2	Regression Splines	9
2.3	Decision Trees	10
2.4	Random Forests	12
2.5	Deep Neural Networks	12
3	Latent Structure Discovery	13
3.1	Discovering Clusters	14
3.2	Discovering Latent Factors	14
3.3	Discovering Graph Structure	16
3.4	Discovering Hierarchical Structure	22
3	RESULTS AND DISCUSSION	25
1	Employed Data	25
1.1	Cohort Study Data	26
1.2	Time-resolved Molecular Data	26
2	Methodology	27
2.1	Adaptive Refinement of Group Bayesian Networks (Article I)	27
2.2	Supporting Random Forest Interpretation with Bayesian Networks (Article II)	29
2.3	Combining Spline Models and Fuzzy Clustering (Article III)	30
3	Medical Conditions	31
3.1	Non-Alcoholic Fatty Liver Disease (Article I)	31
3.2	Hypertension (Article I)	32
3.3	Thyroid (Dys-)Function (Article II)	32
3.4	Staphylococcal Infections (Article III)	33

4 CONCLUSION 35

II THESIS ARTICLES

Author Contributions 39

Article I 43

Article II 65

Article III 93

III REFERENCES

Bibliography 123

List of Abbreviations and Symbols 126

Eigenständigkeitserklärung 127

List of Publications 129

IV APPENDIX

A ADDITIONAL NOTES ON BAYESIAN NETWORKS 133

1 Conditional Probability and Bayes Theorem 133

2 (Conditional) Independence 134

3 Factorization by Chain Rule 134

4 Graph Terminology 134

5 Bayesian Networks 135

6 v-Structures and d-Separation 135

7 Markov Properties 137

8 Faithfulness 139

9 Reading Conditional Independencies from a Network 140

10 Markov Equivalence 140

11 Bayesian Information Criterion and its Properties 141

B ADDITIONAL NOTES ON R-PACKAGE *groupbn* 143

Part I

RESEARCH SUMMARY

BACKGROUND AND MOTIVATION

1 MACHINE LEARNING IN HEALTHCARE

In recent years, the collection of high-dimensional healthcare data has become increasingly simple and cost-efficient. This development is accompanied by a strong need for improved and largely automated data analysis methods to extract hidden knowledge from complex data to make it accessible. Available classes of statistical models tend to focus on the accurate prediction of a particular outcome, and they differ considerably in terms of interpretability.

2 MAJOR HEALTHCARE DATA CHALLENGES

Healthcare data pose substantial challenges for machine learning, while the demands for model interpretability and explainability are much higher than in most other domains. Particular difficulties arise from the diversity of treatments and study designs, data types, and processing. Data may stem from various sources, including electronic health records, patient-reported data, medical imaging, molecular data, cohort studies, and large clinical trials. All of these data sources share a significant amount of measurement noise and uncertainty that is often as high as true effects. High noise may also be present in the labeling of a potential target variable, which is often based on a clinical diagnosis by a physician. Thus, human mistakes cannot be ruled out and come on top of the general uncertainty in diagnosis.

The integration of different data types causes further obstacles, including heterogeneity, incompleteness, imbalance, multicollinearity, and complexity, that inhibit a straightforward application of established machine-learning methods. Furthermore, relatively small sample sizes are a common limitation. At the same time, incorrect prediction comes at an extraordinarily high cost due to high-stakes applications like

clinical decision-making. Thus, healthcare models have to comply with high standards regarding interpretability, regulation, and accountability (Vellido, 2020; Erickson, 2021).

However, the concept of interpretability is complex, as it is closely tied to the human (thus subjective) capacity of understanding and lacks a universal, formalized definition and measure (Lipton, 2018). For example, the *key ethical requirements of EU guidelines on ethics in artificial intelligence* request it to “be possible for [AI systems and related human decisions] to be understood and traced by humans” (Madiega, Madiega). Consequently, a healthcare machine learning model needs to be both, an accurate depiction of reality (based on the given data) on the one hand and as interpretable as possible on the other hand. Only when a good tradeoff is found a model will be practical, accountable, and applicable.

3 BENEFITS OF INCORPORATING LATENT STRUCTURE

Interpretability can, for example, be achieved by choosing a model type, which is intrinsically interpretable, such as a linear model. More complex models can be made interpretable by visualization of model structures or by partitioning them into segments. However, even theoretically well interpretable models get hard to explain if they include a large number of features. For this reason, the approaches presented in this thesis include additional latent structure to reduce the complexity of the final models.

Typically, the first sub-task of model building from high-dimensional healthcare data is an attempt to reduce the number of involved features. The easiest way to streamline this step is to discard redundant or irrelevant features, known as feature subset selection (FSS). FSS can be realized by wrapper or filter strategies, or a feature subset can be selected manually based on prior knowledge (Jović et al., 2015; Hira and Gillies, 2015). This step, however, may have a high impact and lead to the loss of potentially useful information, for example, regarding feature interactions (Haury et al., 2011). As a result, previously unknown or overlooked relations may be discarded, and mutual effects may not be noticed. On the other hand, dimensionality reduction can also be achieved by agglomeration, projection, or transformation of the original features, known as *feature extraction* (FE) (Hira and Gillies, 2015). In this case, new informative and non-redundant features are created from the original data, yielding a compression of the feature space.

Indeed, complex systems are often organized according to a simpler, underlying structure, for example, in the form of clusters or modules (Murphy, 2012). *Latent structure discovery* describes the model-based identification of these underlying patterns or structures in observed data. Incorporating latent structures may help overcome problems raised by multicollinearity, heterogeneity, and complexity. Moreover, by capturing and denoising the main characteristics of the data, the latent structure may even improve discrimination or prediction (Zhou and Nakhleh, 2012). As deep latent structure models like deep neural networks tend to be black-boxes, we focus on identifying interpretable latent structures to offer a compressed representation of complex data and enhance interpretability.

4 THE ROLE OF INTERPRETABILITY FOR INTERDISCIPLINARY COLLABORATION

This thesis describes contributions to the field of latent-structure-based, interpretable machine learning in the healthcare domain. Three practical examples from medicine and healthcare are presented that give an overview of the whole process: from data preprocessing via model training and validation to the interpretation of the results. We show how latent structure-based models offer a flexible opportunity to extract compressed and understandable knowledge from complex data. They appear to be a good starting point for an in-depth interpretation if they are trained carefully and in a constant exchange between informaticians and biomedical researchers. For this step, an intuitive representation of the models is of great advantage.

As a whole, it can be said that AI models in healthcare can only be useful if they allow bridging back to the research questions and hypotheses that arose before data collection. Considering that, it gets clear that model interpretability is an essential characteristic.

METHODOLOGICAL FOUNDATIONS

This chapter gives an overview of related methodological foundations. After an introduction to the taxonomy of model interpretability in Section 1, all relevant supervised model types are briefly presented and discussed in the context of interpretability (Section 2). Section 3 gives an overview of different latent structure types: clusters, latent factors, graph structure, and hierarchical structure. As the core result of this thesis is based on the concept of Bayesian networks, a model to discover graph structure, this model family is introduced in greater detail.

1 MODEL INTERPRETABILITY

To interpret a machine learning model means finding meaning in it and being able to explain how the model works. If a model is based on multiple features, this includes explaining which features contribute to the model and how. Model interpretability is partly subjective as it refers to human understanding of the inner processes of a model. However, as explained above, this understanding is crucial in the application of models in high-stakes domains like healthcare.

In the literature, model interpretability is defined as the degree to 'which a human can understand the cause of a decision' or 'to which a human can consistently predict the model's result' (Linardatos et al., 2021; Masís, 2021; Molnar, 2020; Miller, 2019; Kim et al., 2016). It is, thus, not directly measurable by a metric, and tools for interpretation come in appreciably different forms: They vary from feature importance rankings via surrogate models to visual analytics (Linardatos et al., 2021; Liu et al., 2017; Vellido, 2020; Molnar, 2020).

However, interpretation methods can be vaguely classified according to their type and aim. For example, a machine learning model is called *intrinsically interpretable*, if

interpretability is achieved by restricting the complexity of the model structure. In contrast, so-called *black-box models*, use structures and functions that are too complex for humans to understand. This complexity often arises from models being highly recursive (Rudin, 2019). Such complex models can be made understandable by applying interpretation methods after training, which is called *post-hoc interpretability*. Post-hoc interpretation methods can, for example, be based on summary statistics of model predictions.

Interpretability methods exist for specific model types (*model-specific*) or they can be applicable to any model (*model-agnostic*). Furthermore, interpretation can either be performed on a local level of individual predictions (*local interpretability*) or averaged over entire models (*global interpretability*) (Molnar, 2020; Miller, 2019; Carvalho et al., 2019). For a sound introduction to the taxonomy of interpretability for machine learning, see for example Carvalho et al. (2019); Linardatos et al. (2021).

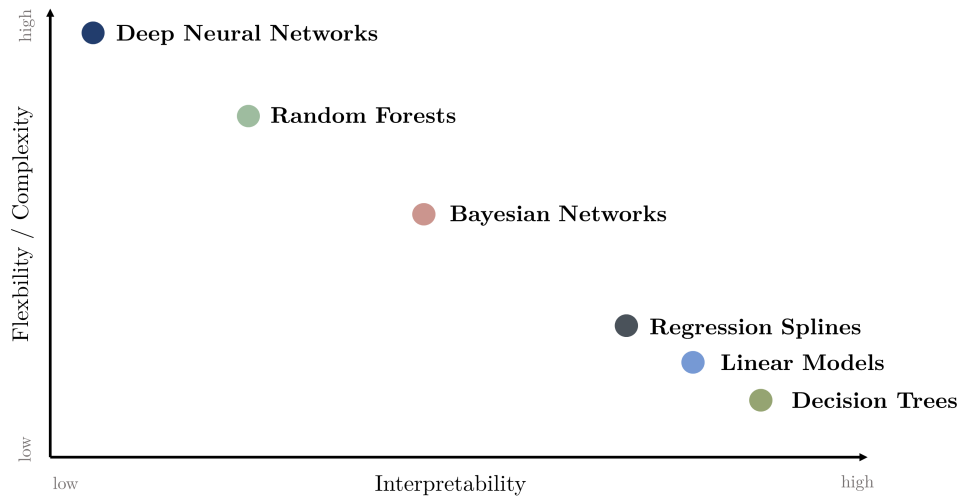
2 INTERPRETING SUPERVISED MODELS

If the data analysis focuses on discrimination or prediction of a certain medical condition, the applications fall under the task of *supervised learning*. The respective medical condition constitutes a *response* or *target variable*, that is usually encoded in the form of a random variable Y . The goal then is to learn a mapping from inputs $X = (X_1, \dots, X_n)$ (also denoted as *predictors* or *features*) to the response Y . Observed data \mathcal{D} are usually collected in a matrix with N rows $j = 1, \dots, N$ (*samples*). A systematic relationship f between X and Y is assumed

$$Y = f(X) + \varepsilon,$$

with ε denoting an independent, random error (*residuals*). If Y is continuous, the problem is referred to as *regression* (such as predicting the concentration of a certain metabolite in blood), in the discrete case as *classification* (such as predicting the presence or stage of a disease). The discrepancy between observed data and model predictions needs to be quantified in order to measure how well a model fits the data. Common measures are the *mean squared error* (MSE) for prediction or the *classification error rate* for classification problems. The shape of the mapping f encodes model assumptions and determines the model's complexity. From a probabilistic view, supervised learning

Figure 1: Tradeoff between interpretability and flexibility of different machine learning models. Figure adapted from (James et al., 2013).



refers to determining the conditional probability distribution $\mathbb{P}(Y|X)$.

Interpretability vs. Flexibility

It is commonly believed that there is a tradeoff between interpretability and performance, such that the restriction to interpretable model structures goes along with an impairment of predictive accuracy (Bratko, 1997; James et al., 2013; Masís, 2021). Although this may be theoretically true, it is in reality often possible to make a model easier interpretable while maintaining its predictive performance on unseen data (Krakovna, 2016; Rudin, 2019). This applies in particular to models with naturally meaningful features and limited data basis. Both is true for the projects discussed here. It is, in this case, more correct to call it a tradeoff between interpretability and flexibility (or complexity) of a model.

Available machine learning models differ significantly in the shape they offer to estimate the relation f between predictors and response. Very restrictive shapes of

f (e.g., linear models) are often ineffective on high-dimensional, complex real-world data, as the model assumptions are too harsh and the distribution they use is too inflexible. On the other hand, an increase in flexibility (e.g., achieved by using nonlinear functional relations or by aggregation of multiple models) may significantly increase a model's complexity, complicate its training, and impair its interpretability. Too flexible models will possibly lead to overfitting, in which case the predictive accuracy on unseen data will even decrease.

Fig 1 illustrates the tradeoff between interpretability and flexibility for those model types discussed within this section. It can be argued that the best-suited models are those that optimally tradeoff interpretability against flexibility. However, what the optimal tradeoff exactly means remains highly subject to the given constraints, and the specific application (Lipton, 2018; Hamon et al., 2020).

2.1 Linear Models

A linear model (LM) assumes the mapping f to be linear and, thus, represents the target as a weighted sum of the predictors

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon = \beta^T X + \varepsilon.$$

The residual error ε is usually assumed to be normally distributed around zero, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The most common way to estimate the parameters β is to compute the maximum likelihood estimate (MLE) using the normal equation $\beta = (X^T X)^{-1} X^T Y$. An example of a linear model is given in Fig 2.

The linearity assumption makes the model rather inflexible but leads to a straightforward interpretation, as linear and additive effects are easy to describe: If X_i is numerical, increasing it by one unit changes the outcome by β_i , given that all other features remain stable. Similarly, if X_i is binary, changing it from 0 to 1 changes the outcome by β_i . For improved interpretation, the influence of the weights can be visualized, e.g. in weight or effect plots. As the weights depend on the scale of features, it can be helpful to standardize features prior to model fitting or to scale the weight by its standard error after model fitting. Latter results in the t -statistic of the estimated weight. Moreover, the model as a whole can be interpreted using summary statistics, like R-squared (R^2), which is calculated as the proportion of variance in the data, that can be explained by the model.

The strong model assumptions, including homoscedasticity, independence, normality, and absence of multicollinearity are, however, often violated in reality. The interpretation of the model can then be inhibited and the manual processing load increases. For example, feature interactions would have to be handcrafted in form of interaction terms. In case of large feature sets, regularization (LASSO, ridge, elastic net) can be used to introduce sparsity to a model and make it easier to interpret (Madsen and Thyregod, 2010; Faraway, 2016; Murphy, 2012).

GENERALIZED LINEAR MODELS The framework of generalized linear models (GLM) allows the predictors to be related to the response via an additional *link function*, such that the residual distribution can be different from Gaussian and the model is applicable to other data types. Binomial residual distributions and the use of a sigmoid link function can be used for binary response variables (*logistic regression*). Coefficients must then be interpreted as odds ratios. For models of count events, residuals may be modeled by Poisson distributions (Madsen and Thyregod, 2010; Faraway, 2016).

2.2 Regression Splines

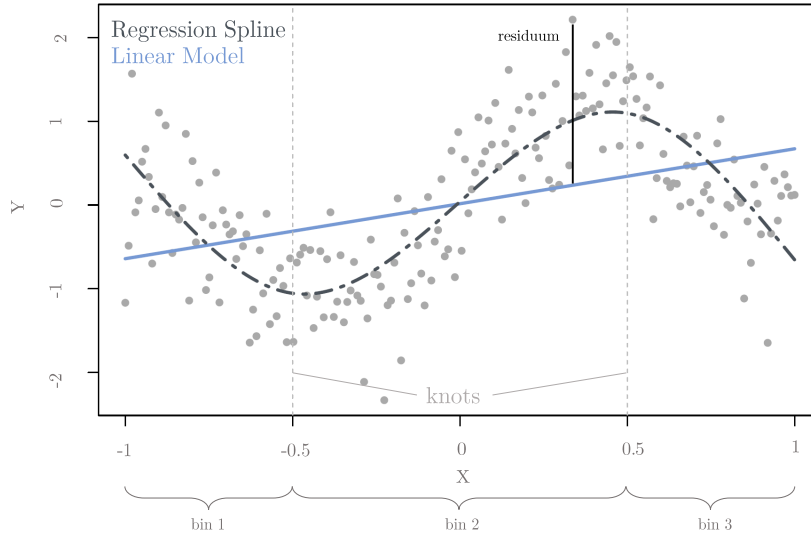
Regression splines do not impose a global structure on the data but divide them into K separate bins to fit linear or low-degree polynomial functions on each bin. To do so, the predictors are transformed using basis functions b_1, \dots, b_n ,

$$Y = \beta_{0k} + \beta_{1k}b_1(X_1) + \dots + \beta_{nk}b_n(X_n) + \varepsilon$$

and weights are determined for every bin $k = 1, \dots, K$ under additional continuity constraints, as illustrated in Fig 2. The cut points between the bins are called *knots*. Splines offer a possibility to smooth discrete input data and overcome hard constraints on the type of the relationship f .

However, the increase in flexibility goes along with a decrease in interpretability. The number and positioning of knots are additional hyperparameters that have to be chosen, and the number of weights gets multiplied by the number of bins in comparison to linear regression. Also, the influence of the weights depends on the types of the used basis functions. To improve the interpretability post-hoc, the behavior of a spline can be summarized, depending on the context of the analysis. Statistical

Figure 2: Linear model and regression spline fitted to synthetic data sampled from the functional relation $Y = 0.5X + \sin(4X) + \varepsilon$ with Gaussian residuals. Gray dots represent sampled data points. The solid blue line shows the fitted linear model ($R^2 = 0.3$) and the dashed gray line shows a natural regression spline ($R^2 = 0.65$) with 4 knots located at the boundaries as well as $x = \pm 0.5$ and respective 3 bins.



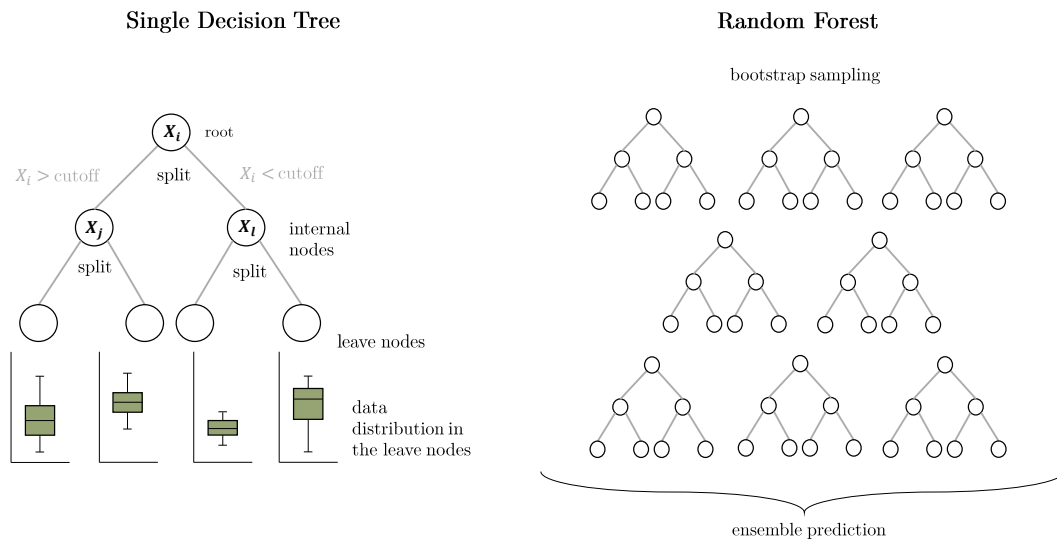
hypothesis tests for properties such as linearity or constancy of a spline can for example act as such summary metrics.

2.3 Decision Trees

Similar to the idea of regression splines, decision trees divide the data into distinct regions. By defining separate models on each of these regions, decision trees can adapt well to heterogeneous data structures. Usually, the mean response μ_k of data in a region R_k is used as prediction for each observation that falls in R_k . The resulting model is then piece-wise constant

$$Y = \sum_{k=1}^K \mu_k \cdot 1_{(X \in R_k)} + \varepsilon, \quad \text{where } K \text{ is the number of regions.}$$

Figure 3: Conceptual visualization of a single decision tree and a random forest



However, more complex leaf models are possible, e.g., linear or logistic models (Breiman et al., 2017; Landwehr et al., 2005). As finding the optimal partition into regions is an NP-hard problem, greedy approaches are used to do successive binary splits, leading to regions in the form of high-dimensional rectangles. The quality of a single split can be evaluated based on a cost function (e.g., Gini index, variance, entropy). Decision trees are able to model highly non-linear and complex feature relations.

The interpretative strength of decision trees lies in their visualization in the form of trees: Binary splits can be represented graphically as branches, with leaves representing the different regions (Fig. 3). Decision trees are popular tools in the healthcare domain, as their recursive structure is close to human decision-making strategies and makes it very easy to interpret and explain the model: For each prediction, there is a path from the root of the tree to a terminal leaf. Such a path consists of a series of decisions concerning only one specific feature.

However, single decision trees are very unstable (i.e., sensitive even to small changes in the input data), and they tend to score poorly for complex problems (Murphy, 2012).

2.4 Random Forests

Random forests (Breiman, 2001) are an ensemble learning method based on decision trees: Instead of training only a single tree, a multitude of T tree models $f_t(x)$ is learned from random sub-samples of the training data and the predictions of the individual models are averaged (known as *bootstrap aggregation* or *bagging*),

$$f(x) = \frac{1}{T} \sum_{t=1}^T f_t(x).$$

This approach helps to decrease the variance of the final estimate. In order to decorrelate the individual trees, for each split, not all but only a random subset of predictors is considered as candidates for splitting. In addition to the described frequentist approach, Bayesian approaches are available to learn single trees (e.g., using MCMC (Wu et al., 2007)) or ensembles of trees (e.g., BART (Chipman et al., 2010)).

Random forests are known to score strongly for very complex and diverse problems (Caruana and Niculescu-Mizil, 2006; Muchlinski et al., 2016; Santhanam et al., 2020). However, it is no longer possible to represent the resulting model as a single tree, so visualization is difficult, and their interpretation gets complicated: Due to the aggregation, random forests do not belong to the class of intrinsically interpretable models. Various heuristic post-hoc measures and model-agnostic approaches are available to interpret a model. Overall feature importance can, for example, be calculated by measuring the decrease in cost due to splits of a specific feature (Breiman, 2001; Fisher et al., 2018). Feature importance may also be computed for feature groups (Wehenkel et al., 2018). However, these measures often suffer from biases if features are heterogeneous or collinear. SHAP values (Lundberg et al., 2018) or surrogate models (Ribeiro et al., 2016; Molnar, 2020) are other interpretation methods that are often applied to random forest models for local interpretation.

2.5 Deep Neural Networks

The previously discussed models are based on different forms of direct mappings from X to Y . In contrast, *deep neural networks* refer to models with many hidden layers in between X and Y that learn representations at increasing abstraction levels, inspired by the visual cortex (Serre et al., 2005; Ranzato et al., 2009). Deep neural network

models tend to have millions of parameters that acquire a high amount of labeled data to be properly trained. Moreover, due to their complexity, they are usually black-box models with the need for further post-hoc methods to achieve interpretability. Suitable measures are, for example, summarized by Montavon et al. (2018) but imply their own difficulties, as they might not be faithful to the model they try to explain (Rudin, 2019; Linardatos et al., 2021).

Deep neural network models perform exceptionally well in image classification, language processing, and similar approaches, where they detect underlying patterns through their complex and recursive latent structure. However, in the case of limited, structured data, they do not tend to be clearly superior in terms of prediction (Caterini and Chang, 2018). Additionally, in all presented healthcare applications, the number of features is very high in relation to the sample size. This dimensionality problem further complicates the training of models with a high number of parameters. To this end, we focus mainly on intrinsically interpretable models, as they are advantageous under the described conditions. Specific models for the detection of simpler and interpretable latent structures are introduced in the following Section 3. For a further introduction to deep learning, see for example Skanski (2018); Caterini and Chang (2018); Maier et al. (2019).

3 LATENT STRUCTURE DISCOVERY

Latent structure discovery refers to the discovery of patterns within a data set (Murphy, 2012). Instead of a mapping f from input X to output Y as before, we aim for a latent structure Z and its relation to the features X . Here, Z is a latent variable that was never explicitly observed, so the task belongs to the category of unsupervised learning. In probabilistic terms, the aim is to describe the unconditional distribution $\mathbb{P}(X)$ in terms of Z .

Latent structure discovery can be coupled with supervised learning, such that discovered latent structures are used as input data for prediction or discrimination tasks afterwards. Depending on the shape of Z , different types of latent structures can be discovered.

3.1 *Discovering Clusters*

Clustering describes the approach to discover homogeneous subgroups within data. It can denote subgroup discovery in the space of observations as well as features. Clusters can be defined in different ways. The most common definition of a cluster is that data points within clusters have small distances to each other. Clustering is one of the simplest forms of latent structure models, as it includes only one single, discrete latent variable Z that assigns an integer $1, \dots, K$ representing the cluster number. A formal definition of model-based clustering can be constructed by the use of finite-mixture models. Mixture models assume the joint distribution to be a convex combination of K base distributions

$$\mathbb{P}(X) = \sum_{k=1}^K \pi_k \mathbb{P}_k(X).$$

Each base distribution \mathbb{P}_k corresponds to one cluster and the mixture weights π_k determine the proportion of the clusters. In the case of Gaussian base distributions, the approach results in a Gaussian mixture model (GMM). The clustering can then be reconstructed by the use of an expectation-maximization (EM) algorithm.

k-means is the most wide-spread variant of the EM-algorithm for GMMs. Its assumptions include fixed covariance $\Sigma_k = \sigma^2 I$ and fixed mixture weights $\pi_k = \frac{1}{K}$, so that only the cluster means are estimated. Evaluation measures of the quality of a clustering include *purity*, *Rand index* or *mutual information* (Hartigan and Wong, 1979; Murphy, 2012). Clusters may be represented by a representative, e.g., the cluster centroids, thereby reducing the dimensionality of the data and supporting interpretation.

Fuzzy or *soft-clustering methods* return membership grades between 0 and 1 instead of a hard cluster assignment. The *fuzzy c-Means* algorithm is an extension of *k-means*, in which such membership scores are used as additional weights. Consequently, the approach is less sensitive to outliers and noise (Dunn, 1973; Bezdek, 2013).

3.2 *Discovering Latent Factors*

In clustering, the latent structure consists of the single, discrete variable Z . As an alternative, Z can be introduced as a random vector $Z \in \mathbb{R}^M$, referring to M *latent factors* (Murphy, 2012). In that case, the idea is to discover a lower number $M \leq n$ of latent factors that approximate the original data and take interrelations among features

into account. The projection of the data into the M -dimensional subspace leads to a *dimensionality reduction*.

Principal component analysis (PCA) is a latent linear model and the most common approach for dimensionality reduction (Murphy, 2012; Bro and Smilde, 2014). PCA identifies orthogonal linear combinations of variables along which most variation occurs and, thereby, allows to summarize a large set of (potentially correlated) numerical variables by a smaller number of representative variables, that collectively explain most of the variability. The first *principal component* of a set of features X_1, \dots, X_n is the normalized linear combination

$$Z_1 = \varphi_{11}X_1 + \dots + \varphi_{n1}X_n$$

that has the largest variance. $\varphi_{11}, \dots, \varphi_{n1}$ are called *loadings* of the first principal component, and they are constrained to sum up to one (Murphy, 2012; Chavent et al., 2017; James et al., 2013; Bro and Smilde, 2014).

The optimization problem of finding this linear combination can be solved via eigen decomposition of the covariance matrix. Similarly to the first, the m th component can be determined by subtracting the first $m - 1$ principal components from X , followed by determination of the weight vector which extracts the maximum variance from the remaining data matrix. Consequently, the first m principal components define the m -dimensional subspace that is the closest to the data in terms of average squared Euclidian distance. Thus, principal component analysis offers an approximation of the data based on orthogonal factors Z :

$$X = Z\Phi^t + \varepsilon,$$

where Φ is the matrix containing all loadings $\varphi_{im}, 1 \leq i \leq n, 1 \leq m \leq M$ and ε denotes the residuals. The principal components describe the directions in which the data vary most. They can be used to produce low-dimensional visualizations of high-dimensional data. Further, the loadings themselves can be used to explain the influence of the respective variable on the principal component.

A description of probabilistic PCA can be found in Murphy (2012). A counterpart of PCA for categorical data is called *Multiple Correspondence Analysis* (MCA) (Kiers, 1991; Greenacre and Blasius, 2006). Several implementations of a combination of PCA

and MCA for mixed data have been developed, for example via Generalized Singular Value Decomposition as described by Chavent et al. (2017).

3.3 *Discovering Graph Structure*

If the focus lies mostly on the interactions of the variables $X = (X_1, \dots, X_n)$, it can be useful to approach the joint distribution $\mathbb{P}(X_1, \dots, X_n)$ with the aid of sparse graphs. In the following, a brief introduction to the field of graphical models is given. Note that Appendix A provides additional information on the mathematical fundamentals.

Formally, two random variables are (conditionally) independent if their (conditional) joint distribution factorizes (see Appendix A1-2 for a detailed definition). In *probabilistic graphical models* (PGM) each random variable X_i is associated with a node i in a graph and edges or their absence represent conditional (in-)dependencies. Thus, conditional independence can be interpreted as separation of nodes in the graph. PGMs allow to visualize inter-dependencies and perform probabilistic inference through graphical manipulations. Due to their clear graphical model interface, PGMs achieve interpretation through visualization. PGMs are advantageous for healthcare problems in many ways: They are statistically grounded, they allow to inherently model uncertainty or missingness and its effects, they provide a framework for probabilistic reasoning, and they have been very actively evolved and adapted to many research questions since their introduction in the 1980s (Pearl, 1985).

A *Bayesian Network* (BN) is a probabilistic graphical model on a *directed acyclic graph* (see Appendix A4). A BN consists of the graph structure G (with nodes and edges) and a parameter set Θ . It has some desirable properties, including the directionality, which may be interpreted as causal, and a hierarchical order. According to this order, neighbors of a node are denoted *parents* and *children* (or *descendants* and *ancestors* if the connection is indirect). In general, the connection between nodes is called *path*. Fig 4A shows an example of a BN structure. Fig 4B and C show two paths consisting of two edges.

A BN encodes the *local Markov property*: each variable X_i is assumed to be independent of its non-descendants conditioned on its parents, $\text{par}(X_i)$, in the graph. This

property allows to effectively factorize the joint distribution of the random vector X according to the graph structure G as follows

$$\mathbb{P}(X) = \prod_{i=1}^n \mathbb{P}(X_i | X_{\text{par}(i)}).$$

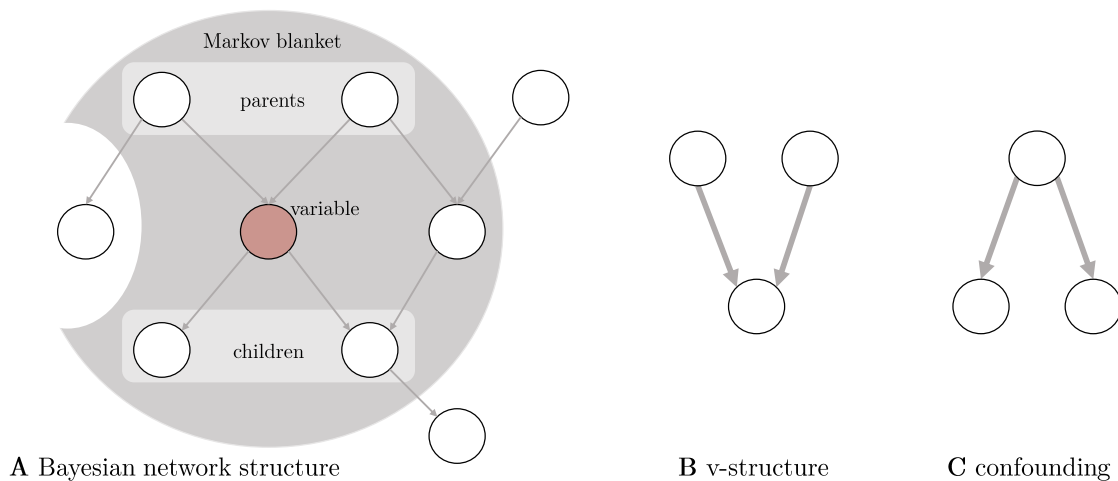
Consequently, the parameter set Θ of a BN consists of the conditional probabilities of all nodes given its parents

$$\theta_i = \mathbb{P}(X_i | X_{\text{par}(i)}).$$

The factorization supports model interpretation, as it allows to transition from the global structure to local (in)dependencies. For a thorough introduction to the Markov property, see Appendix A7.

In the network structure presented in Fig 4A, also the *Markov blanket* of the center node is highlighted. The Markov blanket is a popular concept for feature selection. Conditioning on all nodes that are part of its Markov blanket makes a node independent from all remaining nodes. It can be shown, that the minimal Markov blanket of a node in an BN includes its parents, children and parents of children (Koller and Friedman, 2009), as visualized here.

Figure 4: A) A Bayesian network structure. The parents, children and Markov blanket of the center node are highlighted. B) A network constellation of three nodes and two edges known as *v-structure*. C) A network constellation of three nodes and two edges known as *confounding*.



In a Bayesian network, all independence statements encoded by a joint distribution and the separation of nodes in the graph need to be equivalent. In case of directed acyclic networks the suitable definition of graph separation is called *d-separation* (for directed separation). Two nodes are denoted as d-separated if every *path* between them is *blocked*. A path is called *blocked* in two scenarios: it contains a *v-structure* (Fig 4B), where the central node or its descendants are not observed. Or it contains any other path type (e.g. confounding; Fig 4C) and the central node is observed. The concept of d-separation allows to read probabilistic dependencies directly from the graph. For a more detailed description of d-separation and the validity of theoretical conditions, see Appendix A6-8.

Learning Bayesian Networks

The inference of Bayesian networks from data \mathcal{D} usually happens in two steps in a Bayesian fashion: First a graph structure G is determined that optimally encodes the dependence structure that is present in the data. Then, parameters Θ are estimated given the previously determined graph structure and according to the following Bayesian factorization

$$\mathbb{P}(G, \Theta \mid \mathcal{D}) = \mathbb{P}(G \mid \mathcal{D}) \cdot \mathbb{P}(\Theta \mid G, \mathcal{D}).$$

STRUCTURE LEARNING Bayesian network structure learning is an NP-hard problem (Chickering et al., 1994). There are two main strategies that algorithms pursue (Koller and Friedman, 2009).

Constraint-based algorithms use statistical independence tests to determine conditional dependence structures. They usually start with a complete graph and exclude edges between variables that were found to be conditionally independent. The type of conditional independence test depends mostly on the type of random variables. Common choices include mutual information, correlation coefficients or Fisher's Z test. Constraint-based algorithms are known to be more accurate on small data sets. However, in case of high-dimensional, heterogeneous and noisy data, they often produce networks that are not well connected (Scutari et al., 2019). Then, a propagation of evidences is difficult.

Score-based algorithms handle the problem using a search-and-score approach. They try to optimize the data goodness-of-fit $\mathbb{P}(G \mid \mathcal{D})$ by applying local or global, exact or approximate optimization techniques. Available scores usually penalize the goodness-

of-fit for the complexity of the network. For example, the Bayesian information criterion (BIC) is defined as

$$\text{BIC}(G|\mathcal{D}) := \log \mathbb{P}(\mathcal{D}|G) + \frac{d}{2} \log(N),$$

where d the number of free parameters and N is the number of samples. The BIC is a common choice, as it is *locally and asymptotically consistent*, available in closed form and does not include any hyperparameters (see Section A11). Bayesian scores are also available (Heckerman et al., 1995). In practice, relatively simple score-based algorithms tend to be faster than constraint-based approaches and usually lead to high-likelihood networks with a larger number of edges, that allow for better evidence propagation (Scutari et al., 2019). Moreover, *hybrid approaches* are available that couple both strategies.

Structure learning algorithms are often used in combination with model averaging techniques to minimize noisy, false positive edges (Koller and Friedman, 2009). *Model averaging* describes an approach to identify statistically significant edges. It includes the determination of several model structures based on bootstrap samples of the original data. A *confidence score* for the presence of an edge can be estimated as the proportion of bootstrap models in which the edge is present (Scutari and Nagarajan, 2013). The confidence score can be used to distinguish significant from incidental edges. Unlike for random forests, the bagged models can be integrated into one summarized Bayesian network relatively easily, so that its interpretability is not inhibited.

PARAMETER ESTIMATION Due to the disentangled factorization of the joint distribution, maximizing the total likelihood is equivalent to maximizing each local likelihood separately. That is why, unlike structure learning, the estimation of Bayesian network parameters is straightforward and can happen for each node in parallel. For discrete nodes, the maximum likelihood estimates (MLE) of the conditional probabilities can be calculated based on relative frequencies. Bayesian parameter estimation allows to include a prior distribution for each node and may help to prevent overfitting. It especially prevents an estimation of probability zero for realizations that do not appear in the training data.

Inference in Bayesian Networks

Once determined, a Bayesian network can be used to answer probability queries. As shown, the Bayesian network provides a factorization of the joint probability

distribution. To answer a query means to determine the posterior distribution over the values of one or more query variables Y conditioned on some evidence E

$$\mathbb{P}(Y|E = e) = \frac{\mathbb{P}(Y, E = e)}{\mathbb{P}(E = e)}.$$

Both parts of the fraction can be determined by summing out all entries in the joint distribution for which $E = e$ (and $Y = y$, respectively). Unfortunately, the summation over the joint distribution fastly gets exponentially complex. Thus, it can be shown that both, exact as well as approximate inference are NP-hard problems (Cooper, 1990; Dagum and Luby, 1993). However, inference can be effectively implemented by stochastic simulation, e.g., using likelihood weighting (Koller and Friedman, 2009). In likelihood weighting, the prior distribution given by the Bayesian network is used to weight the importance of generated samples. Evidence nodes are fixed and all remaining nodes are sampled from the network. Finally, each sample is additionally weighted by the local probability of the evidence nodes (Yuan and Druzdzel, 2006).

Bayesian Network Classifiers

Bayesian network classifiers (BNC) make use of the theory of BNs in order to determine the conditional probability distribution of a response given input data $\mathbb{P}(Y|X)$. The conditional distribution can be determined by treating X as evidence. In this case, a network structure is usually fixed beforehand. The simplest Bayesian network classifier is the naive Bayes (NB) model. The NB model assumes that all features are conditionally independent given the response variable. The fixed structure defines accordingly the response variable as parent of all other features. Using Bayes' rule, the conditional distribution factorizes to the simple expression

$$\mathbb{P}(Y|X) = \mathbb{P}(Y) \cdot \mathbb{P}(X_1|Y) \cdot \mathbb{P}(X_2|Y) \cdots \mathbb{P}(X_n|Y).$$

Even though the strong and oversimplified assumptions are often violated, the naive Bayes classifier usually yields relatively good results (Rish et al., 2001; Davies, 2017). However, due to its a priori fixed dependence structure it does not reveal any information about the actual structure, similar to linear models. It also assumes all features to be related to the response variable, necessitating prior feature selection strategies in case of high-dimensional data. There are further approaches of less restrictive classifier model structures that allow additional dependencies among predictors. They are

reaching from tree-based naive Bayes models via k -dependence models to unrestricted Bayesian networks (Rubio and Gámez, 2011; Koller and Friedman, 2009). In principle, every trained Bayesian network structure can be used to predict or classify a chosen response variable using the inference techniques explained above. The approach of learning a nonrestrictive model first and deducing a conditional distribution of the response afterwards can act as regularization and prevent overfitting.

Interpretation of Bayesian Networks

Bayesian networks offer all possibilities for detailed interpretation on a global and on a local level, as they are intrinsically interpretable, yet highly flexible. The largest advantage is the implicit visualization in terms of a network structure that is supportive of human understanding. Additionally, due to the Markov property, it is possible to deduce from local to global structures and back. Further, an important feature of BNs is that conditional independence properties of the joint distribution can be read directly from the graph (see also Appendix A9). However, as for any model, the inner workings of very large networks may be complicated. That is why, several supportive explanation methods have been developed in order to support the interpretation of model output from large models, e.g. by visualizing information flow, by support graphs, or even by output of natural language explanations (Kyrimi et al., 2020; Lacave and Díez, 2002; Timmer et al., 2017; Hennessy et al., 2020).

EDGE DIRECTIONALITY AND CAUSAL INTERPRETATION A Bayesian network model encodes a set of probabilistic dependencies. It is well known that, in general, observational correlation does not imply causation. Similarly, based on purely observational data, the directions of edges in a Bayesian network can not be fully determined, as several differing structures encode the exact same probabilistic dependencies. Such BNs are denoted *Markov equivalent* and share their *skeleton*, which is the undirected graph that remains if all edge directions are ignored.

However, not all networks with the same skeleton are Markov equivalent. In certain scenarios, the directions of edges can be correctly inferred from observational data. That is only the case if more than two variables are involved. The central structure for inferring directions is the previously described v-structure (Fig 4B; also known as common effect, head-to-head node, or collider).

It can be shown, that networks must share their skeleton and additionally they must share the set of v-structures to be Markov equivalent (see Appendix A10).

Consequently, edge directions in a BN can be specified if they are part of a v-structure but not otherwise. To make this clear, BN models are often represented as partially directed networks for further interpretation, where only edges that are part of a v-structure are directed. These graphs are denoted as *completed partially directed acyclic graphs* (CPDAGs). In a large network, usually, multiple v-structures are present so that the direction of multiple edges can be inferred. If possible, the remaining edge directions can be determined by additional intervention experiments.

Furthermore, for an interpretation in terms of causal dependence, the *causal sufficiency assumption* must hold. It says that no latent variables exist (in the domain) that are a parent of at least one observed variable.

3.4 Discovering Hierarchical Structure

Hierarchical models explain complex patterns of observed data in terms of a latent hierarchy of successively higher-level, abstract units. Inferring hierarchical latent structure offers a way to overcome the preliminary specification of the number of clusters, and its nested structure may be close to the generative process of many real-world data.

HIERARCHICAL CLUSTERING Hierarchical clustering is a means to learn not one but multiple, nested partitions that are hierarchically linked. Most hierarchical clustering algorithms are deterministic. They take as input a dissimilarity matrix $D \in \mathbb{R}^{n \times n}$ with d_{ij} measuring the pairwise dissimilarity of X_i and X_j . A hierarchical structure is then inferred either *agglomeratively* (bottom-up) or *divisely* (top-down) (Murphy, 2012). The merges and splits are in general determined using greedy algorithms. The result is a *dendrogram*, that is a binary tree with leafs denoting the observed features. Similar features fuse ascendingly into branches. For any two features, the height of fusion represents a degree of (dis-)similarity. By cutting the dendrogram at a certain height, a hard clustering of the features with arbitrary cluster number is achieved. Conceptually, there is still only one latent variable Z . Classes at different levels of the hierarchy correspond to states of this variable at different resolution levels.

It can be useful to aim for a clustering around latent factors (Vigneau and Qannari, 2003). Cluster similarity is then defined based on the joint linkage to such a central latent factor. In this case, latent factors define natural cluster representatives, and the participation of features or samples in the cluster can be analyzed by their closeness

to the cluster center. The R-package *ClustOfVar* (Chavent et al., 2012) implements this approach for feature clustering of mixed categorical and continuous data.

Hierarchical Bayesian Networks

Hierarchical Bayesian networks (HBNs) are Bayesian networks with rooted tree structure, where all internal nodes refer to latent variables (Gyftodimos and Flach, 2002; Njah et al., 2019). In case of discrete data, they are known as hierarchical *latent class models* (LCMs) (Zhang, 2004). They allow for multivariate clusterings but are difficult to train, due to their potentially deep structure and the high amount of parameters. Moreover, the presence of (a high number of) latent variables inhibits the factorization of the joint distribution. It can be shown that HBNs can under some assumptions be transformed into deep neural networks by reparameterization, so that gradient-based inference is enabled (Kingma and Welling, 2014).

RESULTS AND DISCUSSION

The projects that are part of this thesis combine the introduced methods in different ways in order to extract knowledge from complex healthcare data. Table 1 gives an overview of the employed data, the applied methodology, and the medical conditions on which the projects focused.

		Article I	Article II	Article III
Employed Data	Cohort Study Data Time-res. Molecular Data			
Methodology	Regression Splines Random Forests Bayesian Networks PCA (Hierarchical) Clustering Time-Course Clustering			
Medical Condition	NAFLD Hypertension Thyroid (Dys-)Function Staphylococcal Infections			

Table 1: Overview of the employed data, the applied methodology and the medical conditions on which the thesis projects focused.

1 EMPLOYED DATA

Biomedical and healthcare data may be gathered from a vast amount of sources. Two common types of high-dimensional healthcare data are discussed here.

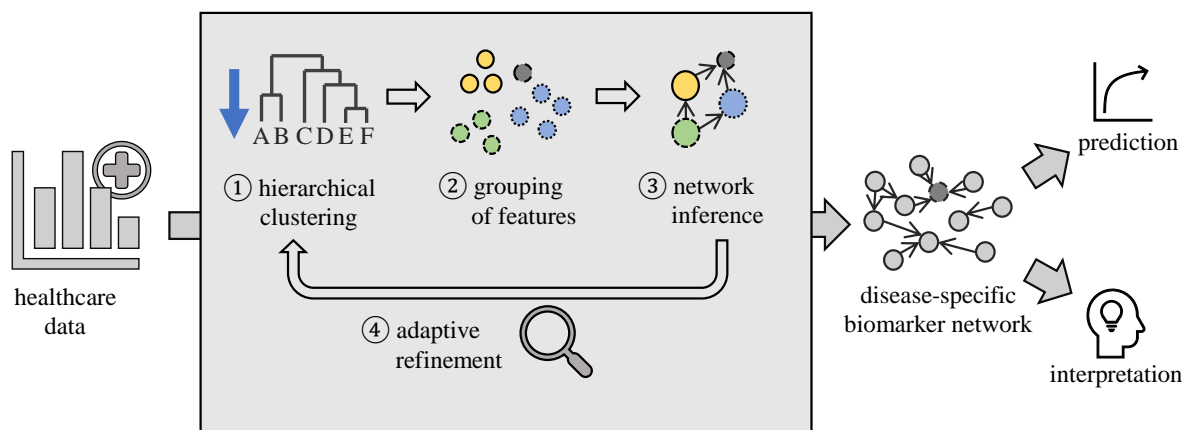
1.1 *Cohort Study Data*

Cohort studies represent a fundamental study design of epidemiology and are a particular form of longitudinal studies. A cohort, which is a group of people sharing a characteristic, such as a demographic similarity, is recruited, sampled, and followed over time. Article I and II deal with data from two different cohorts of the Study of Health in Pomerania (SHIP), which is a population-based cohort study (Völzke et al., 2011). Objectives of SHIP were the provision of prevalence estimates on a broad range of diseases, risk, and health factors for the population of Western Pomerania. Study probands were randomly selected from communities in the region, stratified by age and gender. In Article I, SHIP-TREND data are used. The SHIP-TREND cohort comprises 4420 participants, and examinations were performed between 2008 and 2012. The research in Article II is based on data from the first cohort of SHIP (1997-2001) with 4308 participants. Both data sets are characterized by a wealth of conducted examinations, including nutritional patterns, complete blood counts, sociodemographic data, health status, mood, medication, electrocardiography, echocardiography, sonography, neurological screening, blood and urine sampling, and whole-body MRI scans. The resulting feature set is broad and heterogeneous, with a high amount of intra-individual variation. The broadness allows for the screening of disease-specific patterns while taking numerous potential confounders into account. For both cohorts, 5-year follow-ups were performed that are not included in the discussed analyses. However, the availability of these data opens up the possibility of expanding the presented models by a temporal component in the future.

1.2 *Time-resolved Molecular Data*

Article III deals with a second common healthcare data type, which is time-resolved molecular data. This type of data usually belongs to 'large p , small n '-class, where the number of features greatly exceeds the number of samples. In the presented study, the abundance of $n_h = 3644$ human and $n_{sa} = 930$ staphylococcal proteins was measured over time, while the number of biological replicates was $p = 4$ in both cases. Mass spectrometry was used to quantify peptide abundances using a previously established library. Samples were collected with high temporal resolution up to 4 days post infection with the aim of deciphering the interplay of human bronchial epithelial cells and internalized bacteria.

Figure 5: Outline of the proposed approach to adaptive refinement of group Bayesian networks. Features of the input data are grouped using hierarchical clustering, then a group Bayesian network is learned. Based on the accuracy of the resulting model, the grouping is refined adaptively downwards along the dendrogram. The output is an interpretable disease-specific biomarker network based on feature groups, which has high predictive accuracy. This figure was copied from Becker et al. (2021).



2 METHODOLOGY

The methodological approaches differ between the presented articles depending on data type and goal of the analysis.

2.1 Adaptive Refinement of Group Bayesian Networks (Article I)

In Article I, a workflow to unravel latent interaction networks among feature groups is established. Its steps are shown in Fig 5. The analysis aimed at identifying biomarker and risk factor interactions of specific diseases from high-dimensional cohort study data. Demands on the methodology included the ability to extract relevant predictors and an understandable representation of interactions between predictors. Moreover, the prediction of the outcome should be comparable from different instantiated features based on the same model. Bayesian networks comply with all of these requirements and were chosen as the desired model type.

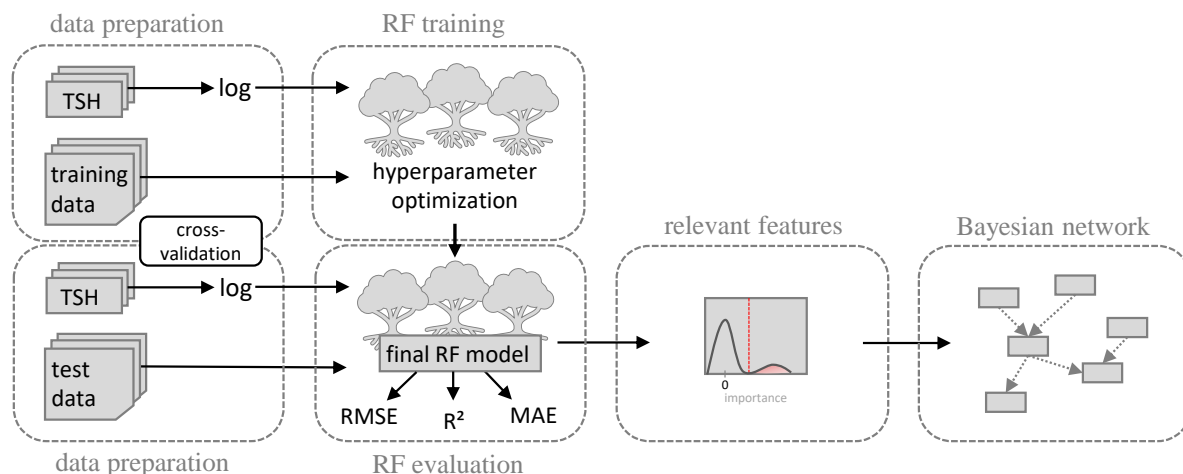
The presented workflow addresses multiple problems: First, Bayesian network structure learning from high-dimensional data is computationally intractable. Available structure learning algorithms tradeoff the model complexity (determined by the number of parameters) with the fit to the data in a generative learning approach. This generative approach is usually quite robust to overfitting but in the case of high-dimensional data, a potential outcome variable may get too much out of focus.

Moreover, throughout the learning process, every possible arc is equally important; thus, the strongest interactions are inserted with priority as they increase the measure of fit the most. In the presence of multicollinearity and modularity, the strongest interactions lie within modules of highly similar variables. That leads to a situation where arcs within modules are learned preferentially and arcs among modules are ignored. Due to these computational properties, resulting networks consist of many disconnected modules. The separation into modules inhibits the propagation of knowledge through the network and may be misleading for interpretation.

In order to overcome these drawbacks, a group-based network learning approach is developed in Article I. In a first step, feature groups are identified and aggregated. A Bayesian network structure is then learned among those groups. By this, the complexity of the model is significantly reduced.

However, in broad data sets like SHIP, which contain information on various diseases and disease states, not one but multiple clusterings may be realistic. The choice of the optimal grouping depends highly on the focus of the analysis. That is why we propose an iterative approach for automatic adaptive refinement that is based on hierarchical feature clustering (Fig 5). The approach uses a hierarchical structure as a basis to optimize the grouping with regard to the outcome of interest. An implementation of the method in R was published on CRAN (Becker and Kaderali, 2020). An overview of the software package can be found in Appendix B. The approach allows to automatically model essential parts of the network in greater detail, while others stay aggregated. The resulting group networks are easily interpretable, as the networks are concise and modules are directly visible. Groups themselves can be analyzed using tools known from PCA, as groups are centered around and represented by their first principal components. In Article I, we show that the resulting group networks also achieve a better prediction on unseen data, than detailed Bayesian networks, logistic regression, and established biomarker scores.

Figure 6: Supporting random forest interpretation with Bayesian Networks: After data preparation, a RF model is trained using nested cross-validation. Relevant predictors are identified based on two feature importance measures and a mixture model approach. Lastly, feature interactions among the relevant predictors are examined in a Bayesian network analysis. This figure was copied from Becker et al. (2021).



2.2 Supporting Random Forest Interpretation with Bayesian Networks (Article II)

In Article II, the proposed workflow consists of random forests and Bayesian networks and is represented in Fig 6. We start by training a full-featured random forest regressor. Random forests internally perform a feature selection while trees are grown. We use two common feature importance measures to identify features that contributed to the final model, an external (*incremental MSE*) and an internal one (*node purity*). These measures can be used to produce a ranking. However, both have shortcomings, especially in the presence of heterogeneity and multicollinearity. Node purity may be biased towards features with many categories or continuous features, as for those, potential splits can happen more flexibly than for binary features. The incremental MSE may especially produce misleading results when features are highly correlated, as it includes the permutation of one feature alone, which might then produce unrealistic data instances.

We thus complement the random forest model by a more detailed Bayesian network analysis. A Bayesian network structure is learned among all features that contributed to the random forest model. Due to the high dimensionality of the data, it could be expected that this applies only to a minority of the features. After identification of

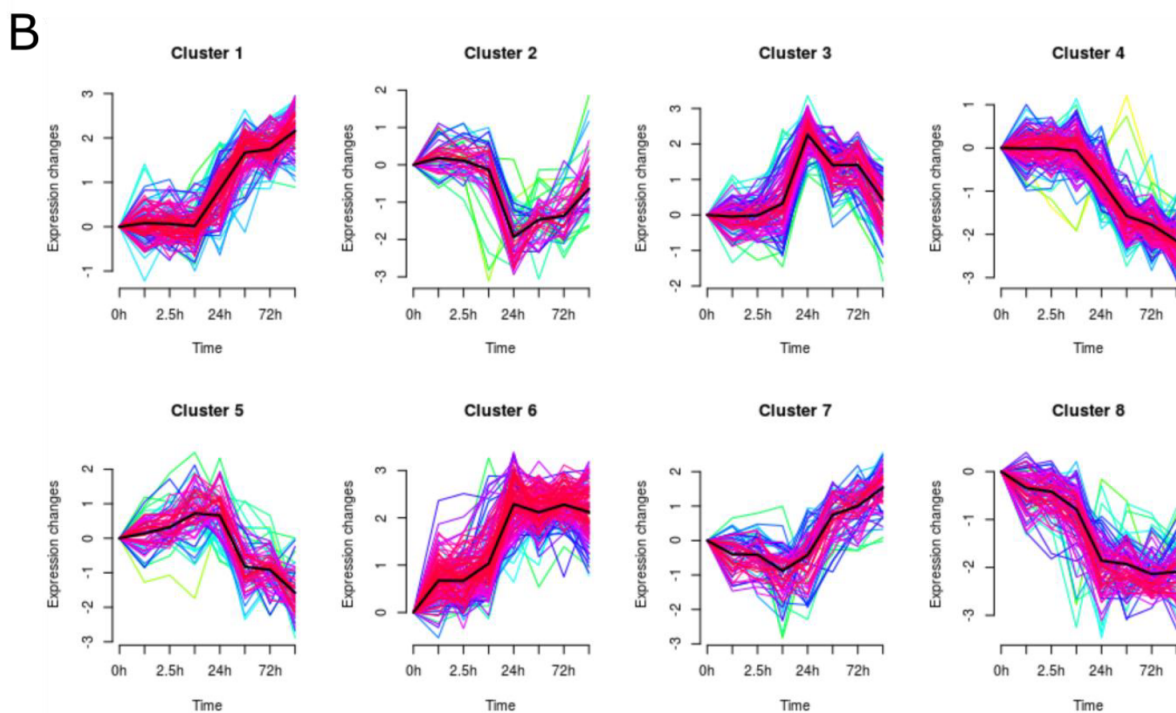


Figure 7: Resulting clusters from time course clustering of bacterial protein abundances. This figure was copied from Palma Medina et al. (2019).

this feature subset, highly similar features were grouped, similar to the procedure in Article I. Then, a network structure was learned. The network structure helps to distinguish between direct and indirect influences and is used to identify broad association patterns within the feature set.

2.3 Combining Spline Models and Fuzzy Clustering (Article III)

In Article III, the low number of samples did not allow to analyze feature interactions at individual time-points in more detail using multivariate methods. However, the temporal resolution and the reduced variation between cells in contrast to patients offered different options for a robust analysis.

The basic research question in Article III aims at the understanding of the metabolic cross-talk between host and bacteria; thus, interpretability of the applied analysis was a key requirement. To smooth the temporal profiles and to reduce data complexity, natural regression splines were fitted to the time-course of each protein. Afterward, an (empirical Bayes moderated) F-test was conducted to identify proteins whose

abundance changes during the measured period (testing for non-constancy of the spline).

The temporal profiles of the proteins that were identified as significantly altered over time were additionally clustered. A fuzzy c-means time course clustering was used. Fig. 7 shows clusters of the bacterial proteins. The data processing and clustering successfully reveals patterns of protein abundances, identifies time points at which significant alterations happen, and allows for better interpretation of the data.

3 MEDICAL CONDITIONS

All reported analyses focus on population-relevant, systemic diseases and related biomarkers.

3.1 *Non-Alcoholic Fatty Liver Disease (Article I)*

As the largest inner organ of the body, the liver is responsible for a broad range of metabolic processes, including lipid metabolism, bile production, protein, and amino acid synthesis, detoxification, glucose control, the recycling of blood cells, or the storage of iron, copper, and vitamins. Today, nonalcoholic fatty liver disease (NAFLD) is the leading cause of chronic liver disease and liver transplantation in Western countries (Burra et al., 2020). It describes a spectrum of disorders characterized by an increase of intrahepatic fatty acids (FA). The accumulation of FAs leads to changes in redox balance and a subsequent increase of reactive oxygen species (ROS) levels. These changes further promote disease progression from simple steatosis to steatohepatitis, advanced fibrosis, and subsequently to irreversible liver damage in the form of cirrhosis or hepatocellular carcinoma (Buzzetti et al., 2016; Drescher et al., 2019).

However, already early-stage NAFLD may cause cell damage, oxidative stress, proinflammatory cytokines, adipokines, and mitochondrial dysfunction (Cazanave et al., 2017). These early changes influence the hepatic function mostly without specific symptoms (Masarone et al., 2018). NAFLD is a multicausal, systemic disease. Its occurrence is strongly associated with metabolic syndrome and type 2 diabetes, which seem to have common pathophysiological mechanisms (Buzzetti et al., 2016). However, its primary causes and triggers of disease progression still lack full understanding.

In Article I, a biomarker network is trained and used to predict and analyze the occurrence of NAFLD based on heterogeneous cohort study data. We show that

the inferred model identifies reasonable biomarker interactions and outperforms established biomarker scores.

3.2 *Hypertension (Article I)*

Blood pressure in arteries is measured by monitoring systolic (contraction) and diastolic (relaxation) pressure. Hypertension describes the condition of persistently elevated blood pressure and is typically diagnosed if the systolic pressure exceeds 140 mmHg or the diastolic pressure exceeds 90 mmHg in repeated measurements. Early detection and treatment of hypertension is essential, as it is a major risk factor for coronary artery disease, stroke, heart failure, and overall end-organ damage (heart, kidneys, brain, and eyes) (Mills et al., 2020; Franceschini et al., 2014; Völzke et al., 2013).

In Article I, a biomarker network of incident hypertension is learned from SHIP-TREND data. The network identifies many well-known risk factors including obesity, age, or chronic conditions and puts them into context. It also reveals the substantial heritability of hypertension.

3.3 *Thyroid (Dys-)Function (Article II)*

The thyroid gland is the largest endocrine gland in humans. Apart from the thyroid gland itself, the hypothalamus and the pituitary gland are part of the thyroid homeostasis in terms of a multi-loop feedback system. The thyroid hormones triiodothyronine (T₃) and thyroxine (T₄), which are produced directly by the thyroid gland, are of great importance for the metabolism and functional state of almost any organ. The hormone thyrotropin (TSH), which constitutes a central biomarker, is secreted by the pituitary gland, and stimulates the secretion of T₄ and T₃ in the thyroid gland. T₃ and T₄, in turn, inhibit the production and secretion of TSH through a negative feedback loop so that an equilibrium among thyroid hormones is usually established and the concentration of thyroid hormones in the blood plasma can be regulated.

The causes and manifestations of thyroid dysfunctions are diverse but can all lead to dysbalances in this hormone homeostasis (Moini et al., 2020; Taylor et al., 2018). The two most common types of thyroid dysfunctions are hyper- and hypothyroidism. They are defined as an increased (or decreased) activity of thyroid hormones. Since receptors for thyroid hormones are widespread within the body, symptoms may occur in almost all organ systems.

In Article II, broad association patterns of individual serum thyrotropin concentrations are analyzed using a machine learning workflow that combines random forests and Bayesian networks. Single nucleotide polymorphisms (SNPs) were also included for the analysis to capture possible genetic influences. The presented predictive random forest model outperforms existing models and the detailed analysis of association patterns is in line with state-of-the-art hypotheses on thyroid function.

3.4 *Staphylococcal Infections (Article III)*

Staphylococcus aureus (*S. aureus*) is a spherical, Gram-positive bacterium, that is often arranged in clusters. *S. aureus* is widespread, occurs in many habitats and belongs to the normal colonizing flora of the skin and mucosa in humans. In most cases, *S. aureus* does not cause any symptoms in humans. However, it is potentially pathogenic and in case of favorable conditions or a weak immune system of the host, the bacterium spreads and causes severe symptoms. In addition to skin and soft tissue infections, it may cause pneumonia, meningitis, endocarditis, toxic shock syndrome and sepsis (Sakr et al., 2018; Palma Medina et al., 2019).

In Article III, the dynamic interplay between host and bacteria was examined over time. The time series clustering of host cell and bacterial protein abundances (Fig 7) reveals significant general changes at 6 and 24 hours post infection, that go along with metabolic differences between a replicating and a persistent subpopulation of bacteria in host-cells: While replicating bacteria induce lysis of the host cells within 24 hours post infection, persistent bacteria adapt to the intracellular environment and reach a state of dormancy.

In general, the quantification of the change in abundance over time (in terms of one p -value for each protein) and the clustering of proteins with similar temporal profiles set the basis for an in-depth analysis. Altered pathways and behavior of bacteria and host were analyzed in detail. For example, the proteomic profiles revealed the competition for nutrients of bacteria and host within cells, resulting in altered glucose uptake and catabolism.

CONCLUSION

High-throughput technologies and electronic health records allow for the generation of large volumes of biomedical data. However, especially in the healthcare domain, the adoption of machine learning methods for data analysis is limited as often the strongly required transparency is missing. Simultaneously, algorithms need to be adapted to common data challenges that naturally occur in this domain.

This thesis reflects recent investigations on interpretability and approaches for learning interpretable models. All presented workflows make use of factorization and actively include latent structure in the form of subgroups, trends, or graph structure. The presented studies underline the complexity of interpreting high-dimensional healthcare data and show that achieving interpretability is a multi-faceted and multi-disciplinary task. Moreover, the studies underline the importance of methods that learn feature interactions: The assumption of independence is usually heavily violated in healthcare data, and interactions of risk factors often play a crucial role in disease development. At the same time, feature interactions are often unknown or such complex that they cannot manually be introduced to the model.

It could be shown that identifying latent structures facilitates a subsequent in-depth analysis. Hereby, visualization and probabilistic model output, as produced by Bayesian networks (Articles I and II), was found to be advantageous, as it provides an intuitive framework directly addressing the inherent uncertainty. Adaptive refinement, as introduced in Article I, offers the possibility to apply even complex latent structure models to high-dimensional data by varying the resolution in which the latent structure is modeled. Compression and stratification constitute as well essential tools if data are time-resolved, like in Article III.

However, balancing between interpretability and flexibility of a model is often not straightforward. Also, the interpretation of complex data may be still complex, even if an interpretable model is used. Thus, it is important that informaticians and biomedical researchers collaborate closely to define a clear goal and find an optimal tradeoff.

Part II

THESIS ARTICLES

AUTHOR CONTRIBUTIONS

ARTICLE I: FROM HETEROGENEOUS HEALTH-CARE DATA TO DISEASE-SPECIFIC BIOMARKER NETWORKS: A HIERARCHICAL BAYESIAN NETWORK APPROACH.

Citation	Becker, A. K., M. Dörr, S. B. Felix, F. Frost, H. J. Grabe, M. M. Lerch, M. Nauck, U. Völker, H. Völzke, and L. Kaderali (2021). From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. <i>PLoS Computational Biology</i> 17(2)
Author Contributions	A.-K. Becker and L. Kaderali conceptualized the presented idea and developed the methodology. M. Dörr, S.B. Felix., F. Frost., H.J. Grabe, M.M. Lerch, M. Nauck, U. Völker and H. Völzke were involved in data collection and curation. A.-K. Becker performed data preprocessing, developed the software, performed the computations and the formal analysis. L. Kaderali supervised the findings of this work. A.-K. Becker drafted the manuscript to which all authors contributed.

ARTICLE II: DISCOVERING ASSOCIATION PATTERNS OF INDIVIDUAL SERUM THYROTROPIN CONCENTRATIONS USING MACHINE LEARNING: AN EXAMPLE FROM THE STUDY OF HEALTH IN POMERANIA (SHIP).

Citation	Becker, A.-K., T. Ittermann, M. Dörr, S. B. Felix, M. Nauck, A. Teumer, U. Völker, H. Völzke, L. Kaderali, and N. Nath (2021). Discovering association patterns of individual serum Thyrotropin concentrations using machine learning: An example from the Study of Health in Pomerania (SHIP). <i>PLoS Computational Biology</i> (under review)
Author Contributions	A.-K. Becker and N. Nath conceptualized the presented idea and developed the methodology. T. Ittermann, M. Dörr, S.B. Felix, M. Nauck, A. Teumer, U. Völker and H.Völzke were involved in data collection and curation. A.-K. Becker and N. Nath performed data preprocessing. N. Nath performed the random forest analysis. A.-K. Becker analyzed the model, performed feature selection and the Bayesian network analysis under supervision of N. Nath. A.-K. Becker, N. Nath and L. Kaderali analyzed the results and drafted the manuscript to which all authors contributed.

ARTICLE III: METABOLIC CROSS-TALK BETWEEN HUMAN BRONCHIAL EPITHELIAL CELLS AND INTERNALIZED *staphylococcus aureus* AS A DRIVER FOR INFECTION.

Citation	Palma Medina, L. M., A.-K. Becker, S. Michalik, H. Yedavally, E. J. Raineri, P. Hildebrandt, M. G. Salazar, K. Surmann, H. Pförtner, S. A. Mekonnen, et al. (2019). Metabolic cross-talk between human bronchial epithelial cells and internalized staphylococcus aureus as a driver for infection. <i>Molecular & Cellular Proteomics</i> 18(5), 892–908
Author Contributions	L.M. Palma Medina, A. Salvati, L. Kaderali, J.M. van Dijl, and U. Völker designed and directed the research. L.M. Palma Medina, H. Yedavally, E.J. Raineri., P. Hildebrandt, M.G. Salazar, K. Surmann, H. Pförtner, and S.A. Mekonnen performed the experiments. L.M. Palma Medina, A.-K. Becker, S. Michalik, L. Kaderali, J.M. van Dijl, and U. Völker analyzed the data. S. Michalik and A.-K. Becker performed data preprocessing, A.-K. Becker performed the spline analysis, statistical testing, and the time-series clustering and contributed to the interpretation of the results. L.M. Palma Medina took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

(doctoral candidate)

(head of thesis committee)

ARTICLE I

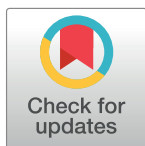
RESEARCH ARTICLE

From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach

Ann-Kristin Becker¹, Marcus Dörr^{2,3}, Stephan B. Felix^{2,3}, Fabian Frost⁴, Hans J. Grabe⁵, Markus M. Lerch⁴, Matthias Nauck⁶, Uwe Völker⁷, Henry Völzke⁸, Lars Kaderali^{1*}

1 Institute of Bioinformatics, University Medicine Greifswald, Greifswald, Germany, **2** Department of Internal Medicine B, University Medicine Greifswald, Greifswald, Germany, **3** German Centre for Cardiovascular Research (DZHK), partner site Greifswald, Greifswald, Germany, **4** Department of Internal Medicine A, University Medicine Greifswald, Greifswald, Germany, **5** Department of Psychiatry, University Medicine Greifswald, Greifswald, Germany, **6** Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany, **7** Interfaculty Institute of Genetics and Functional Genomics, Department of Functional Genomics, University Medicine Greifswald, Greifswald, Germany, **8** Institute of Community Medicine, SHIP/KEF, University Medicine Greifswald, Greifswald, Germany

* lars.kaderali@uni-greifswald.de



OPEN ACCESS

Citation: Becker A-K, Dörr M, Felix SB, Frost F, Grabe HJ, Lerch MM, et al. (2021) From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. *PLoS Comput Biol* 17(2): e1008735. <https://doi.org/10.1371/journal.pcbi.1008735>

Editor: Alison Marsden, Stanford University, UNITED STATES

Received: August 19, 2020

Accepted: January 22, 2021

Published: February 12, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1008735>

Copyright: © 2021 Becker et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data from the Study of Health in Pomerania (SHIP) cannot be shared publicly as they contain potentially identifying and

Abstract

In this work, we introduce an entirely data-driven and automated approach to reveal disease-associated biomarker and risk factor networks from heterogeneous and high-dimensional healthcare data. Our workflow is based on Bayesian networks, which are a popular tool for analyzing the interplay of biomarkers. Usually, data require extensive manual preprocessing and dimension reduction to allow for effective learning of Bayesian networks. For heterogeneous data, this preprocessing is hard to automatize and typically requires domain-specific prior knowledge. We here combine Bayesian network learning with hierarchical variable clustering in order to detect groups of similar features and learn interactions between them entirely automated. We present an optimization algorithm for the adaptive refinement of such group Bayesian networks to account for a specific target variable, like a disease. The combination of Bayesian networks, clustering, and refinement yields low-dimensional but disease-specific interaction networks. These networks provide easily interpretable, yet accurate models of biomarker interdependencies. We test our method extensively on simulated data, as well as on data from the Study of Health in Pomerania (SHIP-TREND), and demonstrate its effectiveness using non-alcoholic fatty liver disease and hypertension as examples. We show that the group network models outperform available biomarker scores, while at the same time, they provide an easily interpretable interaction network.

Author summary

High-dimensional and heterogeneous healthcare data, such as electronic health records or epidemiological study data, contain much information on yet unknown risk factors that

sensitive medical information on study participants. However, data access can be requested from the Forschungsverbund Community Medicine data access committee (online application form at <http://fvcm.med.uni-greifswald.de>) for researchers who meet the criteria for access to confidential data.

Funding: We acknowledge funding by the German BMBF via the LiSyM grant (FKZ 031L0032). AKB holds an add-on fellowship from the Joachim Herz Stiftung. HJG has received travel grants and speakers honoraria from Fresenius Medical Care, Neuraxpharm, Servier and Janssen Cilag as well as research funding from Fresenius Medical Care. SHIP is part of the Community Medicine Research Network of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg- West Pomerania. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

are associated with disease development. The identification of these risk factors may help to improve prevention, diagnosis, and therapy. Bayesian networks are powerful statistical models that can decipher these complex relationships. However, high dimensionality and heterogeneity of data, together with missing values and high feature correlation, make it difficult to automatically learn a good model from data. To facilitate the use of network models, we present a novel, fully automated workflow that combines network learning with hierarchical clustering. The algorithm reveals groups of strongly related features and models the interactions among those groups. It results in simpler network models that are easier to analyze. We introduce a method of adaptive refinement of such models to ensure that disease-relevant parts of the network are modeled in great detail. Our approach makes it easy to learn compact, accurate, and easily interpretable biomarker interaction networks. We test our method extensively on simulated data as well as data from the Study of Health in Pomerania (SHIP-Trend) by learning models of hypertension and non-alcoholic fatty liver disease.

Introduction

High-throughput technologies and electronic health records allow for digital recording and analysis of large volumes of biomedical and clinical data. These data contain plenty of information about complex biomarker interaction systems, and they offer fascinating prospects for disease research. However, to extract this knowledge from the data and make it accessible, we need models that are accurate, easily interpretable, and compact. Bayesian networks (BNs) are popular and flexible probabilistic models that lie at the intersection of statistics and machine learning and can be used to model complex interaction systems. BNs explicitly describe multivariate interdependencies using a network structure in which the measured features are the nodes and directed edges represent the relationships among those features. Thus, they offer an intuitive graphical representation that visualizes how information propagates. This interpretable structure sets them apart from ‘black-box’ concepts of other machine-learning methods. Besides, there are well-established algorithms for the automatic learning of Bayesian networks from data, and they are widely used in Systems Biology, e.g., to model cellular networks [1], protein signaling pathways [2], gene regulation networks [3–5], or as medical decision support systems [6]. For a thorough introduction to Bayesian networks see for example Koski and Noble [7] or Koller and Friedman [8].

However, large volumes of biomedical data raise a challenge for computational inference, as in addition to their high dimensionality, other difficulties, such as incompleteness, heterogeneity, variability, strong feature correlation, and high error rates usually co-occur. Considerable manual time and human expertise are therefore necessary to process and format data, including steps of annotation, normalization, discretization, imputation, and feature selection. In addition to the related expenses, these preprocessing steps have a substantial impact on the resulting model [9, 10]. Therefore, we have developed an entirely automated and data-driven workflow that combines Bayesian network learning with hierarchical variable clustering. Our approach tackles many of the mentioned issues simultaneously, while in manual processing, they are usually approached independently. The combination of the two well-established concepts helps to derive precise biomarker interaction models of manageable complexity from unprocessed biomedical data.

Bayesian network learning usually comprises two separate steps: First, the network structure (a directed acyclic graph) is inferred, then, local probability distributions are estimated.

Structure learning can either be carried out using repeated conditional independence tests (constraint-based learning) or search-and-score techniques (score-based learning) [8]. However, as the number of possible network structures grows super-exponentially, available algorithms usually do not scale well to more than 50 to 100 variables. Various heuristic approaches as well as the incorporation of further information, such as sparseness assumptions or more general restrictions of the search space have led to some progress in learning large Bayesian networks [11, 12]. However, due to the complexity of the underlying statistical problem (non-identifiability, non-convexity, non-smoothness), Bayesian network learning from high-dimensional data remains challenging, and often yields inconsistent results. Moreover, the subsequent interpretation of a giant network is just as complex. Because of the mentioned difficulties, published biomedical Bayesian network models are often based on molecular datasets with homogeneous variables [13–15], as for them, all features can be processed in a similar way. Often, the subsequent analysis concentrates mainly on global network properties. Studies on heterogeneous epidemiological data usually involve smaller models with a preselected set of features, e.g., of cardiovascular risk [16, 17], renal transplantation [18] or liver diseases [19–21].

Because of the way in which biomedical data are gathered, they often contain groups of highly related variables. Some features may be explicitly redundant (like replicated measurements) or multiple features measure the same aspect (like the percentage of body fat and waist circumference). The underlying interaction network (Fig 1A) is then modular or hierarchically modular [22, 23]. This modularity complicates the identification and inference of a Bayesian network even more, as for example many structure learning algorithms penalize for high node degrees that are present in such modules [8]. However, if the modular organization is known, it can be used to simplify the original problem. Instead of aiming for a detailed Bayesian network, a network among groups of similar features can be learned (Fig 1C). Such networks are called *group Bayesian networks*. Group Bayesian networks are smaller and less connected than detailed networks. Moreover, results tend to be more consistent, as the grouping and

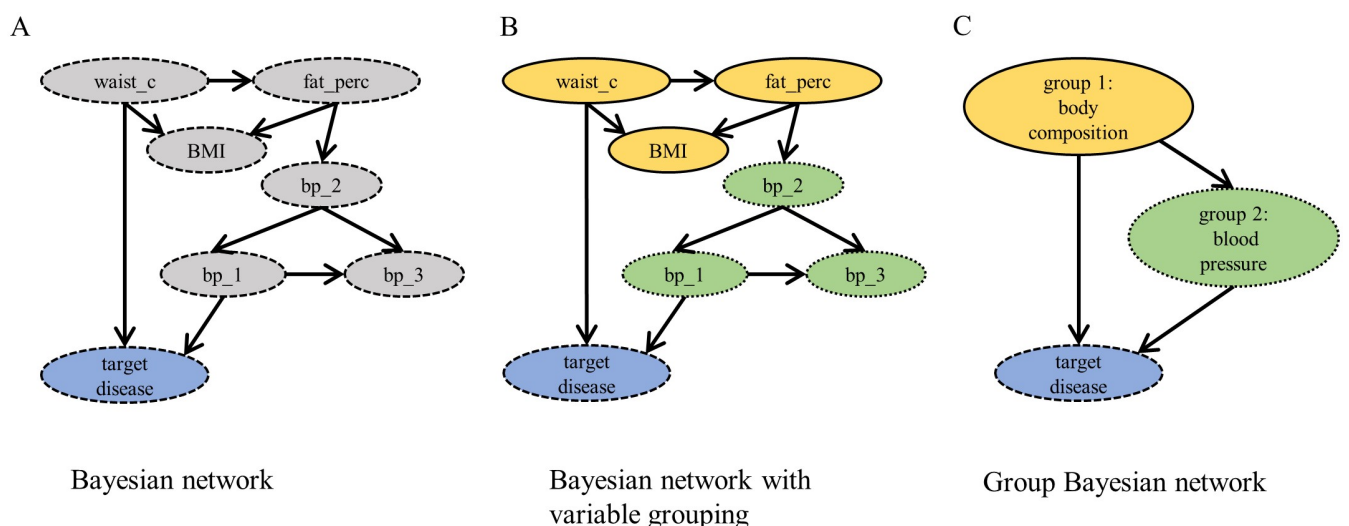


Fig 1. Hypothetical example Bayesian network with and without variable grouping. (A) Example model of a modular detailed Bayesian network with variables waist circumference (*waist_c*), body fat percentage (*fat_perc*), BMI and three blood pressure measurements (*blood_pr1*, *blood_pr2*, *blood_pr3*) as well as a *target disease*. (B) Possible grouping of the variables in the network. (C) Corresponding group Bayesian network among two groups and the target variable.

<https://doi.org/10.1371/journal.pcbi.1008735.g001>

aggregation act as denoise filters. Additionally, the abstraction enables the understanding of the larger picture from a system's point of view.

Most publications that have addressed the question of how to learn Bayesian networks of variable groups discuss the problem for a given grouping. This includes the application to pathway or SNP dependencies given detailed genetic data [24, 25]. However, the determination of the number and type of variable groups is a crucial question itself, and it is unlikely that the correct grouping is known for complex and heterogeneous data. On the other hand, there is the concept of *Module Networks*, which is well studied, and algorithms are available to learn modules and their interactions from data [26–28]. But since Module networks were developed in the context of gene regulatory networks, their structural limitations (variables in modules share set of parents and local probability distribution) do not apply to data as we consider here. Likewise, *hierarchical Bayesian networks* (HBNs) [29] define a related, very general concept of tree-like networks, in which leaf nodes represent observed variables and internal layers represent latent variables. HBNs are usually strictly hierarchical. This means that, similar to the architecture of deep neural networks, they restrict all nodes to have parents only in higher layers [30, 31]. Nevertheless, group Bayesian networks can be seen as a special case of loose HBNs. Latent variables in HBNs can theoretically be identified from detailed Bayesian networks, for example, using subgraph partitioning [32]. However, this approach requires the computationally intensive inference of a large, detailed network, and it suffers from the difficulties mentioned above.

We, instead, propose to combine Bayesian networks with hierarchical clustering to learn a grouping of variables as well as the interplay of groups automatically. Hierarchical clustering is one of the most popular methods of unsupervised learning. The output is a dendrogram, which organizes variables in increasingly broad categories. We propose to build group Bayesian networks by aggregating groups learned from hierarchical clustering. As both methods, BNs and clustering, are unsupervised, we enable focusing on a particular target variable of interest—such as a specific disease or condition—during a step of adaptive refinement. We present an optimization algorithm, that, starting from a coarse network, refines important parts of the network downwards along the dendrogram. It zooms automatically into the relevant parts of a network, which are modeled in detail, while other parts stay aggregated. Thus, refined group Bayesian networks offer a good tradeoff between compactness, interpretability, and predictive power.

While some published approaches make use of variable clustering in order to speed up the learning of detailed networks by going from local (within groups) to global (between groups) connections [4, 33, 34], we are not aware of any study addressing the reverse approach.

Results and discussion

Algorithm

We here introduce a novel algorithm to significantly simplify the use of Bayesian network models for biomarker discovery (Fig 2). It explicitly integrates a target variable of interest that guides the search through the biomarker network. Our approach exploits the modular structure of large biomedical data and models dependencies among groups of similar variables. To keep the combined search for grouping and network structure feasible, a hierarchical structure acts as a basis for the following network inference procedure. Initially, a dendrogram of the feature space is determined via unsupervised, similarity-based hierarchical clustering. A coarse, preliminary grouping of features is identified, and the data are aggregated in groups using principal components. Then, structure and parameters of a Bayesian network model are fitted. The target variable is kept separated during this procedure so that the resulting model

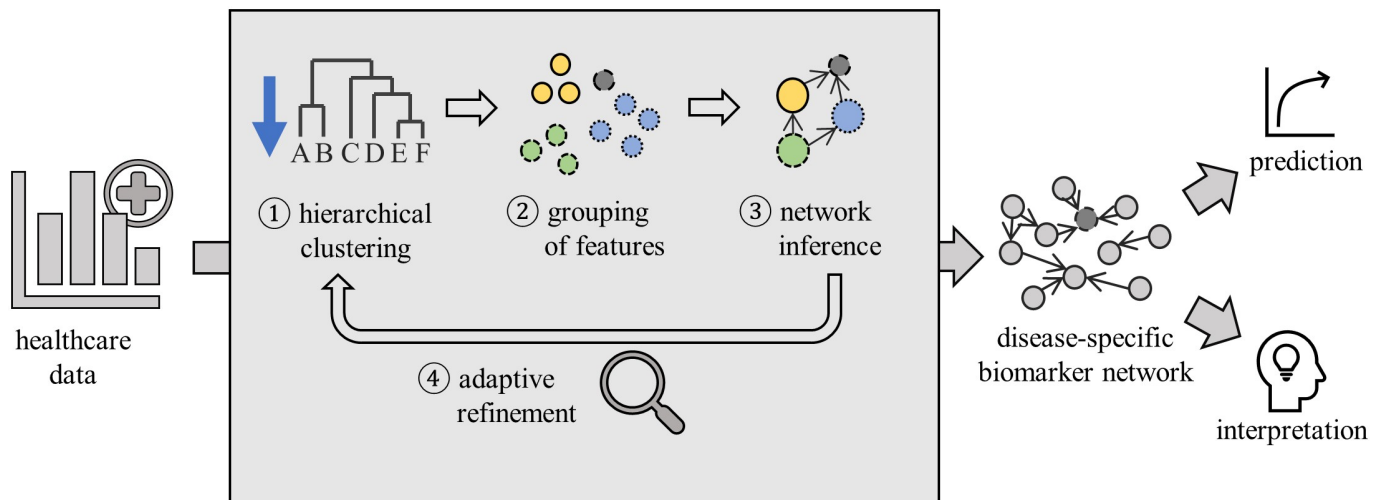


Fig 2. Schematic outline of the proposed approach to learn group Bayesian networks. Features of the input data are grouped using hierarchical clustering, then a group Bayesian network is learned. Based on the accuracy of the resulting model, the grouping is refined adaptively downwards along the dendrogram. The output is an interpretable disease-specific biomarker network based on feature groups, which has high predictive accuracy.

<https://doi.org/10.1371/journal.pcbi.1008735.g002>

can be used for risk prediction and classification. Such groups that were identified to be essential for the prediction of the target variable (i.e., are part of its Markov blanket) are then iteratively refined to smaller clusters. The refinement stops once it no longer helps to improve the predictive performance of the model. We implemented our approach for the construction and refinement of group Bayesian networks using a hill-climbing procedure (Algorithm 1 and 2). The implementation is also available in CRAN from <https://CRAN.R-project.org/package=GroupBN> [35].

Evaluating simulated data

We evaluated the proposed approach using simulated data. To generate noisy and heterogeneous data with latent group structure, we randomly created two-layered Bayesian networks (Fig 3A) with one layer of group variables (*layer 1*) and one layer representing noisy and heterogeneous measurements (*layer 0*). Here, the overall aim was to infer the group structure in layer 1 from data in layer 0. For the analysis, we split the algorithm into its three key-tasks, that we evaluated independently: Inference of groups, inference of group network structure, and prediction of a target variable. In the ‘standard network inference’ approach, the grouping was disregarded for network learning. Instead, a large, detailed Bayesian network was learned, and groups as well as their interactions were only afterwards identified from the network. In the ‘group network inference’ approach, we contrarily learned the grouping prior to network inference using data-based clustering, as proposed above. For group aggregation, we compared cluster medoids (MED) to first principal components (PC). As a baseline comparison for the quality of the network structure, we additionally inferred the network structure directly from data sampled from layer 1 (‘using ground-truth grouping’). We used a partition metric to the ground-truth grouping and the normalized Hamming Distance to the ground-truth network as measures of quality. Lastly, we iteratively chose each variable as target variable and measured the average predictive performance of a detailed network, as well as group networks before and after target-specific refinement. Here, we compared the average prediction error to the applied noise level.

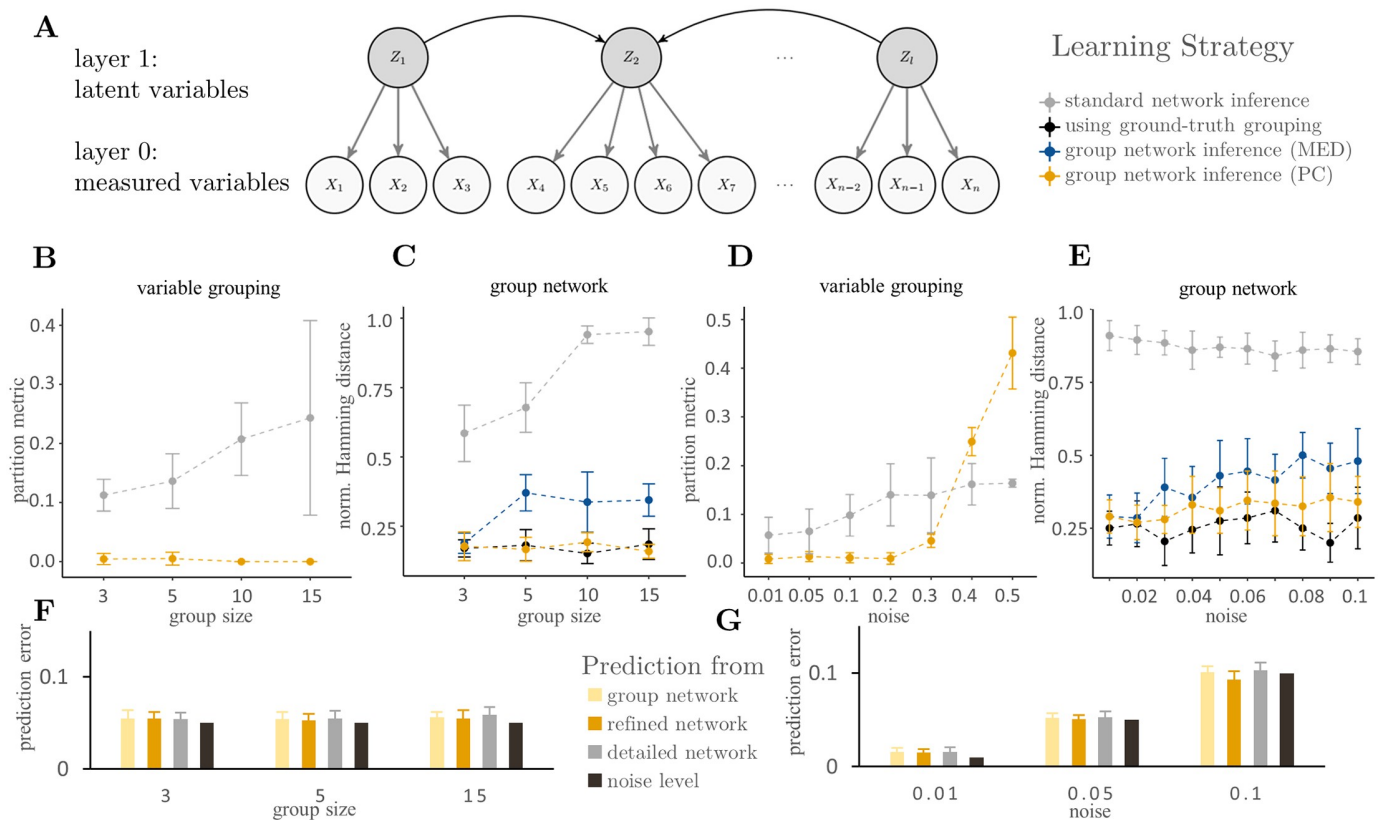


Fig 3. Results on simulated networks. (A) The basic model structure used to simulate random networks with latent group structure. Group networks with 20 nodes in layer 1 were learned from simulated data from layer 0 with varying group sizes and noise levels. (B–C) Results from the reconstruction of variable grouping and group networks for varying group sizes. y-axes showing partition metric and normalized Hamming distance, respectively. Two types of group network inference—aggregation by principal components (PC) and cluster medoids (MED)—as well as a standard network inference approach were used. As a comparison, the ground-truth grouping was used for network inference. (D–E) Results from the reconstruction of variable grouping and group networks for varying noise levels. y-axes showing partition metric and normalized Hamming distance, respectively. (F–G) Results from the prediction of a target variable for varying group sizes and noise levels, and applied noise level as comparison. y-axes showing the average prediction error.

<https://doi.org/10.1371/journal.pcbi.1008735.g003>

Influence of network size and sample size. We first analyzed the influence of network and sample size on the model quality. The results show that the quality of the network structure is best for high sample sizes and small network sizes (S1 Fig). Overall, the PC-based aggregation is close to the baseline results, followed by the medoid-based aggregation, with the network-based aggregation performing worst. Based on these results, we decided to run the remaining simulations with group networks consisting of 20 nodes at layer 0 and a medium sample size of 500.

Influence of group size. Next, we tested the influence of group size on the inference results. We ran simulations with groups ranging from 3 to 15 nodes each. The results show that the identification of variable groups based on a detailed network is impaired with increasing group size. In contrast, data-based clustering enables the detection of the nearly correct grouping independently of the group size. Moreover, even the existence of small groups impedes the inference of the network structure from a detailed network significantly. Especially the number of group connections is underestimated. This effect increases with increasing group size, approaching scores similar to a model without any arcs (Fig 3B). However, data-based clustering enabled the detection of the correct grouping independently of the group size. Aggregation of data before network learning leads to networks that are qualitatively

comparable to networks learned from the group data directly (Fig 3C). Here, the aggregation based on principal components overall achieves better results than with medoids. The prediction of a target variable is mostly positively affected by the grouping (Fig 3F). The refined networks overall perform slightly better when used for predicting the target variable in a cross-validation setting than the detailed models.

Influence of random noise. Finally, we analyzed the influence of random noise. For this purpose, we simulated networks affected by different amounts of random noise in layer 1. The results show that the quality of the inference of groups, as well as group interactions, decreases with increasing noise levels. Data-based clustering outperforms network-based clustering for noise levels up to 35% (Fig 3D). Data aggregation by principal components overall leads to better networks than the use of medoids (Fig 3E). However, a decrease in quality can be noticed for both approaches, as the noise level increases. The average error in prediction of a target variable appears to be in the range of the noise level with slight improvements after target-specific refinement (Fig 3G).

Discussion of simulation results. The simulation results underline, that the aggregation of data increases the quality of the network model compared to group networks that were inferred from detailed networks. This may be explained by the inherent regularization of most structure learning algorithms, that prioritize intra-group interactions in this setting, as those are very strong. Thus, groups tend to be disconnected from each other in a detailed network, even though strong connections are present in the correct network. The proposed combination of hierarchical clustering and network inference puts importance on the inter-group interactions, enabling their accurate inference. Moreover, we observed an overall better performance of aggregation using principal components. This goes along with earlier results on PCA preprocessing for Bayesian networks [10].

Toy example: Wine data

As a first illustration using a small, real-world example, we demonstrate the capability of the proposed method on benchmark data for clustering of heterogeneous variables. The *wine* dataset [36, 37] contains data on the sensory evaluation of red wines from Val de Loire. Variables contain scorings on origin, odor, taste, and visual appearance of the wines. We study the influence of the wine-producing soil on the properties of the wine.

We examine the difference of 7 wines grown in soil type *Env1* to 7 wines of the class *Reference*, an excellent wine-producing soil. In order to learn the links among the variables, we clustered the data subset (14 samples, 29 variables) hierarchically (Fig 4A). We chose 5 clusters for an initial grouping. Fig 4B and 4C show the group Bayesian network model before and after refinement. Line thickness illustrates the confidence of the learned interaction. The neighborhood of the target variable is modeled more detailed in the refined network (Fig 4C). The network revealed two factors, that mainly distinguish wines from *Soil = Reference* and *Soil = Env1*; namely *Acidity* and *Aroma.quality.before.shaking*. Through these variables, the target is further indirectly linked to two kinds of odor (fruity, flower), as well as a larger cluster comprising measures of odor- and aroma intensity. A closer look at the parameters of the Bayesian network revealed that a wine from the reference soil is typically more fruity, less acidic, and has a higher score in aroma quality and floral aroma.

The arc with the highest confidence was learned among *aroma quality before shaking* and *Soil*. Given a wine with a good aroma quality before shaking, there is an 85% probability according to the model, that this wine is from the reference soil and only 15% that it is from soil class *Env1*. On the contrary, soil and spiciness or overall balance of a wine are

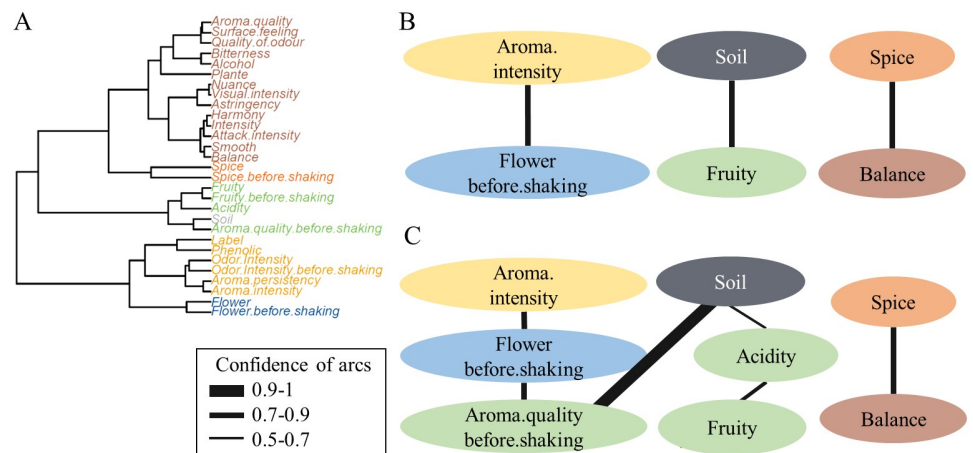


Fig 4. Toy example: Wine dataset. (A) Dendrogram of the wine dataset with 5 groups indicated by colour, and the target variable *Soil* separated. (B) Group Bayesian network learned from the wine dataset with 5 groups, colours refer to the grouping. (C) Group Bayesian network after target-specific refinement.

<https://doi.org/10.1371/journal.pcbi.1008735.g004>

disconnected in the network, indicating that the soil type *Env1* does not influence these characteristics significantly.

Validation with healthcare data: Study of Health in Pomerania (SHIP-TREND)

We further validated the methodology using data from the Study of Health in Pomerania (SHIP-Trend-0) with focus on two common, multifactorial diseases, non-alcoholic fatty liver disease and hypertension. SHIP-Trend is a large-scale cohort study on the general population in Northeast Germany [38]. Interdisciplinary baseline examinations on a total number of 4420 participants were conducted between 2008 and 2012, including a wide variety of assessments. These assessments involve the recording of socioeconomic factors, a detailed questionnaire, measurements of molecular data, preexisting conditions, as well as various clinical tests such as blood counts, imaging techniques, electrocardiography, body impedance analysis and others.

Application 1: Non-alcoholic fatty liver disease. Non-alcoholic fatty liver disease (NAFLD) is widely considered a hepatic manifestation of the metabolic syndrome and represents the most common chronic liver disease worldwide, affecting 15-35% of the general population. Hepatic steatosis is the key feature of NAFLD and describes the excessive accumulation of liver fat. Steatosis is diagnosed if the amount of intrahepatic triglycerides exceeds 5% [39]. Simple hepatic steatosis may progress to non-alcoholic steatohepatitis (NASH), marking the most crucial step in the development of severe liver dysfunction with poor prognosis. Causes of the disease, as well as its progression, are still poorly understood. Today, liver biopsy is the gold standard to diagnose NAFLD [40] and its stage. However, besides its sampling bias, liver biopsy always involves risk of complications. Apart from that, imaging techniques like ultra-scan or magnetic resonance imaging are used. The development of cheaper and reliable noninvasive techniques to diagnose NAFLD are of urgent need—also with regard to prevention. Therefore, several biomarker scores have been proposed in the last years, including the Fatty Liver index [41], Hepatic Steatosis Index [42, 43], and NAFLD ridge score [44], all of which combine 3 to 6 different anthropometric parameters and biochemical tests. They allow for a cheap and noninvasive screening for steatosis in the general population. On their respective

Table 1. Prediction results of NAFLD models.

Model	AUROC	\pm sd	AUPRC	\pm sd
Hepatic Steatosis Index	0.68	± 0.04	0.24	± 0.04
Fatty Liver Index	0.78	± 0.05	0.34	± 0.05
NAFLD ridge score	0.73	± 0.05	0.29	± 0.04
logistic regression	0.78	± 0.03	0.37	± 0.05
detailed Bayesian network	0.74	± 0.02	0.31	± 0.05
group Bayesian network	0.79	± 0.04	0.35	± 0.05
refined group Bayesian network	0.82	± 0.03	0.42	± 0.04

Evaluation of available steatosis scores, logistic regression and different Bayesian network models on SHIP Trend data in terms of discrimination. The table shows area under receiver-operator curve (AUROC), and area under precision-recall curve (AUPRC) under 10-fold cross validation (mean and standard deviation). Predictions from Bayesian network models were obtained using likelihood weighting by taking all nodes but the target as evidence. Best scoring steatosis biomarker score and best scoring Bayesian network model are highlighted.

<https://doi.org/10.1371/journal.pcbi.1008735.t001>

original datasets, these scores achieved an area under the receiver-operator curve (AUROC) between 0.81 and 0.87, thus leaving a substantial proportion of false positive and false negative results. On the SHIP Trend data, the AUROC lies significantly lower, between 0.67 and 0.78 (Table 1). The area under the precision-recall curve (AUPRC), which has its focus on the underrepresented class of positive cases, ranges from 0.24 to 0.34.

We applied the proposed group network approach to the SHIP-Trend data. Compared to a detailed network, the aggregation of data into groups already improved the prediction of steatosis in a Bayesian network (Table 1). The unrefined group network achieved an AUROC score of 0.79 in a cross validation setting. The score is comparable to the one reached by logistic regression and the FLI, which we found to be the best performing biomarker score on the SHIP Trend data of the three tested ones. The refinement procedure resulted in an improved final AUROC score of 0.82 and an AUPRC of 0.42.

We then fit a final model on the complete dataset for interpretation. Hierarchical clustering of the data revealed 17 groups of features. The final network model (Fig 5 and S4 Fig) has an average neighbourhood size of 2.5, an average group size of 16 and also achieved an AUROC of 0.82. Fig 5A shows the complete network structure, in which sex and age are both hubs. Fig 5B shows only the target variable and its surrounding. The group names have been chosen

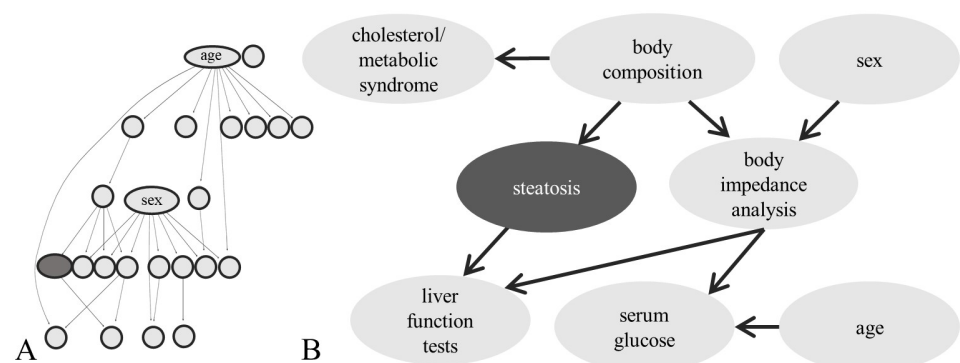


Fig 5. Steatosis network model. (A) Structure of the complete, refined group Bayesian network model for hepatic steatosis. (B) Extract from the group network including the target variable *steatosis*, its Markov blanket and surrounding.

<https://doi.org/10.1371/journal.pcbi.1008735.g005>

manually according to the included variables. The detailed grouping can be found in [S1 Data](#). Steatosis has one parent node, which is a group of variables related to body composition, including body mass index, waist circumference, body fat and others. This group is further linked to a group including cholesterol and triglyceride levels, as well as a group including raw results of the body impedance analysis (BIA). The child node of the target comprises different variables related to serum liver function tests (alanine aminotransferase, aspartate transaminase, gamma-glutamyl transferase). Sex and serum glucose levels are indirectly linked to the group of liver function tests via BIA results.

We further evaluated the distance of the features to the target in the network. In the moralized network, the average distance to the target variable is 2.09. The predictors that have been used in the liver scores are closer than average, with in average only 1.5 arcs distance to the target. For the the FLI, three out of the used four predictors (BMI, waist circumference, triglycerides and GGT) are within the Markov blanket (mean distance 1.25). This overlap might explain the similarity in prediction performance. It shows that meaningful features have been learned by the network. Moreover, the network illustrates clearly the strong relation of steatosis with obesity and the metabolic syndrome. However, different from pure prediction scores, the interpretability of the proposed model enables the understanding of how and why a prediction is made, and, by this, it shows also what may be overlooked. According to the model, and consistent with earlier studies, around 10% of steatosis cases do not go along with multi-organ metabolic abnormalities and obesity [45, 46]. These cases stay hardly detectable without imaging techniques.

Application 2: Hypertension. As a second example, we analyzed the SHIP-Trend data with a focus on hypertension. Hypertension describes the condition of persistently elevated blood pressure in arteries and is a major risk factor for coronary artery disease, stroke, heart failure, and overall end-organ damage (heart, kidneys, brain, and eyes). Blood pressure measurements monitor systolic (contraction) and diastolic (relaxation) pressures. Hypertension is typically diagnosed if the systolic pressure exceeds 140 mmHg or the diastolic pressure exceeds 90 mmHg. It is known to have a substantial heritability (estimated in the range of 30–55%) [47]. Moreover, many risk factors of hypertension are well established, including obesity, age, stress, or chronic conditions, such as diabetes or sleep apnea.

For our analysis, incident hypertension was defined as blood pressure above 140/90 mmHg or self-reported antihypertensive therapy. The target variable was not well connected within a detailed network learned from SHIP-Trend data, which is why a mean AUROC of only 0.55 is achieved in a cross validation setting for training as well as test sets. The refined group network model, however, achieves an AUROC score of 0.84 and an AUPRC of 0.81 (Table 2), which is comparable to other recent hypertension risk-prediction models and results on an earlier SHIP cohort [48, 49].

Table 2. Prediction results of hypertension models.

Model	AUROC	\pm sd	AUPRC	\pm sd
logistic regression	0.82	± 0.02	0.78	± 0.03
detailed Bayesian network	0.55	± 0.04	0.57	± 0.06
group Bayesian network	0.80	± 0.02	0.76	± 0.04
refined group Bayesian network	0.84	± 0.03	0.81	± 0.02

Evaluation of logistic regression and different Bayesian network models on SHIP Trend data for the prediction of hypertension. The table shows area under receiver-operator curve (AUROC), and area under precision-recall curve (AUPRC) under 10-fold cross validation (mean and standard deviation). Predictions from Bayesian network models were obtained using likelihood weighting by taking all nodes but the target as evidence. Best scoring Bayesian network model is highlighted.

<https://doi.org/10.1371/journal.pcbi.1008735.t002>

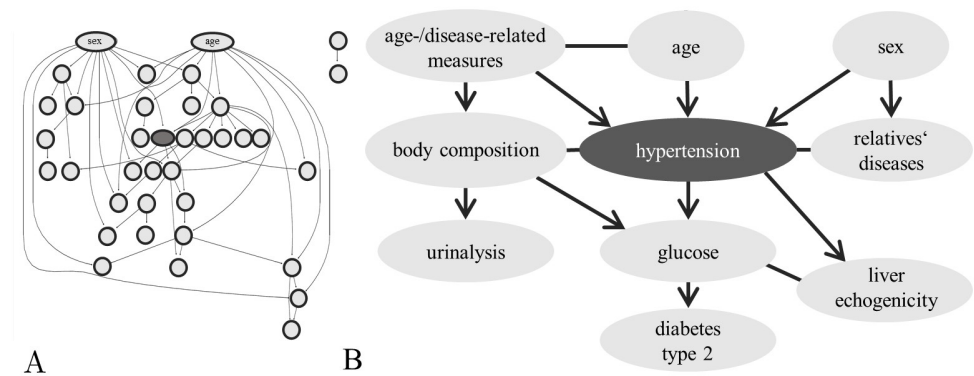


Fig 6. Hypertension network model. (A) Structure of the complete, refined group Bayesian network model for hypertension. (B) Extract from the group network including the target variable *hypertension*, its Markov blanket and surrounding.

<https://doi.org/10.1371/journal.pcbi.1008735.g006>

The final group Bayesian network from 28 groups, as determined by the aggregation levels, is densely connected. After refinement, the network (Fig 6 and S3 Fig) has an average neighbourhood size of 3.5 and an average group size of 9.4. The target variable has three parents in the network, which are age, sex, and a cluster of more general age- and disease-related measures (including the number of doctor visits, and information on employment/retirement). Further, a cluster of diseases of first degree relatives (including hypertension, heart attack, stroke, diabetes) and a cluster of measures of body composition are directly attached to the target. A group around fasting glucose level as well as a group around liver echogenicity are children of the target variable in the network. Via body composition, hypertension is further linked to a group of urinalysis results, as they show frequent consequences of hypertensive kidney injury. The network clearly visualizes the heritability of hypertension, as well as promoting environmental factors. The detailed grouping can be found in S2 Data.

Conclusion

Bayesian networks provide a powerful and intuitive tool for the analysis of the interplay of variables. In this work, we introduced a novel algorithm to infer Bayesian biomarker and risk factor networks from heterogeneous and high-dimensional healthcare data. Our approach combines Bayesian network learning and hierarchical variable clustering. By this means, it supersedes many of the usually necessary manual preprocessing steps and reduces the complexity of the computations, while preserving model interpretability. We introduced an optimization algorithm for adaptive network refinement, which emphasizes a variable of interest and enables the automated refinement leading to small yet precise disease-specific models. The results on simulated data, test data and real-world epidemiological data verify the ability of the approach to successfully reveal important biomarker and risk factor interactions. Moreover, we showed that the increased interpretability of the model does not restrain its predictive performance, which was in both biomedical examples equal or better than well-established purely predictive models. Our method is suitable for an in-depth analysis of biomarker systems, but apart from this, it can also be used as a quick summary and visualization tool for large data prior to further evaluation. Our findings add to a growing body of literature on the use of machine learning and artificial intelligence in medicine, and they facilitate multivariate data analysis, visualization, and interpretation.

The purpose of this study was to investigate how hierarchical variable grouping and Bayesian network learning can be combined to overcome the limitations of network inference on high-dimensional and heterogeneous data. The proposed methodology provides the framework to effectively learn Bayesian networks of manageable complexity without manual steps of feature selection. Our method could be applied to all types of tabular data with many features and high enough sample size for which the interest lies mainly in feature interactions. A crucial step in our procedure is the aggregation of groups for network learning. We found that in the studied data sets, groups of variables were often reflecting highly similar information. The use of single principal components as cluster representatives was therefore mostly sufficient and yielded reasonable clusters. However, depending on the complexity and the aim of the application, the results may be improved by the use of more sophisticated and more accurate aggregations, for example using multidimensional cluster representatives. However, higher precision in the modeling of variable groups would, in turn, significantly increase the computation time and complicate the interpretation. Note also, that data have to satisfy additional assumptions in order to be exactly modeled as group networks with two- or more-dimensional nodes, as studied by Parviainen and Kaski [25]. The same applies to a more complex grouping that allows overlapping clusters. For future studies, in particular a dynamic generalization of the approach using dynamic Bayesian networks is planned to enable the use of longitudinal study data for prognosis. We plan to also include molecular data in order to allow the integrative analysis of multi-omics and epidemiological data. By this, the proposed methodology offers the possibility to reveal yet unknown biomarker and risk factor relations, and to gain new insight into molecular disease mechanisms.

Methods

Group Bayesian networks and adaptive refinement

We implemented an approach to learn group Bayesian networks (Algorithm 1). Prior to the procedure, a hierarchy of the feature space has to be determined by hierarchical clustering. An initial variable grouping is determined by cutting the dendrogram into k clusters and cluster representatives are calculated as first principal components. A target variable can be chosen, which is kept separated. Then, a Bayesian network structure is learned using the discretized version of the cluster representatives and parameters are fitted.

Moreover, we implemented a refinement algorithm for such group Bayesian networks via a divisive hill-climbing approach (Algorithm 2). The current network model is used to predict class probabilities of the target variable using all remaining nodes as evidence, and a prediction score is calculated. As usual for hill-climbing approaches, in each step, all neighbouring states of the current model are evaluated. Those include all models, in which one group was split into two smaller groups along the dendrogram. From the neighbouring states, the model with the highest score improvement is chosen. The procedure is repeated until no further improvement is possible. Random restarts and perturbations are possible to escape from local optima.

To reduce the computation time, the tested splits may be restricted to the Markov blanket of the target variable or a certain maximal distance in the current network. This requires the initial grouping to be detailed enough, so that all important direct relations could be learned. The plot of the aggregation levels for different cluster numbers may help to choose an initial number of clusters that gives a good trade off between data compression and information loss.

If useful, further features besides the target can be chosen to be separated from their groups as well, as for example sex or age that are well-known confounders in many problems.

Objective function. Throughout the refinement, we use the the cross-entropy as objective function for a binary outcome, also known as log-loss, weighted by the class proportions. It

can be calculated as

$$H(o, p) = -\sum_{i=1}^N w_i o_i \log(p_i),$$

where $o \in \{0, 1\}^N$ is the vector of observations, $p \in [0, 1]^N$ is the vector of predicted class probabilities, and $w \in (0, 1)^N$ is the vector of weights with $w_i = \frac{\# \{o=o_i\}}{N}$. Using the class proportions as weights ensures that both outcome classes have an equal share in the total score, independent of their proportion in the training data. The adjustment is important, as often the target variable is heavily imbalanced. Without these weights, the optimization prioritizes models that primarily predict the majority class, as those have high accuracy. In case of a continuous target variable the objective function must be respectively altered.

To account for the stochasticity of the probability estimates p_i , which are based on likelihood-sampling, we estimate an uncertainty range of $H(o, p)$ over 20 runs and accept a more complex model only if its score exceeds this range.

Algorithm 1: Group Bayesian network

```

1: procedure GROUPBN( $D, g, t$ )
2:                                     //  $D$ : dataset,  $g$ : feature grouping
3:                                     //  $t$ : name of target variable
4:
5:    $D_g \leftarrow$  AGGREGATE( $D, g$ )       // aggregate data in groups  $g$ 
6:    $D_{g,t} \leftarrow$  SEPARATE( $D_g, t$ )   // separate  $t$  from its cluster
7:    $S \leftarrow$  BNSL( $D_g, t$ )           // structure learning
8:    $P \leftarrow$  BNPL( $D_{g,t}, S$ )        // parameter learning
9:    $M \leftarrow (S, P)$ 
10:
11:  return  $M$                           // return group BN model
12: end procedure

```

Algorithm 2 Adaptive Refinement

```

1: procedure GROUPBN_REFINEMENT( $D, H, k, t$ )
2:                                     //  $D$ : dataset,  $H$ : feature hierarchy
3:                                     //  $k$ : initial number of groups
4:                                     //  $t$ : name of target variable
5:
6:    $g \leftarrow$  CUT( $H, k$ )              // cut the hierarchy into  $k$  groups
7:    $M \leftarrow$  GROUPBN( $D, g, t$ )       // learn initial group network
8:    $c \leftarrow$  LOSS( $M, t$ )             // calculate loss function for target
9:
10:  repeat                               // refinement step
11:     $B \leftarrow$  MARKOVBLANKET( $M$ )     // set of splits to be tested
12:
13:    for  $b$  in  $B$  do                     // Evaluate all neighbouring models
14:       $g_b \leftarrow$  SPLIT( $H, g, b$ )    // split cluster  $b$  according to  $H$ 
15:       $M_b \leftarrow$  GROUPBN( $D, g_b, t$ ) // and learn new model
16:       $c_b \leftarrow$  LOSS( $M_b, t$ )
17:    end for
18:
19:    if  $\min c_b < c$  then              // if improvement is possible
20:       $b^* \leftarrow$  which.min( $c_b$ )
21:       $g \leftarrow g_{b^*}$ 
22:       $M \leftarrow M_{b^*}$               // Replace  $M$  with best model
23:       $c \leftarrow c_{b^*}$ 
24:    else break
25:
26:  end repeat

```

```

27:
28:   return M                               //return refined group BN model
29: end procedure

```

Hierarchical clustering and data aggregation

To identify groups of similar variables, an agglomerative similarity-based hierarchical variable clustering is used. As the method needs to be applicable to high-dimensional and heterogeneous data (qualitative and quantitative variables), we used the algorithm implemented in the ClustOfVar package in R [50]. A key step of the clustering is the determination of a synthetic central variable for each cluster, which is calculated as the first principal component from the PCAmix method [51]. PCAmix combines principal component analysis and multiple correspondence analysis. For this procedure, the data matrices are internally standardized, concatenated, and factorized respectively. The homogeneity of a cluster is calculated as the distance of all cluster variables and its representative. This distance is based on squared correlation and correlation coefficient.

Bayesian networks

A *Bayesian network* (BN) is a pair (G, Θ) , where G is the structure that represents a random vector $X = (X_1, \dots, X_n)$ and its conditional dependencies via a directed acyclic graph. Θ is the set of parameters. The parameter set Θ consists of the local conditional probabilities of each node X_i given its parents in the graph. Throughout this section, we denote the set of parents of a node X_i by $\Pi(X_i)$. The parameters are of the form

$$\theta_i = \mathbb{P}(X_i | \Pi(X_i)).$$

In case of discrete random variables they are conditional probability tables. A Bayesian network encodes the local Markov property, that is, each variable X_i is independent of its nondescendants conditioned on its parents. A general factorization of the joint probability distribution of X_1, \dots, X_n is given by

$$\mathbb{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbb{P}(X_i | X_{\Pi(i)}) = \prod_{i=1}^n \theta_i$$

accordingly. The *Markov blanket* of a node contains its children, its parents and its children's parents. It can be shown, that given the nodes in the Markov blanket, a node is conditionally independent of all other nodes in the network. It, thus, contains all the nodes that are most important for predicting the node itself. The *moralized* counterpart of a Bayesian network is an undirected graph in which each node is connected to its full Markov blanket. It can be constructed by adding arcs between all nodes that have a common child and are not directly connected.

Data discretization. The majority of the available BN structure learning algorithms assume that all variables in a Bayesian network are discrete. Hybrid approaches that can handle a mixture of discrete and continuous features include parametric models (i.e., Conditional Linear Gaussian Networks), with the drawback that they restrict the type of distribution and the structure space. More complex nonparametric approaches (see for example Schmidt et al. [52]) are computationally demanding and do not scale well to high-dimensional data. As an alternative, continuous features may be discretized. This simplifies the interpretation and enables the use of well-established algorithms for discrete Bayesian networks. Thus, we decided to discretize the cluster representatives prior to structure learning. Note that for clustering itself, the unprocessed data are used. As the cluster representatives are often multimodal,

we use an unsupervised, density-approximative discretization approach. First, significant peaks in the estimated probability density function of a variable are determined. These peaks are then used to initialize a one-dimensional k-means clustering. This procedure allows the binning, and the number of bins itself, to be directly estimated from the data. If only one significant peak is present, distribution quartiles are used for binning.

Equivalence classes of Bayesian networks (CPDAGs). As several graph structures encode the same conditional independence statements (*Markov Equivalence*), they cannot be distinguished based on observational data alone. As usual, we use completed partially directed acyclic graphs (CPDAG) to represent the inferred equivalence class. In a CPDAG, arcs with undetermined direction are drawn as undirected arcs.

Bayesian network structure learning. Our general approach, does not depend on a specific structure learning algorithm, but works with every available one. For the reported applications, we used the score-based hill-climbing algorithm, as implemented in the bnlearn package [53]. The BIC was chosen as the target function, as it is locally and asymptotically consistent and does not include any hyperparameters. The BIC of a model structure G is defined as

$$\text{BIC}(G | \mathcal{D}) := \log \mathbb{P}(\mathcal{D} | G) + \frac{d}{2} \log(N),$$

where d is the model dimension (the number of free parameters) and N is the number of observations. The BIC is asymptotically and locally consistent and decomposes to parts that are only dependent on one variable X_i and its parents $\Pi(X_i)$. For categorical random variables X_1, \dots, X_m , these parts can be calculated as

$$\text{BIC}(X_i, \Pi(X_i) | \mathcal{D}) := - \sum_j \sum_k N_{ijk} \log \frac{N_{ijk}}{\sum_j N_{ijk}} - \frac{q_i(r_i - 1)}{2} \log(N), \quad (5)$$

where N_{ijk} is the number of observations in which $X_i = k$ and $\Pi_G(X_i) = j$, q_i is the number of possible states of the parents $\Pi_G(X_i)$ and r_i the number of possible states of X_i itself.

Throughout the adaptive refinement steps, the hill-climbing procedure was initialized with the current network structure and the two new groups, formed by splitting, were embedded into this structure. To escape from local optima, 10 restarts were performed in each run with a number of perturbations depending on the total network size (10% of current number of arcs, at least 1).

Structure learning was repeated 200 times using nonparametric bootstrapping to reduce the number of false positive arcs and add only arcs with high confidence to the model (*model averaging*). The confidence threshold for inclusion of an arc was determined using adaptive thresholding, as suggested in [54].

Bayesian network parameter learning. A Bayesian parameter estimation was performed using the previously determined structure. We used a uniform prior and an imaginary sample size of 1.

Simulating networks

To generate noisy and heterogeneous data with latent group structure, we sampled two-layered Bayesian networks (Fig 3A) with a layer of (latent) group variables (layer 1), as well as a layer of noisy child variables, reflecting the information of the group variables plus measurement noise (layer 0). Arcs among group variables were sampled using Melancon's and Philippe's Uniform Random Acyclic Digraphs algorithm, which generates graphs with a uniform probability distribution over the set of all directed acyclic graphs. Child nodes were then connected to every group node. We parameterized the group variables using a randomly chosen Dirichlet

distribution, whereas the child nodes could have both, a continuous or discrete range, to simulate heterogeneity. Random noise was introduced via the parameters. For continuous features, a Gaussian noise was added; for discrete features, the distribution was respectively altered. We used these network models to simulate random samples from the joint distribution using forward sampling. By this, several simulated datasets could be created based on the same network model. They were used to assess the quality of the different approaches of group network inference under varying group size, noise level, sample size and network size. Data sampling and network learning were repeated 100 times for each scenario. In the standard network inference approach, the grouping was disregarded for network learning. Instead, a detailed network structure was learned among all variables in layer 0, which was afterwards used to identify groups and group network structure. For identification of the groups, hierarchical community detection was used. The resulting dendrogram was cut at each level, and the grouping that was closest to the true grouping in terms of the evaluation metric was chosen. To aggregate the detailed network into a group network, the ground-truth grouping was applied. As arcs between variables of different groups were only rarely learned, an arc was added to the group network, whenever at least one arc between any two variables from two groups was present. For the group network inference approach, the respective steps of the proposed algorithm were applied.

Evaluation metrics

Partition metric. To compare different variable groupings, we used an entropy-based partition metric [55]. It is zero, if two groupings are identical, and returns a positive value otherwise.

Structural hamming distance (SHD). To compare learned Bayesian network structures to the true latent structure, we used the Structural Hamming Distance (SHD). The SHD of two CPDAGs is defined as the number of changes that have to be made to a CPDAG to turn it into the one that it is being compared to. It can be calculated as the sum of all false positive, false negative and wrongly directed arcs. In order to evaluate the quality of inferred group networks, we calculated the SHD of the inferred network and the ground-truth model, and normalized it to the number of arcs within the ground-truth model.

Area under the curve. To evaluate the discriminative performance of a model, we compared the area under the receiver-operator (AUROC) as well as the precision-recall curve (AUPRC) in a 10-fold cross validation setting. We calculated the metrics using the PRROC package [56, 57].

SHIP-trend data preprocessing

The initial set of features was the same for both SHIP Trend examples. As a first step, the set of participants was reduced to those, for which the related diagnosis was present. Further steps included the removal of context-specific variables and features with high amounts of missing values.

NAFLD. As target variable for the NAFLD-specific analysis of the SHIP Trend data, we chose the presence of hepatic steatosis diagnosed based on liver MRI. An MRI of the liver was conducted and evaluated for a subset of 2463 participants of the cohort. Probands with a significant intake of alcohol (more than 20 g/day in women, more than 30 g/day in men based on the last 30 days) were excluded from the analysis. Features related to sonography of the liver or earlier diagnoses of steatosis were removed, too ($n = 14$). From the original dataset, we further removed features that contained more than 20% of missing values ($n = 59$, S4 Fig). The threshold was chosen to remove measurements that were done for specific patient subgroups only

Table 3. Processing times.

	NAFLD	Hypertension
number of features	407	328
number of probands	2311	4403
hierarchical clustering	9m 55s	13m 26s
initial group BN	1m 08s	2m 57s
group BN refinement (per iteration)	2m 34s	5m 11s

Individual processing times are stated for initial hierarchical clustering, learning of an initial group BN, and the average time needed for one refinement iteration. It must be noted that processing times depend highly on the chosen structure learning algorithm, the number of groups and the number of neighbored models.

<https://doi.org/10.1371/journal.pcbi.1008735.t003>

(like, e.g., hormone measurements, differential haematology). Our final dataset comprises 2311 participants and 407 features. The prevalence of NAFLD is 18%.

Hypertension. In SHIP Trend, blood pressure of each proband was measured three times. The average pressure of the latter two measurements was used for diagnosis of hypertension. Proband were classified as hypertensive if their measured systolic pressure exceeded 140 mmHg or the diastolic pressure exceeded 90 mmHg or they reported to receive antihypertensive treatment. Our hypertension model is based on data of 4403 participants (2123 cases of hypertension). From the original dataset we excluded features, that had more than 20% of missing values ($n = 63$, [S4 Fig](#)). We removed all features that contain further information on the blood pressure and earlier diagnoses or treatment of hypertension ($n = 35$), as well as 54 features related to medication that was related to treatment of hypertension or had extremely low variance (e.g., multiple forms of beta blockers).

Cross-validation

For comparison of the predictive power of different liver scores, logistic regression and Bayesian network models, we split the data into 10 folds. The liver scores did not have to be trained and were applied to all 10 folds separately to obtain mean and standard deviation. Bayesian network models were trained ten times on 9 of 10 folds and tested on the remaining fold, as usual. As comparison, a regularized logistic regression model was trained and tested. The same folds were used for all tests

Computations and code availability

All computations were performed using R version 3.6.2 [58] on a Unix workstation with 16 GB RAM and an eight-core Xeon E5-1620 v3 processor running Ubuntu 16.04.6. An implementation of Algorithms 1 and 2 is available on CRAN [35]. Processing times for Hypertension and NAFLD-models are given in [Table 3](#).

Supporting information

S1 Fig. Simulation results: Influence of sample size and network size. Results of the reconstruction of group networks for varying sample sizes. **A** Group networks with 5 nodes. **B** Group networks with 20 nodes. On the basis of these simulations, we decided to run the remaining simulations with group networks of size 20 and a medium sample size of 500. (TIF)

S2 Fig. Steatosis network. Group Bayesian network with target variable steatosis. (TIF)

S3 Fig. Hypertension network. Group Bayesian network with target variable hypertension. (TIF)

S4 Fig. Missing values in SHIP Trend data. Histograms of missing values in % per variable **A** for the subset of participants included in steatosis model **B** for the subset of participants included in hypertension model. (TIF)

S1 Data. Steatosis grouping. Grouping of steatosis network. Features are sorted by their centrality in the cluster. (CSV)

S2 Data. Hypertension grouping. Grouping of hypertension network. Features are sorted by their centrality in the cluster. (CSV)

Author Contributions

Conceptualization: Ann-Kristin Becker, Lars Kaderali.

Data curation: Ann-Kristin Becker, Marcus Dörr, Stephan B. Felix, Fabian Frost, Hans J. Grabe, Markus M. Lerch, Matthias Nauck, Uwe Völker, Henry Völzke.

Formal analysis: Ann-Kristin Becker.

Funding acquisition: Lars Kaderali.

Investigation: Ann-Kristin Becker.

Methodology: Ann-Kristin Becker, Lars Kaderali.

Project administration: Lars Kaderali.

Resources: Marcus Dörr, Stephan B. Felix, Fabian Frost, Hans J. Grabe, Markus M. Lerch, Matthias Nauck, Uwe Völker, Henry Völzke, Lars Kaderali.

Software: Ann-Kristin Becker.

Supervision: Lars Kaderali.

Validation: Ann-Kristin Becker.

Visualization: Ann-Kristin Becker.

Writing – original draft: Ann-Kristin Becker.

Writing – review & editing: Marcus Dörr, Stephan B. Felix, Fabian Frost, Hans J. Grabe, Markus M. Lerch, Matthias Nauck, Uwe Völker, Henry Völzke, Lars Kaderali.

References

1. Markowitz F, Spang R. Inferring cellular networks—A review. *BMC Bioinformatics*. 2007; <https://doi.org/10.1186/1471-2105-8-S6-S5> PMID: 17903286
2. Amin MT, Khan F, Imtiaz S. Fault detection and pathway analysis using a dynamic Bayesian network. *Chemical Engineering Science*. 2019; <https://doi.org/10.1016/j.ces.2018.10.024>
3. Kaderali L, Radde N. Inferring gene regulatory networks from expression data. *Studies in Computational Intelligence*. 2008;

4. Liu F, Zhang SW, Guo WF, Wei ZG, Chen L. Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLoS Computational Biology*. 2016; <https://doi.org/10.1371/journal.pcbi.1005024> PMID: [27479082](https://pubmed.ncbi.nlm.nih.gov/27479082/)
5. Chen YC, Wheeler TA, Kochenderfer MJ. Learning discrete Bayesian networks from continuous data. *Journal of Artificial Intelligence Research*. 2017; <https://doi.org/10.1613/jair.5371>
6. Lakho S, Jalbani AH, Vighio MS, Memon IA, Soomro SS, Soomro QuN. Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network. *Sukkur IBA Journal of Computing and Mathematical Sciences*. 2017; <https://doi.org/10.30537/sjcms.v1i2.51>
7. Koski TJ, Noble J. A review of Bayesian networks and structure learning. *Mathematica Applicanda*. 2012; 40(1).
8. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT press; 2009.
9. Nojavan A F, Qian SS, Stow CA. Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling and Software*. 2017; <https://doi.org/10.1016/j.envsoft.2016.10.007>
10. Sturlaugson LE, Sheppard JW. Principal component analysis preprocessing with Bayesian networks for battery capacity estimation. In: *Instrumentation and Measurement Technology Conference (I2MTC)*, 2013 IEEE International. IEEE; 2013. p. 98–101.
11. Aragam B, Gu J, Zhou Q. Learning Large-Scale Bayesian Networks with the sparsebn Package. *Journal of Statistical Software*. 2019; 91(11):1–38. <https://doi.org/10.18637/jss.v091.i11>
12. Gámez JA, Mateo JL, Puerta JM. Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*. 2011;
13. Li J, Shi J, Satz D. Modeling and analysis of disease and risk factors through learning Bayesian networks from observational data. *Quality and Reliability Engineering International*. 2008; <https://doi.org/10.1002/qre.893>
14. Rodin A, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). *Bioinformatics*. 2005;
15. Gendelman R, Xing H, Mirzoeva O, Sarde P, Curtis C, Feiler H, et al. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. *Cancer Research*. 2017; <https://doi.org/10.1158/0008-5472.CAN-16-0512> PMID: [28087598](https://pubmed.ncbi.nlm.nih.gov/28087598/)
16. Srinivas K, Rani BK, Govrdha A. Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*. 2010; 2.02:250–255.
17. Fuster-Parra P, Tauler P, Bennasar-Veny M, Ligeza A, López-González AA, Aguiló A. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Computer Methods and Programs in Biomedicine*. 2016; <https://doi.org/10.1016/j.cmpb.2015.12.010> PMID: [26777431](https://pubmed.ncbi.nlm.nih.gov/26777431/)
18. Bayat S, Cuggia M, Kessler M, Briançon S, Le Beux P, Frimat L. Modelling access to renal transplantation waiting list in a French healthcare network using a Bayesian method. *Studies in Health Technology and Informatics*. 2008; PMID: [18487797](https://pubmed.ncbi.nlm.nih.gov/18487797/)
19. Onisko a, Druzdzal MJ, Wasyluk H. A Bayesian network model for diagnosis of liver disorders. *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering*. 1999; 2.
20. Multani P, Niemann U, Cypko M, Kuehn J, Voelzke H, Oeltze-Jafra S, et al. Building a Bayesian Network to Understand the Interplay of Variables in an Epidemiological Population-Based Study. In: *Proceedings—IEEE Symposium on Computer-Based Medical Systems*; 2018. p. 88–93.
21. Völzke H, Fung G, Ittermann T, Yu S, Baumeister S, Dörr M, et al. A new, accurate predictive model for incident hypertension. *Journal of Hypertension*. 2013; PMID: [24077244](https://pubmed.ncbi.nlm.nih.gov/24077244/)
22. Lo L, Wong ML, Lee KH, Leung KS. Exploiting modularity and hierarchical modularity to infer large causal gene regulatory network. *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2015*. 2015; p. 1–8.
23. Meunier D, Lambiotte R, Bullmore ET. Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*. 2010; <https://doi.org/10.3389/fnins.2010.00200> PMID: [21151783](https://pubmed.ncbi.nlm.nih.gov/21151783/)
24. Nefian AV. Learning SNP dependencies using embedded Bayesian networks. In: *IEEE Computational Systems, Bioinformatics Conference*; 2006. p. 1–6.
25. Parviainen P, Kaski S. Learning Structures of Bayesian Networks for Variable Groups. *Int J Approx Reasoning*. 2017; 88(C):110–127. <https://doi.org/10.1016/j.ijar.2017.05.006>
26. Michael T, Maere S, Bonnet E, Joshi A, Saey Y, den Bulcke T, et al. Validating module network learning algorithms using simulated data. *BMC bioinformatics*. 2007; 8(2):S5. <https://doi.org/10.1186/1471-2105-8-S2-S5> PMID: [17493254](https://pubmed.ncbi.nlm.nih.gov/17493254/)
27. Segal E, Pe'er D, Regev A, Koller D, Friedman N. Learning module networks. *Journal of Machine Learning Research*. 2005; 6(Apr):557–588.

28. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*. 2003; 34(2):166. <https://doi.org/10.1038/ng1165> PMID: 12740579
29. Gyftodimos E, Flach P. Hierarchical Bayesian Networks: A Probabilistic Reasoning Model for Structured Domains. Proceedings of the ICML-2002 Workshop on Development of Representations. 2002;
30. Mourad R, Sinoquet C, Leray P. A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC bioinformatics*. 2011; 12(1):16. <https://doi.org/10.1186/1471-2105-12-16> PMID: 21226914
31. Njah H, Jamoussi S, Mahdi W. Deep Bayesian network architecture for Big Data mining. *Concurrency and Computation: Practice and Experience*. 2019; 31(2):e4418. <https://doi.org/10.1002/cpe.4418>
32. Ong MS. A Bayesian network approach to disease subtype discovery. *Methods in Molecular Biology*. 2019;
33. Bouhamed H, Masmoudi A, Lecroq T, Rebaï A. Structure space of Bayesian networks is dramatically reduced by subdividing it in sub-networks. *Journal of Computational and Applied Mathematics*. 2015; 287:48–62. <https://doi.org/10.1016/j.cam.2015.02.055>
34. Zainudin S, Deris S. Combining clustering and Bayesian network for gene network inference. In: *Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on*. vol. 2. IEEE; 2008. p. 557–563.
35. Becker AK. GroupBN: Learn Group Bayesian Networks using Hierarchical Clustering, R package version 0.2.0; 2020. Available from: <https://CRAN.R-project.org/package=GroupBN>.
36. Lê S, Josse J, Husson F. FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*. 2008; 25(1):1–18.
37. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software, Articles*. 2012; 50(13):1–16.
38. Völzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology*. 2010; 40(2):294–307.
39. Drescher HK, Weiskirchen S, Weiskirchen R. Current status in testing for nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH). *Cells*. 2019; 8(8):845. <https://doi.org/10.3390/cells8080845>
40. Buzzetti E, Pinzani M, Tsochatzis EA. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism: Clinical and Experimental*. 2016; <https://doi.org/10.1016/j.metabol.2015.12.012>
41. Bedogni G, Bellentani S, Miglioli L, Masutti F, Passalacqua M, Castiglione A, et al. The fatty liver index: A simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterology*. 2006; <https://doi.org/10.1186/1471-230X-6-33> PMID: 17081293
42. Lee J, Kim D, Kim H, Lee C, Yang J, Kim W, et al. Hepatic steatosis index: A simple screening tool reflecting nonalcoholic fatty liver disease. *Digestive and Liver Disease*. 2010; <https://doi.org/10.1016/j.dld.2009.08.002> PMID: 19766548
43. Meffert PJ, Baumeister SE, Lerch MM, Mayerle J, Kratzer W, Völzke H. Development, external validation, and comparative assessment of a new diagnostic score for hepatic steatosis. *The American journal of gastroenterology*. 2014; 109(9):1404. <https://doi.org/10.1038/ajg.2014.155> PMID: 24957156
44. Yip TCF, Ma A, Wong VWS, Tse Y, Chan HLY, Yuen P, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. *Alimentary Pharmacology and Therapeutics*. 2017; <https://doi.org/10.1111/apt.14172>
45. Margariti E, Deutsch M, Manolakopoulos S, Papatheodoridis GV. Non-alcoholic fatty liver disease may develop in individuals with normal body mass index. *Annals of gastroenterology*. 2012; 25(1):45. PMID: 24713801
46. Kim D, Kim WR. Nonobese fatty liver disease. *Clinical Gastroenterology and Hepatology*. 2017; 15(4):474–485. <https://doi.org/10.1016/j.cgh.2016.08.028> PMID: 27581063
47. Franceschini N, Chasman DI, Cooper-DeHoff RM, Arnett DK. Genetics, ancestry, and hypertension: implications for targeted antihypertensive therapies. *Current hypertension reports*. 2014; 16(8):461. <https://doi.org/10.1007/s11906-014-0461-9> PMID: 24903233
48. Völzke H, Fung G, Ittermann T, Yu S, Baumeister SE, Dörr M, et al. A new, accurate predictive model for incident hypertension. *Journal of hypertension*. 2013; 31(11):2142–2150. <https://doi.org/10.1097/HJH.0b013e328364a16d> PMID: 24077244
49. Sun D, Liu J, Xiao L, Liu Y, Wang Z, Li C, et al. Recent development of risk-prediction models for incident hypertension: An updated systematic review. *PloS one*. 2017; 12(10). <https://doi.org/10.1371/journal.pone.0187240>

50. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*. 2012; <https://doi.org/10.18637/jss.v050.i13>
51. Chavent M, Kuentz V, Labenne A, Liquet B, Saracco J. Multivariate Analysis of Mixed Data. R package. 2017;.
52. Schmidt M, Morup M. Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*. 2013; <https://doi.org/10.1109/MSP.2012.2235191>
53. Scutari M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*. 2010; 35(3):1–22. <https://doi.org/10.18637/jss.v035.i03>
54. Scutari M, Nagarajan R. Identifying significant edges in graphical models of molecular networks. *Artificial Intelligence in Medicine*. 2013; 57(3):207–217. <https://doi.org/10.1016/j.artmed.2012.12.006> PMID: [23395009](https://pubmed.ncbi.nlm.nih.gov/23395009/)
55. Weisman D, Simovici DA. Several remarks on the metric space of genetic codes. *International Journal of Data Mining and Bioinformatics*. 2012; <https://doi.org/10.1504/IJDMB.2012.045534> PMID: [22479816](https://pubmed.ncbi.nlm.nih.gov/22479816/)
56. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. *PLoS ONE*. 2014; <https://doi.org/10.1371/journal.pone.0092209> PMID: [24651729](https://pubmed.ncbi.nlm.nih.gov/24651729/)
57. Grau J, Grosse I, Keilwagen J. PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. 2015;
58. R Core Team. *R: A Language and Environment for Statistical Computing*; 2020.

ARTICLE II

1 **Discovering association patterns of individual serum Thyrotropin concentrations using**
2 **machine learning: An example from the Study of Health in Pomerania (SHIP)**

3 Ann-Kristin Becker¹, Till Ittermann², Markus Dörr^{3,6}, Stephan B. Felix^{3,6}, Matthias Nauck^{4,6},
4 Alexander Teumer², Uwe Völker^{5,6}, Henry Völzke^{2,6}, Lars Kaderali^{1,6,*}, Neetika Nath^{1,*}

5
6 ¹Institute of Bioinformatics, University Medicine Greifswald, Felix-Hausdorff-Str. 8,
7 17475 Greifswald, Germany

8 ²Institute for Community Medicine, SHIP/Clinical-Epidemiological Research, University
9 Medicine Greifswald, Greifswald, Germany

10 ³Department of Internal Medicine B, University Medicine Greifswald, Greifswald,
11 Germany and German Centre for Cardiovascular Research (DZHK), partner site
12 Greifswald, Greifswald, Germany

13 ⁴Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald,
14 Greifswald, Germany

15 ⁵Interfaculty Institute of Genetics and Functional Genomics, Department of Functional
16 Genomics, University Medicine Greifswald, Greifswald, Germany

17 ⁶DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald,
18 Greifswald, Germany

19
20 *Corresponding Authors: neetika.nath@uni-greifswald.de; lars.kaderali@uni-greifswald.de

21 **Abstract**

22 **Background:** Thyrotropin, also known as thyroid-stimulating hormone (TSH), is the primary
23 diagnostic and an important monitoring target in thyroid disease. The main treatment goal of
24 thyroid dysfunction is the renormalization of the thyroid function, monitored by serum TSH
25 concentration. However, its strong dependence on external factors makes it challenging to
26 specify a generally valid optimum. The goal of our study was to discover and decipher
27 association patterns of individual serum TSH concentrations from broad cohort study data to
28 complement the study situation with a wholistic view.

29 **Method:** We here propose a machine learning workflow that includes random forests and
30 Bayesian networks to allow for automatic data processing and a more straightforward

31 interpretation of complex association patterns. Based on this workflow, we discover and
32 interpret broad patterns of individual serum TSH concentration using data from the large
33 population-based Study of Health in Pomerania (SHIP).

34 **Results:** The presented model achieves good predictive accuracy and outperforms existing
35 models (root mean square error of 0.66, mean absolute error of 0.55, coefficient of
36 determination of $R^2=0.15$). Moreover, we identify 62 relevant features from the final random
37 forest model, ranging from general health variables over dietary and genetic factors to
38 physiological, hematological and hemostasis parameters. A Bayesian network model is used to
39 put these features into context and make the black-box random forest model more
40 understandable.

41 **Conclusion:** We demonstrate that the combination of random forest and Bayesian network
42 analysis is helpful to reveal broad association patterns of individual thyrotropin concentrations.
43 The discovered patterns are in line with state-of-the-art literature. They may be useful for future
44 thyroid research and improved dosing of therapeutics.

45

46 **Introduction**

47 Machine learning (ML) based on epidemiological data offers an attractive approach to discover
48 predictive patterns in complex biomedical systems, such as thyroid homeostasis. However,
49 many ML models mainly aim at high predictive accuracy and do not offer easy model
50 interpretability and explainability, which is often necessary for healthcare applications and
51 research. Random forests (RF) [1] are such an example. They achieve high predictive accuracy
52 by ensemble learning of a multitude of decision trees. However, in contrast to single decision
53 trees, RFs may lead to a variety of complex decision paths, which makes them challenging to
54 interpret. Therefore, they are usually considered black-box models. This work aims at
55 complementing a black-box RF model with post-hoc Bayesian network analysis.

56 Bayesian networks are probabilistic models describing (in-)dependence structures among
57 random variables. They are well interpretable and offer an intuitive visualization of feature
58 interactions. The presented combined workflow allows identifying predictive patterns from
59 epidemiological and clinical data and permits it to visualize and understand the nature of these
60 patterns. We apply the proposed workflow to data from the Study of Health in Pomerania
61 (SHIP) [2] in order to identify broad predictive patterns of individual serum thyroid-stimulating
62 hormone (TSH) concentrations.

63 TSH is a central component of the thyroid homeostatic system and a major diagnostic as well
64 as an important therapy monitoring target in thyroid dysfunction. With a prevalence estimated
65 at around 1-10%, thyroid dysfunction is one of the major endocrine disorders in Europe [3] and
66 worldwide [4]. It causes a wide range of symptoms, including changes in the gastrointestinal
67 system, heart rate, mood, skin, sexual function, and sleep. However, due to the mild or
68 unspecific nature of these symptoms, thyroid dysfunction often stays undetected. Yet, it has
69 been shown that even mild long-term imbalances of thyroid hormone levels increase
70 cardiovascular risk, risk of dementia, and bone disorders, amongst others [5,6]. TSH is
71 produced by the anterior pituitary gland and stimulates the thyroid gland to secrete thyroxine
72 (T4), which is then further converted to triiodothyronine (T3). Elevated levels of free T3 and
73 free T4 in the blood plasma, in turn, inhibit the production of TSH via a negative feedback loop.
74 Thus, serum TSH is a sensitive and easily accessible indicator of thyroid (dys-) function.

75 However, TSH is not steadily released from the pituitary gland but follows circadian and
76 ultradian rhythms [7]. Moreover, TSH levels in serum fluctuate depending on life phases,
77 reaching exceptionally high levels during periods of growth, stress, or pregnancy.
78 Consequently, the individual TSH level depends on various external factors, including sex, age,
79 diet, or stress level [8]. The treatment goal in thyroid dysfunction is the renormalization of the
80 thyroid function, monitored by the TSH level, but the optimum seems to be highly individual

81 and may even be genetically predetermined [9–11]. With TSH being a central marker and
82 treatment target for thyroid dysfunction, there is considerable interest in investigating patterns
83 associated with the individual TSH concentration in serum. Identified patterns may be highly
84 valuable for therapeutic decision-making.

85 The relation of TSH to other thyroid hormones is complex and nonlinear [12,13]. Such complex
86 relations can best be investigated by taking advantage of flexible ML models. Consequently,
87 advanced ML methods, especially RFs, outperform simpler models in predicting TSH, as
88 recently shown by Santhanam et al. [14]. In their study, the best scoring model was RF and
89 achieved a coefficient of determination of $R^2=0.13$. The model was based on a small set of
90 preselected thyroid-related features, including free thyroxine (FT4), free triiodothyronine
91 (FT3), autoantibodies to thyroid peroxidase (anti-TPO), as well as Body Mass Index (BMI),
92 age and ethnicity.

93 Apart from that, in large parts, existing related literature focuses only on the ML-based
94 classification of thyroid disease [15,16]. However, clinical reference ranges sometimes fail to
95 distinguish actual disease states from ordinary fluctuations in case of complex, multifactorial
96 diseases like thyroid dysfunction. Especially since serum TSH levels within the reference range
97 are also known to vary by age, sex, the applied assay, and the population's background iodine
98 status [17], labels derived from TSH alone may be imprecise. Therefore, we considered the
99 problem as a regression problem. Nevertheless, also in the classification case, decision-tree-
100 based algorithms were found to score superiorly.

101 To date, it is still unclear how to automatize the prediction from high-dimensional clinical data,
102 and how to present results of a complex ML model, such that it can be interpreted easily by
103 medical professionals as well as non-experts. The state-of-the-art to interpret RFs is to use
104 global feature importance (FI). FI measures the global influence of every individual feature on

105 the model and may be used to create a ranking. Model internal measures may be used as FI,
106 such as the increase in homogeneity in the trees' leaves. Apart from that, external measures are
107 available that evaluate the FI on out-of-bag data. One such example is permutation-based
108 feature importance, which was introduced initially for RFs and later generalized [18]. However,
109 all these FI measures neglect feature interactions. Thus, the resulting ranking may suffer from
110 disruptive effects in the presence of heterogeneity and multicollinearity, as present in SHIP
111 data: Continuous features or features with many categories offer more flexibility and may gain
112 higher importance than binary features. Moreover, permutation of one feature alone may result
113 in unrealistic data instances, so associated features may bias the importance score. Lastly,
114 indirect effects and confounders cannot be noticed from the ranking alone. To take feature
115 associations into account and offer an interpretation that goes beyond a ranking, we present a
116 workflow that complements the RF model by a Bayesian network analysis. As Bayesian
117 network structure learning from data is highly computationally expensive, we reduce the feature
118 set by extracting potentially relevant features from the RF model for this step.

119 The presented workflow allows to identify and explain broad patterns from high-dimensional
120 data. We apply this workflow to predict the individual serum TSH concentration from clinical
121 data and to identify broad and interpretable clinical patterns of thyroid functionality from the
122 model. As a basis, we use data from the Study of Health in Pomerania (SHIP) [2], which
123 includes nutritional patterns, complete blood counts, sociodemographic data, health status,
124 mood, medication, and detailed thyroid examinations of 4308 adult individuals. Additionally,
125 genetic information in the form of single nucleotide polymorphisms (SNPs) is used. While
126 many of the discovered factors have been analyzed in univariate studies before, to our best
127 knowledge, this is the first thyroid study applying ML-based algorithms to identify multivariate
128 patterns of such broadness.

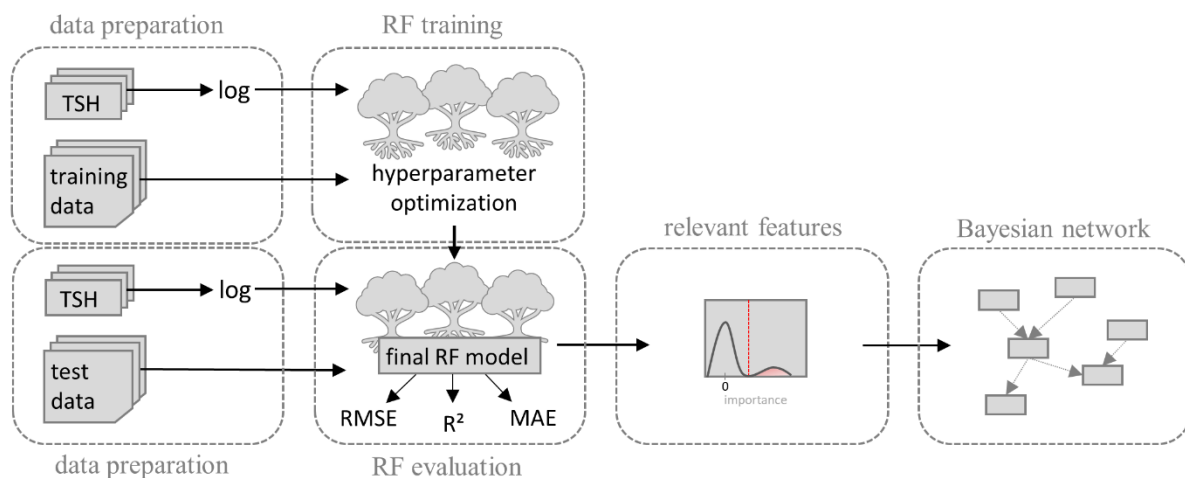
129

130 **Results and Discussion**

131 **Workflow**

132 We propose a workflow that complements a RF model with a Bayesian network analysis for
133 the post-hoc interpretation of inferred global predictive patterns. The network analysis allows
134 an interpretation that goes beyond a ranking by global feature importance.

135 The workflow consists of three steps (Fig 1) and starts with careful training of a RF model.
136 Afterward, relevant features are identified from the model using two different feature
137 importance measures. Due to the high number of features included, we use a statistical mixture
138 model approach to distinguish relevant from irrelevant features. The latter are modeled by a
139 component around a FI of zero. This step is followed by Bayesian network structure learning
140 among all relevant predictors, yielding an interpretable feature association model.



141

142 **Fig 1. Workflow.** Schematic representation of the workflow. After data preparation, a RF
143 model is trained using nested cross-validation. Relevant predictors are identified based on two
144 feature importance measures and a mixture model approach. Lastly, feature interactions among
145 the relevant predictors are examined in a Bayesian network analysis.

146

147 **Data Description**

148 The preprocessed dataset includes 602 features and 3,989 probands (49% female, 51% male,
149 for details on data preprocessing, see Methods). The mean age of study participants is 49 years
150 (min=20, max=81). In addition to serum TSH, thyroid function evaluation in SHIP includes
151 FT3, FT4, and anti-TPO measurements. Besides, urinary iodine and enlargement of the thyroid
152 ('goiter') were assessed. Results of a sonography examination of the thyroid were included in
153 terms of thyroid volume, echogenicity, and the presence of at least one thyroid nodule.
154 Descriptive statistics about these features are reported in Table 1. Additional features from
155 SHIP include nutritional patterns, complete blood counts, sociodemographic information,
156 health status, and medication. Moreover, we added 67 prefiltered SNPs to examine possible
157 genetic predispositions [19].

158 **Table 1. Descriptive statistics of thyroid examination results from SHIP.** Mean, standard
159 deviation, median and skewness are presented for continuous features. For categorical features,
160 the exact distribution is shown. The analysis is based on n=3,989 probands.

SHIP Variable	Description	Mean	StDev	Median	Skewness
tsh	Thyroid stimulating hormone (TSH) [mU/l]	0.89	2.28	0.66	25.5
log_tsh	log-transformed TSH	-0.45	0.73	-0.41	-0.65
ft3	free triiodothyronine [pmol/l]	5.25	0.88	5.2	1.24
ft4	free thyroxine [pmol/l]	12.84	3.82	12.5	1.24
sd_volg	total sonography volume of the thyroid	21.54	12.57	18.8	3.35
jodid_u	Iodide (urine) [µg/dl]	14.42	11.64	12.5	5.1
tpo_ak	anti-TPO antibodies [IU/l]	90.28	294.28	45.1	25.47
SHIP Variable	Description	No		Yes	
node_s0	presence of thyroid nodule(s)	3299 (77.2%)		975 (22.8%)	
echogenthyr_s0	hypoechoic thyroid pattern	3958 (92.7%)		313 (7.3%)	
goiter_s0	enlargement of the thyroid gland	2660 (62.2%)		1611 (37.8%)	

161

162 **Random Forest Predictions**

163 Our RF model achieves an RMSE of 0.663 (± 0.003), coefficient of determination (R^2) of 0.15
 164 (± 0.002), and a mean absolute error (MAE) of 0.55 (± 0.003) on unseen data. For evaluation
 165 purposes, we compared the achieved prediction scores to a baseline model trained on the same
 166 dataset with TSH values randomly shuffled. The shuffling breaks all relations of TSH to the
 167 remaining data; thus, the baseline predictions can be interpreted as scores achieved by random
 168 guessing (Table 2).

169

170 **Table 2. Evaluation of the final RF model for the prediction of TSH.** As a baseline
 171 comparison, we trained a similar model on the same dataset where TSH values have been
 172 randomly shuffled. The scores given in the column (random) baseline prediction thus represent
 173 scores achieved by random guessing. Average results are presented together with standard
 174 deviations given in brackets.

Evaluation criteria	prediction of TSH (\pm SD)	(random) baseline prediction (\pm SD)
RMSE Training	0.63 (± 0.041)	0.70 (± 0.004)
RMSE Test	0.66 (± 0.003)	0.72 (± 0.301)
R^2 Training	0.23 (± 0.003)	0.0001 (± 0.002)
R^2 Test	0.15 (± 0.002)	0.0004 (± 0.011)
MAE Training	0.52 (± 0.002)	0.52 (± 0.001)
Test Test	0.55 (± 0.003)	0.62 (± 0.111)

175 **Extraction of Relevant Features**

176 We identified 62 from the originally 602 features as relevant in the RF model. For the
 177 identification, we used two different FI measures and a statistical mixture model approach (see
 178 Methods). All relevant features are reported in S1 Table. The highest importance scores were
 179 found for age, FT3, FT4, anti-TPO antibodies, goiter, nodules, and thyroid hypoechogenicity
 180 in sonography (S1 and S2 Fig). The 62 extracted relevant predictors are distributed across the
 181 categories basic patient information (8), information about the general health status (5), thyroid

182 examinations (8), metabolism (9), SNPs (3), socioeconomic status (8), diet (5), immune system
183 (6), hematological and hemostasis parameters (5), hormones (3), and electrolyte levels (2).

184 **Bayesian Network Analysis**

185 In order to investigate the association patterns of TSH, we complement the RF model with a
186 better interpretable Bayesian network. Whereas RFs are optimized with respect to high
187 predictive power alone, Bayesian networks are probabilistic models of feature interactions.
188 They allow examining how features are associated and how these associations affect the
189 outcome. Their graphical representation allows for intuitive interpretation, also for non-experts.
190 Based on the methodology described in an earlier study [20], we train a Bayesian network,
191 including those features that were identified as relevant in the RF model. To reduce the
192 network's complexity, we aggregated highly collinear features to represent them as one single
193 node in the network. The network structure among the resulting 54 nodes, which are reported
194 in S1 Fig, was then learned by a score-based structure learning approach. The final Bayesian
195 network structure (Fig 2) has 128 edges. The nodes *sex* and *age*, number of medications taken
196 during the last seven days (*medic7d_s0*), and hip circumference (*som_huef*) are hub nodes in
197 the network. Additionally, we identified four different clusters within the network that refer to
198 the categories socioeconomic status, metabolism, hematological and hemostasis factors, and
199 thyroid examinations (S1 Table). The Markov blanket of a node in a Bayesian network is the
200 set of directly dependent variables in the network, i.e., those features that are most important to
201 predict a particular variable and have a direct influence on it. The Markov blanket of the TSH
202 level contains the predictors that have also been identified as the most important predictors
203 based on global feature importance in the RF model. The average Markov blanket size of the
204 whole network is 7.3, showing a relatively strong association among all predictors.

205 The extracted relevant features and their associations reveal broad clinical patterns of thyroid
206 functionality. As expected, the top features include a person's age and thyroid-specific

207 examinations (FT3, FT4, anti-TPO, and sonography results, see S1 Table). The strong
208 association of these features with the individual TSH level is also reflected by their closeness
209 in the Bayesian network (nodes colored in red, TSH is dark red, Fig 2).

210 In addition to age, the set of relevant features includes a set of further general patient
211 information and clinical measures (e.g., sex, body height, hip circumference, heart rate, blood
212 pressure) as well as measures describing the health status of a person (number of doctoral visits,
213 subjective physical and mental health, medication). As expected, sex and age influence nearly
214 every other feature; they are hub nodes in the Bayesian network, as are the amount of
215 medication and hip circumference (blue nodes, Fig 2).

216 Moreover, eight of the 62 relevant features describe a person's socioeconomic status and are
217 related to occupation, education, or family status. The Bayesian network shows that they are
218 closely tied to a person's age and health status so that their influence on the TSH level is
219 presumably only indirect (nodes colored in gray, Fig 2).

220 The same holds for most included dietary factors (amount of grey bread, cake, fresh fruits) that
221 we found mainly related to health status and age. On the contrary, we found the daily amount
222 of alcohol influencing the mean corpuscular volume and the liver status, both well-established
223 markers of alcohol use [21]. Alcohol consumption is thereby indirectly linked to TSH. Also,
224 coffee consumption was indirectly linked via the serum potassium level. Its association with
225 the thyroid was studied extensively in targeted studies [22,23].

226 Moreover, factors related to the status of liver and kidney, hematological and hemostasis status,
227 immunity, hormones, lipid and glucose metabolism, as well as electrolyte levels have been
228 identified as relevant. Three of the 67 SNPs are included as well; two of them are highly related
229 and occur in the phosphodiesterase 8B gene; the third one is a variant near FOXE1, which is
230 also known as thyroid transcription factor 2. All three have been associated with altered TSH
231 levels in earlier studies [19]. The active inclusion of SNPs into the RF model affirms the genetic
232 component of thyroid (dys-) function.

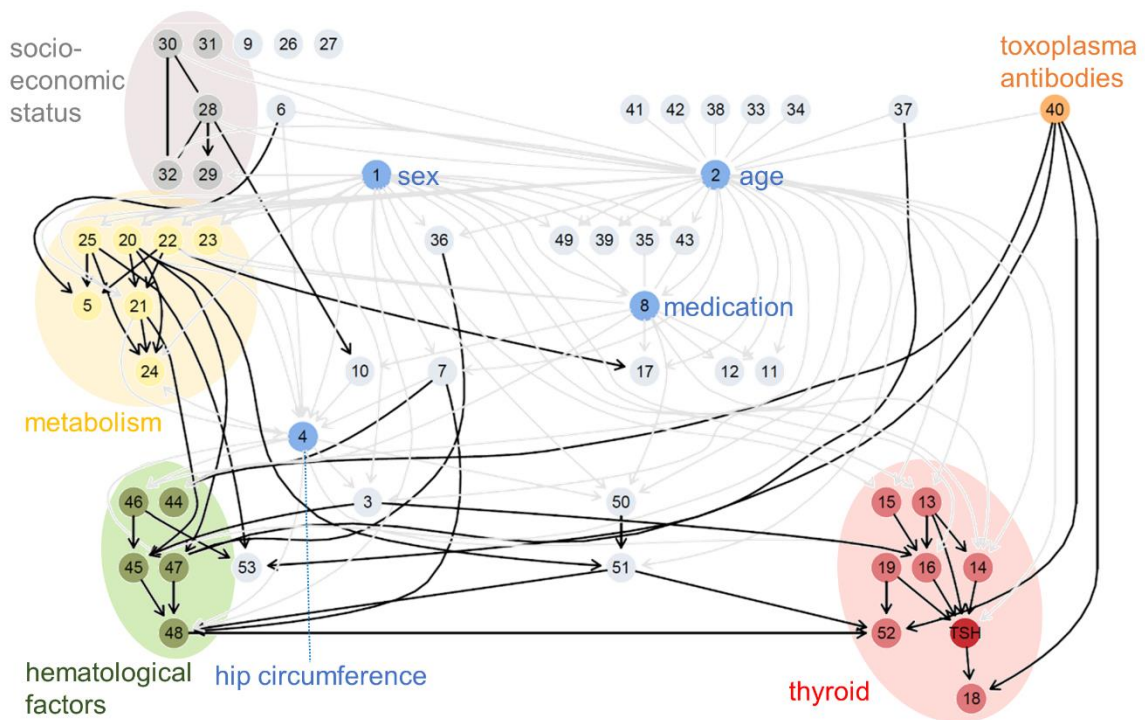
233 The association of TSH with liver and kidney markers can be explained physiologically. It is
234 well known that thyroid hormones affect renal physiology, hepatic function, and bilirubin
235 metabolism. We here identified an altered glomerular filtration rate, altered serum creatinine
236 levels, and levels of serum uric acid as associated with TSH. A correlation in the case of healthy
237 as well as diseased thyroid states has been observed before [24,25]. With thyroid hormones
238 regulating the basal metabolic rate of hepatocytes, it is no surprise that changes in the thyroid
239 homeostatic system also go along with hepatic disorders. In our model, the presence of hepatic
240 steatosis, serum aspartate-aminotransferase levels, ferritin levels, but also serum glucose,
241 lipase, and triglycerides appear to be relevant predictors. While an association of liver and
242 thyroid disease has been examined for a long time, it is still under debate if this correlation is
243 independent of the metabolic syndrome or can be fully explained by alterations in glucose and
244 lipid metabolism [26–29].

245 In the Bayesian network, the correlation between metabolic measures (nodes colored in yellow,
246 Fig 2) and thyroid is predominantly explained by health status and hematological and
247 hemostasis parameters (nodes colored in green, Fig 2). Hematological parameters (like mean
248 corpuscular volume, hematocrit, number of leukocytes) and hemostasis parameter (like partial
249 thromboplastin time and fibrinogen) are widespread general measures for health and disease.
250 However, it is also well-known that overt hypothyroidism is associated with a bleeding
251 tendency, while hyperthyroidism leads to increased coagulation and decreased fibrinolysis.
252 Recent studies suggest that coagulation factors have a mediating role between thyroid and
253 cardiovascular abnormalities [30,31]. Our Bayesian network model is in line with this
254 hypothesis, as it links metabolic factors to thyroid hormones mainly via hematological and
255 hemostasis factors. However, more targeted studies are needed to evaluate this association
256 further.

257 Additionally, several included features are related to vaccinations or infections, including
258 immunity against rubella, measles, toxoplasmosis, or helicobacter pylori. Most of them appear

259 to be primarily correlated to age with only minor independent effects on TSH (e.g., insufficient
 260 immunization against rubella or measles is more frequent in older people). However, the level
 261 of antibodies against toxoplasmosis (node colored in orange, Fig 2) is strongly related to thyroid
 262 hormone levels, predominantly to free T3. The association between *Toxoplasma gondii*
 263 infections and thyroid dysfunction was observed in earlier studies as well, underlining the role
 264 of toxoplasmosis antibodies as an independent predictor of thyroid hormone levels [32,33].
 265 Thyroid function also plays an essential role in the balancing of sex hormones. We identified
 266 serum prolactin, sex hormone-binding globulin, and insulin-like growth factor 1 (IGF-1) levels
 267 as predictors of TSH. Already earlier, subclinical hypothyroidism was shown to increase the
 268 prevalence of overt hyperprolactinemia, particularly in women. It was also shown to indirectly
 269 influence IGF-1 and sex hormone-binding globulin (SHBG) levels [34–36].

270



271

272 **Fig 2. Inferred Bayesian network structure among the extracted relevant predictors and**
 273 **the TSH level.** The four hub nodes, sex, age, medication (taken during the last seven days), and
 274 hip circumference are colored in blue. Arcs originating from the hub nodes are plotted in light

275 gray to make the network more readable. The TSH level is colored in dark red, thyroid-related
 276 examinations in red. Yellow nodes refer to metabolic factors, green nodes to hematological and
 277 hemostasis factors, and grey nodes to socioeconomic parameters. Antibody titer against
 278 toxoplasmosis is presented in orange. Further information on the features can be found in Table
 279 S1.

280 **Random Forest Prediction from Feature Subgroups**

281 To complement the analysis, we examined how well a RF can predict individual TSH
 282 concentrations from one of the identified feature subgroups 'metabolism', 'socioeconomic
 283 status', and 'hematological factors' alone. Table 3 reports the performance scores of predicting
 284 TSH measures when a new RF model was trained using only sex, age, and features from one
 285 category. The best performance was achieved in case of the feature subgroup 'metabolism' (Test
 286 RMSE of 0.703). However, all metrics and (especially R^2) decreased considerably due to the
 287 reduced feature sets.

288 **Table 3. RF prediction results for different feature subgroups.** Columns refer to models
 289 built based on different feature subgroups. The first two rows show the respective RF
 290 hyperparameters. The remaining six rows contain the prediction metrics achieved by the
 291 models. Average results are stated with standard deviations given in brackets.

Model	(random) baseline prediction (± SD)	All Features (± SD)	metabolism [yellow nodes] (± SD)	socioeconomic status [grey nodes] (± SD)	hematological factors [green nodes] (± SD)
RMSE Training	0.703 (± 0.003)	0.632 (± 0.003)	0.697 (± 0.003)	0.702 (± 0.003)	0.702 (± 0.003)
RMSE Test	0.719 (± 0.029)	0.662 (± 0.032)	0.703 (± 0.003)	0.704 (± 0.032)	0.705 (± 0.032)
MAE Training	0.599 (± 0.104)	0.515 (± 0.11)	0.592 (± 0.105)	0.598 (± 0.104)	0.598 (± 0.104)

MAE Test	0.618 (± 0.106)	0.551 (± 0.11)	0.599 (± 0.111)	0.601 (± 0.111)	0.601 (± 0.111)
R ² Training	0.045 (± 0.008)	0.229 (± 0.0035)	0.061 (± 0.002)	0.046 (± 0.002)	0.046 (± 0.002)
R ² Test	-0.0004 (± 0.002)	0.149 (± 0.023)	0.042 (± 0.021)	0.037 (± 0.015)	0.037 (± 0.015)

292 **Conclusion**

293 In summary, the presented model successfully predicts individual TSH concentrations from a
294 broad set of features. We demonstrate that the combination of RF and Bayesian network
295 analysis is useful to reveal and interpret broad association patterns. A complementary network
296 analysis can overcome classical drawbacks of RF interpretation based on feature importance
297 only and is helpful for high-quality interpretation. The identified predictive patterns are in line
298 with recent findings and give new insights into thyroid functionality. The most important
299 predictors included a person's age and thyroid-specific parameters: FT3, FT4, anti-TPO, and
300 sonography results (S1 Fig). It must be noted that relevant predictors were successfully revealed
301 automatically from an extensive set of features. Yet, the presented model yields better
302 prediction results than models built from a small, manually chosen feature set. Sex, age, hip
303 circumference, and medication intake during the last seven days were further identified as hub
304 nodes in the Bayesian network of relevant predictors. Based on the network, clusters of related
305 features could be identified and further tested for their predictive capacity. However, a large
306 fraction of variance in the data remains unexplained. Possibly, parts of the leftover variance are
307 due to temporal fluctuations. The inclusion of individual temporal profiles, or at least
308 measurements at more than one time point, could further increase the accuracy. Due to the large
309 number of features, the final model lacks validation on external data. Such validation is
310 challenging, as an appropriate dataset would need to include all of the used features with similar
311 measurement protocols. However, the discovered patterns are well supported by state-of-the-
312 art literature.

313 In contrast to often-used classification models, the presented regression model is independent
314 of clinical TSH reference limits or the diagnosis of specific dysfunctions. Thus, it may also be
315 used to detect disease initiation or minor abnormalities. Our study underlines the need for
316 careful interpretation of complex models. It shows that a ranking by global feature importance
317 is not enough to interpret intricate predictive patterns and can be misleading. The identified
318 association patterns may be useful for future thyroid research and improved dosing of
319 therapeutics.

320 **Materials and Methods**

321 **Study Population**

322 The Study of Health in Pomerania (SHIP) is a population-based study carried out in West
323 Pomerania, the north-east area of Germany [2,37]. A sample from the population aged 20 to 79
324 years was drawn from population registries. First, the three cities of the region (with 17,076 to
325 65,977 inhabitants) and the 12 towns (with 1,516 to 3,044 inhabitants) were selected, and then
326 17 out of 97 smaller towns (with less than 1,500 inhabitants) were drawn at random. Second,
327 from each of the selected communities, subjects were drawn at random, proportional to the
328 population size of each community and stratified by age and gender. Only individuals with
329 German citizenship and main residency in the study area were included. Finally, 7,008 subjects
330 were sampled, with 292 persons of each gender in each of the twelve five-year age strata. In
331 order to minimize dropouts by migration or death, subjects were selected in two waves. The net
332 sample (without migrated or deceased persons) comprised 6,267 eligible subjects. Selected
333 persons received a maximum of three written invitations. In case of non-response, letters were
334 followed by a phone call or by home visits if contact by phone was not possible. The SHIP
335 population finally comprised 4,308 participants (corresponding to a final response of 68.8%).

336 **Data Preprocessing**

337 In addition to phenotypical features from SHIP (including nutritional patterns, complete blood
338 counts, sociodemographic data, health status, mood, medication, and detailed thyroid
339 examinations), additional data about 67 SNPs were included in our analysis, that have
340 previously been shown to be associated with thyroid dysfunction in a genome-wide association
341 study (GSWA) [19]. From the original dataset, features with more than 20% of missing values
342 were removed, and the remaining data were imputed using a nonparametric, RF-based
343 imputation procedure [38]. We further removed participants under anti-thyroid medication
344 (n=280) and those for which the information about anti-thyroid medication was missing (n=37).
345 We also removed participants with extremely high TSH measurements that exceeded 60 mU/l
346 (n=2). In total, 3989 participants and 602 features (67 SNPs) were used for further analysis. As
347 the distribution of TSH concentrations is heavily right-skewed, we log-transformed the target
348 variable for further analysis to make its distribution more symmetric.

349 **Genotyping**

350 Non fasting blood samples were drawn from the cubital vein in the supine position. The samples
351 were taken between 07:00 AM and 04:00 PM, and serum aliquots were prepared for immediate
352 analysis and storage at -80 °C in the Integrated Research Biobank (Liconic, Liechtenstein). The
353 SHIP samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0.
354 Hybridization of genomic DNA was done following the manufacturer's standard
355 recommendations. Genetic data were stored using the database Caché (InterSystems).
356 Genotypes were determined using the Birdseed2 clustering algorithm. For quality control
357 purposes, several control samples were added. On the chip level, only subjects with a
358 genotyping rate on QC probe sets (QC call rate) of at least 86% were included. Finally, all
359 arrays had a sample call-rate of above 92%. The overall genotyping efficiency of the GWA was

360 98.55 %. Imputation of genotypes in SHIP was performed using the software IMPUTE v2.2.2
361 based on the 1000 Genomes release Mar 2012 ALL populations reference panel. SNPs with a
362 Hardy-Weinberg-Equilibrium p-value <0.0001 or a call rate <0.8 were removed before
363 imputation.

364 **Evaluation Criteria**

365 To evaluate the predictive capacity of a model, we use the root mean square error (RMSE),
366 coefficient of determination (R^2), and mean absolute error (MAE). The MAE is the mean of the
367 absolute differences between data and predictions. It is non-negative, with a value of zero
368 indicating a perfect prediction. Its value is dependent on the scale of the outcome variable.
369 Similarly, the RSME can be calculated as the quadratic mean of these differences. Lastly, the
370 coefficient of determination measures the proportion of the variance that is explained by the
371 model; it ranges from 0 to 1.

372 **Random Forests**

373 RFs are ensemble models that combine a multitude of decision trees [1]. They output the mean
374 prediction of the individual trees, which are guaranteed to be decorrelated due to the use of
375 bootstrap samples and random feature subsets of the training data. RFs are considered black-
376 box models, as it is very difficult to retrace how the model came to a specific prediction. In
377 order to reduce the bias in model selection, we applied nested 10-fold cross-validation for
378 hyperparameter optimization. The results from hyperparameter optimization and the final
379 parameter settings are reported in S3 Fig and S2 Table. For training of the RF model, we used
380 the R-package randomForest [39]. We optimize the three main hyperparameters, which control
381 the structure and depth of the forest, based on internal 10-fold cross-validation in a grid-search:
382 the minimum size of terminal nodes (*nodesize*, tested values: 15, 40, 65), the maximal number
383 of terminal nodes (*maxnodes*, tested values: 15, 40, 65), and the number of variables randomly

384 sampled as candidates at each split (*mtry*, tested values 3-50). External cross-validation was
385 then applied to evaluate the performance of the final model on unseen data. The RSME was
386 chosen as an objective function; additionally, MAE and R^2 of a model are used as evaluation
387 criteria.

388 **Measures of Feature Importance**

389 Feature importance in the RF model was assessed using two different measures: node purity
390 and incremental mean square error (IncMSE). Node purity measures the increase in
391 homogeneity (here in terms of variance) of the labels at the respective node. The final node
392 purity value of a feature is the sum over all splits in which the feature is chosen, averaged over
393 all trees. Conversely, the IncMSE is a permutation-based FI measure, and it is calculated as the
394 difference in the overall out-of-bag error before and after permutation of the feature (S1 Fig).

395 **Extraction of relevant predictors**

396 Due to the high number of included features, likely, most of them are not relevant for the
397 prediction of TSH in the RF model. Thus, most features have importance scores close to zero
398 S1 Fig. That is why we used an approach based on linear mixture models to distinguish between
399 relevant and irrelevant features. We assume that irrelevant features are closely distributed
400 around zero, with a small amount of variation due to the inherent randomness of the FI
401 measures. In the case of the IncMSE, we used a normal distribution around zero to model all
402 irrelevant features (S2 A Fig). As the node purity takes only positive values, we assume it to be
403 gamma-distributed instead (S2 B Fig). We consider those features as relevant for which the
404 mean importance is larger than the respective 0.999-quantile, which means they most likely
405 stem from the set of features with importance greater than zero. Fitting resulted in a mean of
406 0.1 and a standard deviation of 0.85 for the component around zero in case of IncMSE, and

407 shape parameter 0.62 and scale parameter 4.65 for the component around zero in case of node
408 purity.

409 **Bayesian Network Analysis**

410 We additionally analyze the feature interrelations of relevant features using a Bayesian network
411 approach. Bayesian networks are a prominent tool for probabilistic reasoning in artificial
412 intelligence, and they model the joint distribution of a feature set in terms of a directed acyclic
413 graph. In the graph, the nodes refer to the features (resp. random variables) X_1, \dots, X_n and arcs
414 model conditional (in)dependencies among them. The joint distribution factorizes efficiently
415 according to the graph structure, allowing for efficient computation and approximate inference.
416 It can be evaluated based on local probabilities depending only on a node's parents in the graph:

$$417 \quad P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \mathit{parents}(X_i)) \quad (1)$$

418 The visualization in terms of a network is intuitive even for non-experts and supports human
419 interpretation. Thus, a Bayesian network structure can help to understand complex associations,
420 identify confounding factors, and make multicollinearity visible. We use a score-based hill-
421 climbing approach to learn a conditional Gaussian Bayesian network that models heterogeneous
422 data [40,41]. For the learning of the Bayesian network, we used the Bayesian information
423 criterion (BIC) as objective function, which is a penalized likelihood criterion. We also apply a
424 model averaging approach to reduce false-positive arcs [42]. Before learning the network, we
425 aggregate highly collinear features based on feature similarity for heterogeneous variables to
426 reduce the network's complexity [41,43]. A list of features, including aggregated features, is
427 given in S1 Table.

428 **Computations**

429 We report the different R packages that were used in our analysis with their version in S3 Table.

430 **Acknowledgment**

431 SHIP is part of the Community Medicine Research network of the University of Greifswald,
432 Germany, which is funded by the Federal Ministry of Education and Research (grants no.
433 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social
434 Ministry of the Federal State of Mecklenburg-West Pomerania, and the network' Greifswald
435 Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of
436 Education and Research (grant 03IS2061A). Genome-wide data have been supported by the
437 Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from
438 Siemens Healthineers, Erlangen, Germany and the Federal State of Mecklenburg- West
439 Pomerania. The University of Greifswald is a member of the Caché Campus program of the
440 InterSystems GmbH. This work was further supported by the project Superthyreose, funded by
441 the German “Innovationsfonds des Gemeinsamen Bundesausschusses” (grant no. VSF2_2019-
442 167)

443 AKB received funding from the BMBF (LiSyM, grant number 031L0032) and gratefully
444 acknowledges an add-on-fellowship from the Joachim Herz Stiftung. LK acknowledges
445 funding from the European Union (EuCanShare, grant 825903), as well as the State of Lower
446 Saxony and the Volkswagenstiftung (Indira, grant number ZN3437). The funders had no
447 influence on study design, data analysis, study interpretation, decision to publish, and writing
448 of the manuscript.

449 **Bibliography**

450 1. Breiman L. Random forests. Mach Learn. 2001;45: 5–32.
451 doi:10.1023/A:1010933404324

- 452 2. Völzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort profile:
453 The study of health in Pomerania. *Int J Epidemiol.* 2011;40: 294–307.
454 doi:10.1093/ije/dyp394
- 455 3. Madariaga AG, Santos Palacios S, Guillén-Grima F, Galofré JC. The incidence and
456 prevalence of thyroid dysfunction in Europe: A meta-analysis. *J Clin Endocrinol Metab.*
457 2014;99: 923–931. doi:10.1210/jc.2013-2409
- 458 4. Taylor PN, Albrecht D, Scholz A, Gutierrez-Buey G, Lazarus JH, Dayan CM, et al.
459 Global epidemiology of hyperthyroidism and hypothyroidism. *Nat Rev Endocrinol.*
460 2018;14: 301–316. doi:10.1038/nrendo.2018.18
- 461 5. Biondi B, Cooper DS. The clinical significance of subclinical thyroid dysfunction.
462 *Endocr Rev.* 2008;29: 76–131. doi:10.1210/er.2006-0043
- 463 6. Peppas M, Betsi G, Dimitriadis G. Lipid Abnormalities and Cardiometabolic Risk in
464 Patients with Overt and Subclinical Thyroid Disease. *J Lipids.* 2011;2011: 1–9.
465 doi:10.1155/2011/575840
- 466 7. Ikegami K, Refetoff S, Van Cauter E, Yoshimura T. Interconnection between circadian
467 clocks and thyroid function. *Nat Rev Endocrinol.* 2019;15: 590–600.
468 doi:10.1038/s41574-019-0237-z
- 469 8. Boucai L, Hollowell JG, Surks MI. An approach for development of age-, gender-, and
470 ethnicity-specific thyrotropin reference limits. *Thyroid.* 2011;21: 5–11.
471 doi:10.1089/thy.2010.0092
- 472 9. Lee YK, Shin DY, Shin H, Lee EJ. Sex-specific genetic influence on thyroidstimulating
473 hormone and free thyroxine levels, and interactions between measurements: KNHANES
474 2013 2015. *PLoS One.* 2018;13: 1–12. doi:10.1371/journal.pone.0207446
- 475 10. Medici M, Visser TJ, Peeters RP. Genetics of thyroid function. *Best Pract Res Clin*

- 476 Endocrinol Metab. 2017;31: 129–142. doi:10.1016/j.beem.2017.04.002
- 477 11. Razvi S, Hostalek U. Therapeutic challenges in the application of serum thyroid
478 stimulating hormone testing in the management of patients with hypothyroidism on
479 replacement thyroid hormone therapy: a review. *Curr Med Res Opin.* 2019;35: 1215–
480 1220. doi:10.1080/03007995.2019.1570769
- 481 12. Brown SJ, Bremner AP, Hadlow NC, Feddema P, Leedman PJ, O’Leary PC, et al. The
482 log TSH–free T4 relationship in a community-based cohort is nonlinear and is influenced
483 by age, smoking and thyroid peroxidase antibody status. *Clin Endocrinol (Oxf).* 2016;85:
484 789–796. doi:10.1111/cen.13107
- 485 13. Clark PM, Holder RL, Haque SM, Hobbs FDR, Roberts LM, Franklyn JA. The
486 relationship between serum TSH and free T4 in older people. *J Clin Pathol.* 2012;65:
487 463–465. doi:10.1136/jclinpath-2011-200433
- 488 14. Santhanam P, Nath T, Mohammad FK, Ahima RS. Artificial intelligence may offer
489 insight into factors determining individual TSH level. *PLoS One.* 2020;15: e0233336.
490 doi:10.1371/journal.pone.0233336
- 491 15. Raisinghani S, Shamdasani R, Motwani M, Bahreja A, Raghavan Nair Lalitha P. Thyroid
492 prediction using machine learning techniques. *Communications in Computer and*
493 *Information Science.* 2019. doi:10.1007/978-981-13-9939-8_13
- 494 16. Mir YI, Mittal S. Thyroid disease prediction using hybrid machine learning techniques:
495 An effective framework. *Int J Sci Technol Res.* 2020;9: 2868–2874.
- 496 17. Ittermann T, Khattak RM, Nauck M, Cordova CMM, Völzke H. Shift of the TSH
497 reference range with improved iodine supply in Northeast Germany. *Eur J Endocrinol.*
498 2015;172: 261–267. doi:10.1530/EJE-14-0898
- 499 18. Fisher A, Rudin C, Dominici F. Model Class Reliance: Variable Importance Measures

- 500 for any Machine Learning Model Class, from the “Rashomon” Perspective. *J Mach*
501 *Learn Res.* 2018;20.
- 502 19. Teumer A, Chaker L, Groeneweg S, Li Y, Di Munno C, Barbieri C, et al. Genome-wide
503 analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat*
504 *Commun.* 2018;9: 1–14. doi:10.1038/s41467-018-06356-1
- 505 20. Becker AK, Dörr M, Felix SB, Frost F, Grabe HJ, Lerch MM, et al. From heterogeneous
506 healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network
507 approach. *PLoS Comput Biol.* 2021;17: 1–21. doi:10.1371/JOURNAL.PCBI.1008735
- 508 21. Aoun EG, Lee MR, Haass-Koffler CL, Swift RM, Addolorato G, Kenna GA, et al.
509 Relationship between the thyroid axis and alcohol craving. *Alcohol Alcohol.* 2015;50:
510 24–29. doi:10.1093/alcalc/agu085
- 511 22. Pietzner M, Köhrle J, Lehmpfuhl I, Budde K, Kastenmüller G, Brabant G, et al. A thyroid
512 hormone-independent molecular fingerprint of 3,5-diiodothyronine suggests a strong
513 relationship with coffee metabolism in humans. *Thyroid.* 2019;29: 1743–1754.
514 doi:10.1089/thy.2018.0549
- 515 23. Han MA, Kim JH. Coffee consumption and the risk of thyroid cancer: A systematic
516 review and meta-analysis. *Int J Environ Res Public Health.* 2017;14.
517 doi:10.3390/ijerph14020129
- 518 24. Den Hollander JG, Wulkan RW, Mantel MJ, Berghout A. Correlation between severity
519 of thyroid dysfunction and renal function. *Clin Endocrinol (Oxf).* 2005;62: 423–427.
520 doi:10.1111/j.1365-2265.2005.02236.x
- 521 25. Kimmel M, Braun N, Alscher MD. Influence of thyroid function on different kidney
522 function tests. *Kidney Blood Press Res.* 2012;35: 9–17. doi:10.1159/000329354
- 523 26. Kim HJ. Importance of thyroid-stimulating hormone levels in liver disease. *J Pediatr*

- 524 Endocrinol Metab. 2020;33: 1133–1137. doi:10.1515/jpem-2020-0031
- 525 27. Kim D, Kim W, Joo SK, Bae JM, Kim JH, Ahmed A. Subclinical Hypothyroidism and
526 Low-Normal Thyroid Function Are Associated With Nonalcoholic Steatohepatitis and
527 Fibrosis. Clin Gastroenterol Hepatol. 2018;16: 123-131.e1.
528 doi:10.1016/j.cgh.2017.08.014
- 529 28. Malik R, Hodgson H. The relationship between the thyroid gland and the liver. QJM -
530 Mon J Assoc Physicians. 2002;95: 559–569. doi:10.1093/qjmed/95.9.559
- 531 29. Jang J, Kim Y, Shin J, Lee SA, Choi Y, Park EC. Association between thyroid hormones
532 and the components of metabolic syndrome. BMC Endocr Disord. 2018;18: 1–9.
533 doi:10.1186/s12902-018-0256-0
- 534 30. Elbers LPB, Fliers E, Cannegieter SC. The influence of thyroid function on the
535 coagulation system and its clinical consequences. J Thromb Haemost. 2018;16: 634–
536 645. doi:10.1111/jth.13970
- 537 31. Bano A, Chaker L, De Maat MPM, Atiq F, Kavousi M, Franco OH, et al. Thyroid
538 Function and Cardiovascular Disease: The Mediating Role of Coagulation Factors. J Clin
539 Endocrinol Metab. 2019;104: 3203–3212. doi:10.1210/jc.2019-00072
- 540 32. Alvarado-Esquivel C, Ramos-Nevarez A, Guido-Arreola CA, Cerrillo-Soto SM, Pérez-
541 Álamos AR, Estrada-Martínez S, et al. Association between *Toxoplasma gondii*
542 infection and thyroid dysfunction: A case-control seroprevalence study. BMC Infect Dis.
543 2019;19: 1–5. doi:10.1186/s12879-019-4450-0
- 544 33. Shapira Y, Agmon-Levin N, Selmi C, Petříková J, Barzilai O, Ram M, et al. Prevalence
545 of anti-toxoplasma antibodies in patients with autoimmune diseases. J Autoimmun.
546 2012;39: 112–116. doi:10.1016/j.jaut.2012.01.001
- 547 34. Tseng FY, Chen YT, Chi YC, Chen PL, Yang WS. Serum levels of insulin-like growth

- 548 factor 1 are negatively associated with log transformation of thyroid-stimulating
549 hormone in Graves' disease patients with hyperthyroidism or subjects with
550 euthyroidism: A prospective observational study. *Medicine (Baltimore)*. 2019;98:
551 e14862. doi:10.1097/MD.00000000000014862
- 552 35. Bahar A, Akha O, Kashi Z, Vesgari Z. Hyperprolactinemia in association with
553 subclinical hypothyroidism. *Casp J Intern Med*. 2011;2: 229–233.
- 554 36. Selva DM, Hammond GL. Thyroid hormones act indirectly to increase sex hormone-
555 binding globulin production by liver via hepatocyte nuclear factor-4 α . *J Mol Endocrinol*.
556 2009;43: 19–27. doi:10.1677/JME-09-0025
- 557 37. John U, Hensel E, L demann J, Piek M, Sauer S, Adam C, et al. Study of Health in
558 Pomerania (SHIP): A health examination survey in an east German region: Objectives
559 and design. *Sozial- und Prventivmedizin SPM*. 2001;46: 186–194.
560 doi:10.1007/BF01324255
- 561 38. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for
562 mixed-type data. *Bioinformatics*. 2012;28: 112–118. doi:10.1093/bioinformatics/btr597
- 563 39. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2:
564 18–22.
- 565 40. Scutari M. bnlearn: Bayesian network structure learning. *R Packag*. 2010.
566 doi:10.1007/s10337-017-3440-x
- 567 41. Becker A-K, Kaderali L. GroupBN: Learn Group Bayesian Networks using Hierarchical
568 Clustering. R package version 0.2.0, 2020 Available from: [https://CRAN.R-](https://CRAN.R-project.org/package=GroupBN)
569 [project.org/package=GroupBN](https://CRAN.R-project.org/package=GroupBN) .
- 570 42. Scutari M, Nagarajan R. Identifying significant edges in graphical models of molecular
571 networks. *Artif Intell Med*. 2013;57: 207–217. doi:10.1016/j.artmed.2012.12.006

572 43. Chavent M, Kuentz-Simonet V, Lique B, Saracco J. ClustOfVar: An R package for the
573 clustering of variables. *J Stat Softw.* 2012;50: 1–16. doi:10.18637/jss.v050.i13

574 **Supplementary Material**

575 **S1 Fig. Variable importance measures of the top 20 features from the random forest**
576 **model.** Based on A) incremental mean square error (IncMSE) and B) Node purity, the highest
577 importance scores were found for age, FT3, FT4, anti-TPO antibodies, goiter, thyroid nodules,
578 and thyroid hypoechogenicity in sonography. The important features were extracted from
579 random forest based on the selected parameter nodesize of 15, and maxnodes of 15.

580 **S2 Fig. Estimated density of feature importance measures.** A statistical mixture model was
581 used, the component around zero (red dashed lines) was modeled as A) a normal distribution
582 for IncMSE (mean 0.1, standard deviation) B) a Gamma distribution for node purity (shape
583 parameter $p=0.62$ and scale parameter $b=4.65$). Features were identified as relevant if they had
584 feature importance larger than the respective 0.999-quantile.

585 **S3 Fig. Hyperparameter optimization.** Prediction results from the grid-based hyperparameter
586 optimization of the random forest model using 10-fold nested cross-validation. Rows show
587 results for varying values of the parameter nodesize (tested values: 15, 40, 65), and columns
588 show results for varying values of the maximal number of terminal nodes (maxnodes, tested
589 values: 15, 40, 65). The parameter mtry was optimized separately and set to 5.

590 **S1 Table. List of SHIP variables that have been identified as relevant in the RF model.**
591 The IDs refer to nodes in Fig 2, which are categorized into basic information, general health
592 status, Thyroid examination, Metabolism, SNPs, Socioeconomic Status, Diet, immune system,

593 Hematological and Hemostasis parameters, Hormones and Electrolytes. The relevant features
594 from RF model are described in third column.

595 **S2 Table. Hyperparameters of a random forest model and their description.** This table
596 reports the different RF hyperparameters that were optimized, with describing the features in
597 second column. The third column contains the parameter ranges for parameters for
598 optimization. The last column contains the final parameter selected for final RF model
599 presented above.

600 **S3 Table.** R Software packages and versions.

601

ARTICLE III

Metabolic Cross-talk Between Human Bronchial Epithelial Cells and Internalized *Staphylococcus aureus* as a Driver for Infection

Authors

Laura M. Palma Medina, Ann-Kristin Becker, Stephan Michalik, Harita Yedavally, Elisa J.M. Raineri, Petra Hildebrandt, Manuela Gesell Salazar, Kristin Surmann, Henrike Pfortner, Solomon A. Mekonnen, Anna Salvati, Lars Kaderali, Jan Maarten van Dijk, and Uwe Völker

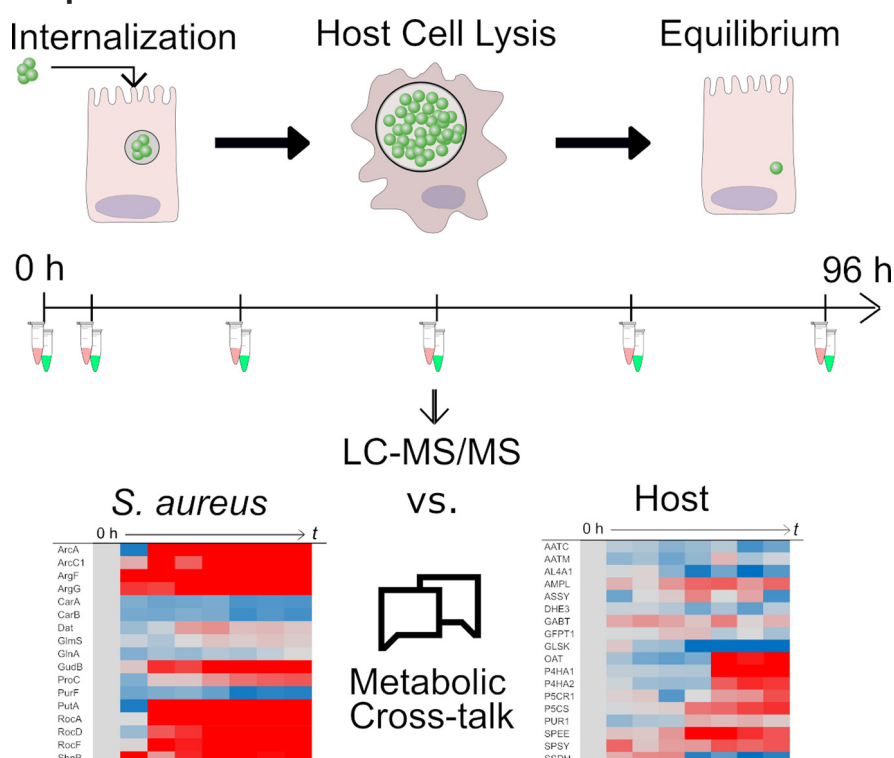
Correspondence

j.m.van.dijk01@umcg.nl;
voelker@uni-greifswald.de

In Brief

Staphylococcus aureus invades bronchial epithelial cells to reach underlying lung tissue and to escape from the human immune defenses or antibiotic therapy. The internalized pathogen achieves these objectives by differentiation into growing and dormant subpopulations. Here we tracked the dynamic interactions between internalized bacteria and their host over four days by quantitative proteomics. The results highlight metabolic cross-talk between host and pathogen as a key driver for mutual adaptation and the outcome of infection.

Graphical Abstract



Highlights

- Interplay of epithelial cells and internalized *S. aureus* was dissected over 96 h.
- Surviving host cells contain nonreplicating bacteria that persists in the cytoplasm.
- Competition over resources triggers temporal metabolic changes.
- Metabolic adaptation of host and bacteria determines the outcome of infection.



Metabolic Cross-talk Between Human Bronchial Epithelial Cells and Internalized *Staphylococcus aureus* as a Driver for Infection*

✉ Laura M. Palma Medina‡§, Ann-Kristin Becker¶, Stephan Michalik‡, Harita Yedavally||, Elisa J.M. Raineri§, Petra Hildebrandt‡, Manuela Gesell Salazar‡, Kristin Surmann‡, Henrike Pförtner‡, Solomon A. Mekonnen‡§, Anna Salvati||, Lars Kaderali¶, Jan Maarten van Dijk§**, and Uwe Völker‡‡

***Staphylococcus aureus* is infamous for causing recurrent infections of the human respiratory tract. This is a consequence of its ability to adapt to different niches, including the intracellular milieu of lung epithelial cells. To understand the dynamic interplay between epithelial cells and the intracellular pathogen, we dissected their interactions over 4 days by mass spectrometry. Additionally, we investigated the dynamics of infection through live cell imaging, immunofluorescence and electron microscopy. The results highlight a major role of often overlooked temporal changes in the bacterial and host metabolism, triggered by fierce competition over limited resources. Remarkably, replicating bacteria reside predominantly within membrane-enclosed compartments and induce apoptosis of the host within ~24 h post infection. Surviving infected host cells carry a subpopulation of non-replicating bacteria in the cytoplasm that persists. Altogether, we conclude that, besides the production of virulence factors by bacteria, it is the way in which intracellular resources are used, and how host and intracellular bacteria subsequently adapt to each other that determines the ultimate outcome of the infectious process. *Molecular & Cellular Proteomics* 18: 892–908, 2019. DOI: 10.1074/mcp.RA118.001138.**

Staphylococcus aureus is a Gram-positive opportunistic pathogen of humans, but also a commensal of the human body. Specifically, *S. aureus* is commonly found in the anterior nares of around 30% of the human population (1). Although most *S. aureus* carriers do not present any clinical symptoms, *S. aureus* can cause a wide range of diseases such as skin

and soft tissue infections, osteomyelitis, septic arthritis and pneumonia (2, 3). This pathogen has gained notoriety in recent years because of its prevalence in nosocomial infections and the rise of methicillin-resistant *S. aureus* (MRSA) (3–5).

Although *S. aureus* often acts as an extracellular pathogen, it can evade immune responses and antibiotic therapy by entering human cells. The latter strategy is also used by the bacteria as a mechanism to spread to other tissues and both professional as well as non-professional phagocytic cells are used for internalization (6–8). After the bacteria have been taken up by the host cells, they will initially be localized in vesicles, which subsequently might fuse with lysosomes or be engulfed by an isolation membrane because of autophagy, and the bacteria inside them may prevail or escape into the cytosol. Although the internalization by host cells is potentially lethal for the bacteria, the survivors will have two options: proliferation or persistence. In the first case, the bacteria replicate intracellularly and subsequently induce lysis of the host cells. The released bacteria search for new host cells to be infected and spread into new tissues (6–8). In the second case, the persistent bacteria do not multiply, but adapt to the intracellular environment and may survive intracellularly without causing clinical symptoms for extended time periods. This pattern has been linked to relapse of infections or emergence of small colony variants (SCVs)¹ of *S. aureus* which display reduced metabolic activity (9–11).

Despite strain-specific differences in overall virulence, all *S. aureus* strains, including laboratory strains, can display proliferative and persistent phenotypes. Although this phenomenon has been known (6), the actual adaptations either ena-

From the ‡Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany; §Department of Medical Microbiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; ¶Institute of Bioinformatics, University Medicine Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany; ||Division of Pharmacokinetics, Toxicology, and Targeting, Groningen Research Institute of Pharmacy, University of Groningen, Groningen, Groningen, The Netherlands

Received October 17, 2018, and in revised form, February 25, 2019

Published, MCP Papers in Press, February 26, 2019, DOI 10.1074/mcp.RA118.001138

bling active intracellular proliferation or reduced metabolic activity and persistence are still poorly understood. The precise outcome of the interplay between the bacterium and the host depends on the type of host cell involved and, perhaps most importantly, the physiological states of both parties (12, 13). The main challenge in obtaining a detailed understanding of the adaptive behavior of internalized *S. aureus* lies in the fact that it is essential to study quantitative changes over an extended period, not only in one of the two interacting parties but simultaneously in both. Previous studies have addressed these aspects only partially either by focusing on the internalized bacteria only, or over only short periods of time post infection (p.i.) (11, 12, 14–17). Yet, it is important to get the “complete picture” of such an infection scenario, because the invasion and destruction of lung epithelial cells is representative for some of the most serious staphylococcal diseases possible, especially necrotizing pneumonia.

The present study was designed to close the current knowledge gap on the interplay between *S. aureus* and lung epithelial cells by a time-resolved analysis of both parties over the longest possible period. The limits for such an analysis are set by the amount of material that can be extracted for bacteria- and host cell-specific analyses, and the parameters to be measured. This led us to a proteomics approach, where adaptations of the bronchial epithelial cell line 16HBE14o- and *S. aureus* were followed up to 4 days p.i. using a data independent acquisition (DIA) method. Importantly, our findings highlight dynamic adaptive changes, in both the host and the internalized pathogen, and describe the active cross-talk between them at different stages of infection. Additionally, we correlate these adaptations with the intracellular localization of the bacteria p.i. and the epithelial cells’ response. The observations suggest that, after a period of violent conflict, both parties reach an equilibrium phase where they are apparently at peace and the bacteria have reached a persistor status.

MATERIALS AND METHODS

Bacterial Strains—*S. aureus* strain HG001 (18) was used to perform all experiments. The bacteria carried plasmid pJL-sar-GFP to constitutively express the green fluorescent protein (GFP; Liese *et al.*, 2013). For the immunostaining protocols, a *spa* mutant was used to prevent unspecific binding of marker antibodies to protein A. The HG001 Δspa strain was kindly provided by Dr. Jan Pané-Farré, University of Greifswald. Cultivation of bacteria was performed in prokaryotic minimal essential medium (pMEM): 1x MEM without sodium bicarbonate (In-

vitrogen, Karlsruhe, Germany) supplemented with 1x non-essential amino acids (PAN-Biotech GmbH, Aidenbach, Germany), 4 mM L-glutamine (PAN-Biotech GmbH), 10 mM HEPES (PAN-Biotech GmbH), 2 mM L-alanine, 2 mM L-leucine, 2 mM L-isoleucine, 2 mM L-valine, 2 mM L-aspartate, 2 mM L-glutamate, 2 mM L-serine, 2 mM L-threonine, 2 mM L-cysteine, 2 mM L-proline, 2 mM L-histidine, 2 mM L-phenylalanine and 2 mM L-tryptophan (All from Sigma-Aldrich, Schnelldorf, Germany), adjusted to pH 7.4 and sterilized by filtration. One day before the infection of epithelial cells, bacterial overnight cultures in pMEM supplemented with 0.01% yeast extract (Sigma-Aldrich) and 10 $\mu\text{g/ml}$ erythromycin (Sigma-Aldrich) were prepared by serial dilutions (1×10^{-6} up to 1×10^{-10}) of a 100 μl glycerol stock of a bacterial culture with an OD₆₀₀ of 1.2. Incubation was performed at 37 °C and 220 rpm. The following day, the main culture was inoculated from an overnight culture with an OD₆₀₀ between 0.3 to 0.8. The starting OD₆₀₀ of the main culture was set to 0.05 and it was incubated for ~2 h in a shaking water bath at 150 rpm and 37 °C until it reached the mid-exponential phase at an OD₆₀₀ of ~0.4 (supplemental Fig. S1). The bacteria were then harvested and used for preparation of the master mix for infection as explained below in the Internalization Experiments” paragraph.

Cell Line—The human epithelial cell line 16HBE14o- is a transformed bronchial epithelial cell line originally derived from a 1-year-old heart-lung transplant patient (20). This cell line is known for its ability to form tight junctions and to differentiate. The cells were cultured at 37 °C in 5% CO₂ in eukaryotic minimal essential medium (eMEM): 1x MEM (Biochrom AG, Berlin, Germany) supplemented with 10% (v/v) fetal calf serum (FCS; Biochrom AG), 2% (v/v) L-glutamine 200 mM (PAN-Biotech GmbH) and 1% (v/v) nonessential amino acids 100x (PAN-Biotech GmbH). The splitting of cells was carried out every 3 days with 0.25% trypsin-EDTA (Gibco®, Grand Island, NY). After thawing of frozen stocks (in liquid N₂) the cells were maintained for 20 additional passages. The cell lines stocks used are not authenticated.

Experimental Design and Statistical Rationale—Four independent biological replicates of the infection set-up were used for quantification of bacterial and host cell populations, and for mass spectrometry measurements. The number of replicates was selected to ensure that at every time point there were at least three consistent measurements for every protein. The sampling of each independent infection consisted of 8 samples taken over the course of 4 days, including a 0 h sample which is the control condition. In total, 32 samples of the cytosolic proteome of bacteria and 32 samples of the human bronchial epithelial cell proteome were measured. To avoid measuring replicates of the same condition sequentially, the measuring order of each set of samples was determined by assigning a random number between 1 and 32 to each sample (function *sample* in R version 3.4.4 (21)). Additionally, samples for imaging were collected from three independent infection experiments.

To determine changes over time in protein abundance, an empirical Bayes moderated F-test was conducted for each protein profile. This test also evaluates the similarity of the replicates. The moderated *p* values were corrected for multiple testing using Benjamini and Hochberg’s multiple testing correction.

Internalization Experiments—Internalization experiments were performed essentially as described by Pfürtnner *et al.* (22). Briefly, internalization was performed using a confluent 16HBE14o- cell layer seeded at a density of 1×10^5 cells/cm² in 12-well plates, 3 days before infection. The infection was carried out at a multiplicity of infection (MOI) of 25 bacteria per host cell. The master mix for infection was prepared from a mid-exponential (OD₆₀₀ of 0.4) culture of *S. aureus* HG001 diluted in eMEM, buffered with 2.9 μl sodium hydrogen carbonate (7.5%, PAN-Biotech GmbH) per ml bacterial culture added. The growth medium over the confluent epithelial layer

¹ The abbreviations used are: SCVs, small colony variants; p.i., post infection; DIA, data independent acquisition; GFP, green fluorescent protein; pMEM, protein minimal essential medium; eMEM, eukaryotic minimal essential medium; FCS, fetal calf serum; MOI, multiplicity of infection; LC3, microtubule-associated protein 1A/1B-light chain; LAMP-1, lysosomal-associated membrane protein 1; PFA, para-formaldehyde; TEM, transmission electron microscopy; ROS, reactive oxygen species; RNS, reactive nitrogen species; BCAAs, branched-chain amino acids.

was replaced with the master mix, and the coculture was incubated for 1 h at 37 °C in 5% CO₂. Afterward, the medium was collected (non-adherent sample) and replaced with eMEM medium containing 10 μg 47 ml of lysostaphin (AMBI Products LLC, Lawrence, NY). The medium was replaced every 2 days.

For collection of the proteome samples, the culturing medium was aspirated, and the epithelial cell layers were treated for 5 min at 37 °C with UT buffer (8 M urea, 2 M thiourea in MS-grade water; Sigma-Aldrich) to generate samples for analysis by mass spectrometry (MS). If samples were intended for the collection of bacteria, the disruption of epithelial cells was performed for 5 min at 37 °C in 0.05% sodium dodecyl sulfate (SDS; Carl Roth, Karlsruhe, Germany). Samples were collected at 2.5 h, 6.5 h, 24 h, 48 h, 72 h, and 96 h p.i.

To monitor changes in the abundance of human and bacterial cells, counting was performed at the times of sample collection. Epithelial cells were counted after staining with trypan blue dye using a Countess[®] system (Invitrogen). Quantification of intracellular bacteria and infected epithelial cells was performed with a Guava[®] easyCyte flow cytometer (Merck Millipore, Darmstadt, Germany) by excitation of the GFP with a 488 nm laser and detection at 510–540 nm.

Preparation of Proteome Samples—After disruption of epithelial cells with 0.05% SDS, two million liberated bacteria were sorted by flow cytometry using a FACSAria IIIu cell sorter (Becton Dickinson Biosciences, Franklin Lakes, NJ) per time point. The recognition of bacteria was carried out by excitation with a 488 nm laser and the emission was detected in the range of 515–545 nm. The bacterial cells were collected on low protein binding filter membranes with a pore size of 0.22 μm (Merck Millipore). These bacteria-containing filters were immediately placed in Eppendorf tubes that were then frozen by transferring them to a –20 °C freezer for the course of the experiment and then kept at –80 °C until use. The bacteria on the filter were lysed by incubation for 30 min at 37 °C with 7.4 μg/ml lysostaphin in 50 mM ammonium bicarbonate (Sigma-Aldrich) (23). Digestion of bacterial proteins on the filter was performed overnight at 37 °C with 0.1% Rapigest SF surfactant (Waters, Eschborn, Germany) and 0.3 μg of trypsin (Promega, Madison, WI).

For human proteome analyses, the protein content of samples was quantified using a Bradford assay (Bio-Rad, Hercules, CA). Four μg of protein per sample were prepared for MS measurements by reduction with 2.5 mmol/L dithiothreitol (Thermo Fisher Scientific, Idstein, Germany) for 1 h at 60 °C and alkylation with 10 mmol/L iodoacetamide (Sigma-Aldrich) for 30 min at 37 °C. Then, the samples were digested overnight with trypsin (protein/trypsin 25:1) at 37 °C.

The following 16HBE14o- samples were used for the construction of the spectral library of the host: a confluent cell layer cultured in a 10 cm dish for 3 days, an apoptotic cell layer in a 10 cm dish cultured for a week, non-polarized cells cultured for 3 days over Transwells[®], and lastly polarized cells cultured for 11 days over Transwells[®] (Corning, Schnellendorf, Germany). The last two conditions were grown over 12 mm inserts with 0.4 μm pores, and with media volumes of 400 μl on the apical side and 1300 μl on the basal side of the cultures. Furthermore, to expand the host proteome library, published reads (24) of the bronchial epithelium cell line S9 were also used. These cells are immortalized cells isolated from a patient with cystic fibrosis that were transformed with a hybrid virus adeno-12-SV40 (ATCC[®] number CRL-2778) (25). For the construction of the host proteome spectral library, aliquots of the different samples of whole cell lysates of 16HBE14o- in UT buffer were mixed, and then 25 μg of the extract mixture was fractionated by SDS-PAGE. The gel was partitioned into ten protein-containing pieces that were destained by 15 min washes with ammonium bicarbonate solution (200 mM) in 50% acetonitrile (Mallinckrodt Baker, Inc., Deventer, Netherlands) at 37 °C and 500 rpm. Then, the gel pieces were dehydrated by incubation with acetonitrile at 37 °C and 500 rpm. The supernatant was discarded after-

ward. Proteins in each gel piece were in-gel digested overnight at 37 °C with 20 μl of trypsin (10 ng/μl) and 30 μl ammonium bicarbonate solution (20 mM). Lastly, the peptides were extracted by addition of 0.1% acetic acid (Carl Roth) and incubation in an ultrasound bath for 30 min. Afterward, the supernatant was collected, 50% acetonitrile with 0.05% acetic acid were added to the gel pieces for another 30 min incubation, and both supernatant fractions were united. Two of the supernatants of the ten SDS-PAGE fractions were mixed to generate five final samples with essentially the same protein quantity, which were then used for further processing and DDA-measurements.

The tryptic peptides derived from bacterial or human proteins were concentrated and purified using C₁₈ ZipTip columns (Merck Millipore). All samples were resuspended in a buffer consisting of 2% acetonitrile and 0.1% acetic acid in MS-grade water. Indexed Retention Time (iRT) peptides (Biognosys AG, Schlieren, Switzerland) were added to the samples for feature alignment, peak calibration and signal quantification. The spike in of the samples was carried out according to the manufacturer's instructions assuring that the injected volumes have one IE (injection equivalent) of iRT peptide mix. The final volume of the samples and the injection volumes were set to 12 μl and 10 μl for *S. aureus* samples, 20 μl and 5 μl for 16HBE14o- samples, and 20 μl and 4 μl for the spectral library samples, respectively.

Mass Spectrometry Measurements—Tryptic peptides were separated on an Accucore 150-C18 analytical column of 250 mm (25 cm × 75 μm, 2.6 μm C18 particles, 150 Å pore size, Thermo Fischer Scientific, Waltham, MA) using a Dionex Ultimate 3000 nano-LC system (Thermo Fischer Scientific). Peptides were eluted at a constant temperature of 40 °C and a flow rate of 300 nL/min with a 120 min linear gradient (2% to 25%) of buffer (acetonitrile in 0.1% acetic acid).

To design a spectral library MS/MS data were recorded on a Q Exactive mass spectrometer (Thermo Fischer Scientific) in data dependent mode (DDA). The MS scans were carried out in a *m/z* range of 300 to 1650 *m/z*. Data was acquired with a resolution of 70,000 and an AGC target of 3 × 10⁶. The top 10 most abundant isotope patterns with charge ≥ 2 from the survey scan were selected for fragmentation by high energy collisional dissociation (HCD) with a maximal injection time of 120 ms, an isolation window of 3 *m/z*, and a normalized collision energy of 27.5 eV. Dynamic exclusion was set to 30 s. The MS/MS scans had a resolution of 17,500 and an AGC target of 2 × 10⁵.

MS/MS analyses of samples were performed in data independent mode (DIA) on a Q Exactive Plus mass spectrometer (Thermo Fischer Scientific) following the method described by Bruderer *et al.* (26). Briefly, the data was acquired in the *m/z* range from 400 to 1220 *m/z*, the resolution for MS and MS/MS was 35,000, and the AGC target was 5 × 10⁶ for MS, and 3 × 10⁶ for MS/MS. The number of DIA isolation windows was 19 with 2 *m/z* overlap. For further details to the instrumental setup and the parameters for LC-MS/MS analysis in DDA and DIA mode see [supplemental Tables S1 and S2](#).

Immunofluorescent Confocal Microscopy—Time-lapse imaging was carried out with a DeltaVisionRT deconvolution microscope (GE Healthcare Europe GmbH, Freiburg im Breisgau, Germany). To perform the imaging, the actual infection experiment was carried out on a glass bottom 35-mm plate (MatTek, Ashland, MA). After a change of media with lysostaphin, the plate was transferred to the microscope base under incubation conditions. Imaging of the epithelial layer was performed by light microscopy, whereas GFP fluorescent bacteria were observed by excitation with a 490/20 nm mercury vapor lamp and detection of fluorescence at 528/38 nm. Image acquisition was performed every 5 min for the first 48 h, then from 48 h to 72 h and finally from 92 h to 96 h. Picture processing was performed with Fiji (<http://fiji.sc/Fiji>).

Subcellular localization of microtubule-associated protein 1A/1B-light chain (LC3) and lysosomal-associated membrane protein 1 (LAMP-1) by immunofluorescence microscopy was performed using a Leica TCS SP8 Confocal laser scanning microscope (Leica Microsystems B.V., Amsterdam, Netherlands). The cells were seeded over coverslips of 18 mm diameter 3 days before infection as described above. However, in this case a HG001 Δ spa mutant was used to avoid aspecific IgG binding. The samples were collected at 0 h, 1 h, 2.5 h, 6.5 h, 24 h, 48 h, 72 h, and 96 h by fixation with 2% para-formaldehyde (PFA, Merck Millipore) for 20 min at room temperature. Preparation of the samples for the actual microscopy was performed simultaneously after conclusion of the experiment. The samples were permeabilized with 0.5% Tween 20 (Sigma-Aldrich) for 30 min at room temperature and then nonspecific binding sites were blocked with 1% bovine serum albumin, 10% FCS in 0.07% Tween 20 for 120 min at room temperature. All antibodies were diluted in blocking solution. Primary rabbit anti-LC3B (Cat. No. 1384; Novus Biologicals, Oxon, England) and mouse CD107a (LAMP-1; Cat. No. 555798; BD, Drachten, Netherlands) antibodies were used at 1:500 and 1:100 dilutions, respectively. The incubation was carried out simultaneously for 1 h at room temperature in a humidified chamber. The secondary Goat anti-rabbit antibody conjugated with Alexa Fluor 594 (A11012; Invitrogen) and goat anti-mouse antibody conjugated with Alexa Fluor 647 (A-21236; Invitrogen) were used, both, at a 1:500 dilution with incubation for 1 h at room temperature. Lastly, the DNA was stained with 4',6-diamidino-2-phenylindole (DAPI), the slides were mounted with Mowiol® 4–88 mounting medium (EMD Chemical, Inc., Temecula, CA) and stored at -20°C until microscopic visualization.

Transmission Electron Microscopy (TEM)—After the invasion assay, 16HBE14o- cells were fixed with 0.2% glutaraldehyde Polyscience, Inc., Warrington, PA and 2% PFA in 0.1 M sodium cacodylate buffer (pH 7.4; Sigma-Aldrich) for 10 min. Subsequently, the fixative solution was replaced with new fixative solution and incubation was continued for 30 min at room temperature. The cells were rinsed twice for 5 min each in 0.1 M cacodylate buffer at room temperature followed by post-fixation in 1% Osmium tetroxide (Electron Microscopy Sciences, Hatfield, PA) 1.5% potassium ferrocyanide (Merck Millipore) in 0.1 M sodium cacodylate at 4°C for 30 min. The 16HBE14o-cells were then incubated with 1% tannic acid in 0.05 M of sodium cacodylate buffer for 5 min to enhance the color of the 16HBE14o-cell membranes and demonstrate the internalization of *S. aureus*. The cells were washed with Milli-Q water, dehydrated through serial incubation in graded ethanol (30%, 50%, 70%, and 100%) and lastly embedded in EPON resin (Hexion, Columbus, OH). Ultrathin sections (80 nm) were cut with an UC7 ultramicrotome (Leica, Vienna, Austria) and contrasted using 5% uranyl acetate for 20 min, followed by Reynolds lead citrate for 2 min. Images were recorded with a FEI CM100 transmission electron microscope operated at 80 kV using a Morada digital camera.

Identification and Quantification of Proteins—Human proteins were identified using Spectronaut™ Pulsar 11 (v11.0.18108.11.30271) software (Biognosys AG) against a human bronchial epithelial cell line generated from data-dependent acquisition measurements of 16HBE14o- samples and 12 additional data sets of the cell line S9 that were previously published (24). The spectral library construction was based on a Comet database search using a human protein database in a target-decoy approach using the trans-proteomic-pipeline (TPP) version 4.8.0 PHILAE (27, 28). The raw files were converted to mzML files with msconvert (Proteowizard version 3.0.11537; November 31, 2017) using a vendor peak picker on spectra with MS level 1–2. Then, the mzML files were searched with the Comet search engine (2014.02 rev. 0) against a human data base that comprised 20,217 Uniprot-reviewed entries (February 2018), 102 cRAP common contaminants entries (<https://www.thegpm.org/crap/>)

and 1 entry for the concatenated iRT peptides. The target-decoy version of this database was generated by adding all reverse entries resulting in 40,640 entries in total. The target-decoy search was performed with a parent mass error of ± 20 ppm, fragment mass error of 0.01 Da, and allowing full-tryptic peptides (trypsin/P cleavage rule) with up to two internal cleavage sites. The search included fixed modification of +57.021464 for carbamidomethylated cysteine and variable modification of +15.9949 for oxidized methionine. The search results were scored using PeptideProphet (29) and iProphet (30) with a minimal peptide length of 7 amino acids. The global protein FDR calculation of 0.01 was assessed with MAYU (31). The filtered peptide spectrum matches were used to generate the ion library in Spectronaut™ v11.0.15038.14.27660 (Asimov; Biognosys AG, Switzerland) setting the m/z mass range from 300 to 1,800, 6 to 10 fragments per peptide, removing fragments smaller than 3 amino acids, no segmented regression, and a minimum root mean square error of 0.5.

The *S. aureus* ion library used in this study was generated previously (32). Briefly, the data sets used for the construction of this library comprise samples of the cytosolic and exoproteomes of *S. aureus* HG001 and the isogenic Δ rho mutant ST1258. This ion library was constructed with a similar protocol as the one applied for the host library. The Comet database search was based on the fasta file from AureoWiki (33). This database comprises 2852 *S. aureus* protein entries. The final database used for the target-decoy approach contains 5944 entries in total (including cRAP contaminants). The peptide spectral matching was performed using the trypsin/P digest rule with a number of tolerable termini (NTT) of 2, a parent mass error of ± 30 ppm, fragment mass tolerance of 0.01 Da, variable modification of +15.9949 for methionine oxidation and fixed modification of +57.021464 for carbamidomethylation (sample preparation-dependent). The generation of the ion library in Spectronaut™ v11.0.15038.14.27660 resulted in a constructed library consisting of 2154 proteins with 38,570 tryptic peptides.

The Spectronaut DIA-MS analysis was carried out using dynamic MS1 and MS2 mass tolerance, dynamic XIC RT extraction window, automatic calibration, dynamic decoy strategy (library size factor = 0.1, minimum limit = 5000), protein Q-value cutoff of 0.01, precursor Q-value cutoff of 0.001. The search included variable modifications of +15.9949 for oxidized methionine, and if reduction and alkylation was used, fixed modifications of +57.021464 for carbamidomethylated cysteine. A local cross run normalization was performed using complete profiles with a Q-value < 0.001. The MS2 peak area was quantified and reported. Missing values were parsed using an iRT profiling strategy with carry-over of exact peak boundaries (minimum Q-value row selection = 0.001). Only nonidentified precursors were parsed with a Q-value > 0.0001. Ion values were parsed when at least 25% of the samples contained high quality measured values. The settings for the Spectronaut™ analyses are available in [supplemental Table S3](#). Afterward, parsed values were filtered out again, if they were more than 2-fold higher than the measured values preventing false positives because of parsing. For further analyses proteins with at least two identified peptides were considered. The identified proteins were annotated based on the Uniprot database and the gene names for *S. aureus* were extracted from AureoWiki (33).

Statistical Testing of Changes in Protein Abundances Over Time—Peptide intensities were normalized by the mean of all time points. The time-course data of each protein were then calculated as the median of all normalized peptide data belonging to that protein. Subsequently, the protein data were analyzed using the LIMMA package version 3.34.9 (34) in R version 3.4.4 (21). First, a natural regression spline with four degrees of freedom was fitted for each protein, considering that expression changes smoothly over time. Then, a linear model (35) with eight parameters was fitted to the data, includ-

ing an intercept, four spline parameters and three parameters corresponding to comparability of the four replicates. An empirical Bayes moderated F-test was conducted for each protein, which can detect very general changes over time while simultaneously testing for similarity of the replicates. Because of general differences in replicate 1 data at 6.5 h and 24 h, these 16HBE14o- cell samples were excluded for the testing only, but not for calculating the median values. The moderated p values were corrected for multiple testing using Benjamini and Hochberg's multiple testing correction. Finally, proteins with an adjusted p value below $\alpha = 0.01$ were assumed to be significantly changed during the time course of infection. For further analyses, the data were averaged over the replicates using the median for each protein and shifted to start in zero at time point 0 h.

Afterward the time course data of all significantly changed proteins were clustered using a noise-robust soft-clustering (fuzzy c-means) implemented in the Mfuzz package version 2.38.0 (36, 37) in R version 3.4.4 (21). Because the clustering is performed in Euclidian space, the expression values of proteins were standardized for clustering only to have a standard deviation of 1, so that only the dynamics were considered. The optimal number of clusters was chosen using both a cluster validity index (minimum centroid distance) and a manual validation of biologically meaningful groupings.

RESULTS

S. aureus Population Displays a Dormant State After 48 h p.i. in Human Epithelial Cells—To analyze physiological changes in *S. aureus* internalized by 16HBE14o- lung epithelial cells over an extended period of time, we applied a previously established infection assay (22) and assessed the longest time span over which bacteria could be reproducibly isolated from the infected cells in sufficient amounts to enable proteome profiling. Pilot experiments showed that this is possible for at least 96 h, which was therefore, selected as the end point for all further analyses. During these 96 h, the fate of the infected cells and the internalized bacteria was recorded through *time-lapse* microscopy as shown in the three Supplemental movies and illustrated through snapshots in Fig. 1A. Additionally, the abundance of the bacterial population during infection was measured by quantification of the mean fluorescence intensity (Fig. 1B). These data show that the internalized bacteria do not multiply significantly during the first 8 h p.i. Subsequently, most bacteria start to proliferate leading to lysis of their respective host cells. Of note, in the applied experimental setup, host cell lysis is suicidal for the bacteria because of the presence of lysostaphin in the medium. Accordingly, the liberated bacteria disappear and cannot reinfect other cells. After ~30 h p.i., nonreplicating bacteria remain detectable within the surviving host cells until the end of the experiment at 96 h p.i. These findings were corroborated by flow cytometric counting of the bacterial numbers (Fig. 1C). Notably, two distinctive bacterial phenotypes were observed over time, namely one involving high rates of intracellular bacterial replication and another one involving internalized bacteria in a dormant state. Further, although the number of host cells remains essentially constant during the first 24 h p.i., this number increases during later stages (Fig. 1C). The percentage of host cells that contain bacteria gradually de-

creases over time, starting with a maximum of ~27% at 2.5 h p.i. and ending with ~13% at 96 h p.i.

As microscopy and flow cytometry cannot reveal the intricacies of the processes occurring in host cells and internalized bacteria, we employed quantitative proteomics profiling. This approach has the advantage of quantifying many different proteins belonging to the adaptive processes in bacteria and the human host cells with good reproducibility. Thus, changes in each protein's abundance can be followed over time and compared. Through data-independent acquisition, application of an in-house built *S. aureus* HG001 spectral library (24, 32) and human cell line spectral libraries, we achieved a complete data collection for 3644 human and 930 staphylococcal proteins at all time points during 96 h p.i. (supplemental Tables S6 and S7). Because the bacterial population displays two different phenotypes with different prominence at the different stages of infection, the observed proteome dynamics describe the adaptive changes in the bacteria at each stage and the final "shape" of the persistent bacteria. Likewise, these measurements record the influence of the intracellular bacteria on the host's physiological state over time. To determine the weight of the observed changes, the generated bacterial and host proteome data were fitted to spline lines and tested against constant linearity. This linearity test considers the replicates' variability, rendering it an indicator of the reliability of the observed changes. Furthermore, all proteins that show significant changes were clustered depending on their behavior over time, thereby marking the stages p.i. at which important changes take place (supplemental Fig. S2).

The differences in bacterial and host replication rates as observed by microscopy are mirrored in the abundance of the respective ribosomal proteins (Fig. 1D–1E). Consistent with their unimpaired growth, the host cells do not show major regulatory changes in ribosomal protein abundance over the time course of infection. In contrast, the amounts of bacterial proteins that make up the small and large ribosomal subunits start to decrease after internalization, being detectable at substantially lowered amounts at 48 h p.i. This time point correlates with the absence of bacterial growth observed in the live cell imaging. An exception to this general trend is the "*Staphylococcus aureus* hibernation promotion factor" (SaHPPF), which shows significantly increased levels right from the start of infection (Fig. 1E). This protein represses translation, and therefore, it is important for survival in the stationary phase and during conditions of nutrient deprivation (38, 39).

Nutrient Competition Affects Primarily S. aureus, but Staphylococcal Persistence Triggers Changes in Host Metabolism at 48 h p.i.—The acquisition of metabolites and their subsequent conversion to cellular constituents is a fundamental feature of all living organisms. This implies that different organisms residing in the same system are likely to compete for nutrients. Accordingly, internalized *S. aureus* will compete with the host for carbon sources. This is clearly reflected in the

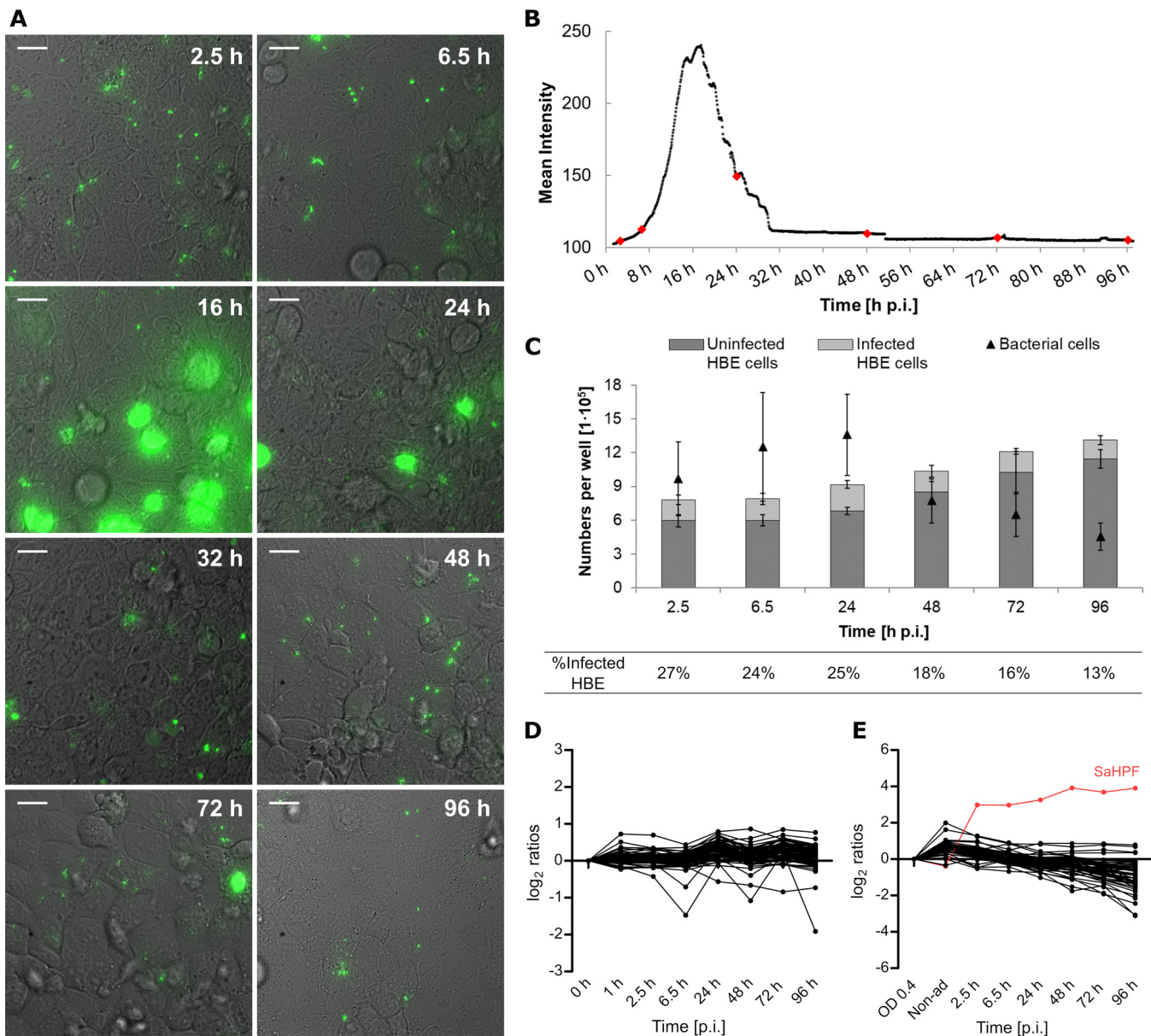


FIG. 1. After internalization two subpopulations of *S. aureus* can be distinguished by differences in replication rate. The progression of the infection was followed by time-lapse microscopy (A; Supplemental videos; scale bar = 20 μ m) and the GFP fluorescence intensity was quantified to elucidate the dynamics of the bacterial population (B). Most of the bacteria replicate intracellularly during the first day of infection, but a secondary subpopulation remains in a dormant state during the whole time of observation. The red dots indicate the time points of sample collection for further experiments. Counting of bacterial cells, and uninfected and infected host cells, as jointly presented in (C) confirmed the changes in both populations. Average values of four replicates are presented (B-C). Mass spectrometry quantification of ribosomal proteins of the host (D) or *S. aureus* (E). The list of proteins presented in the line plots is available in supplemental Table S8. The levels of the proteins were calculated based on the mid-exponential phase or the 0 h time point as reference, for the bacteria and the host, respectively. The SaHPF protein detected in higher amounts is colored in red.

proteins required for central carbon metabolism, both in the bacteria and the host. At first instance, the proteins related to the bacterial glycolytic pathway remain fairly stable after internalization displaying similar levels as encountered in exponentially growing bacteria (Fig. 2). Yet, at 6.5 h p.i., the pathogen displays changes in the levels of most of the proteins related to the glucose activation phase. Correspondingly, the

host presents similar changes, but only from 48 h p.i. onwards, which coincides with transition of the *S. aureus* population into a dormant state. Remarkably, during the entire course of infection, neither the pathogen nor its host regulates proteins related to the production of acetyl-CoA from pyruvate. During nutrient-rich conditions, *S. aureus* will produce acetate from acetyl-coA, thereby providing the NAD^+ neces-

Cross-talk of Bronchial Cells and *S. aureus* During Infection

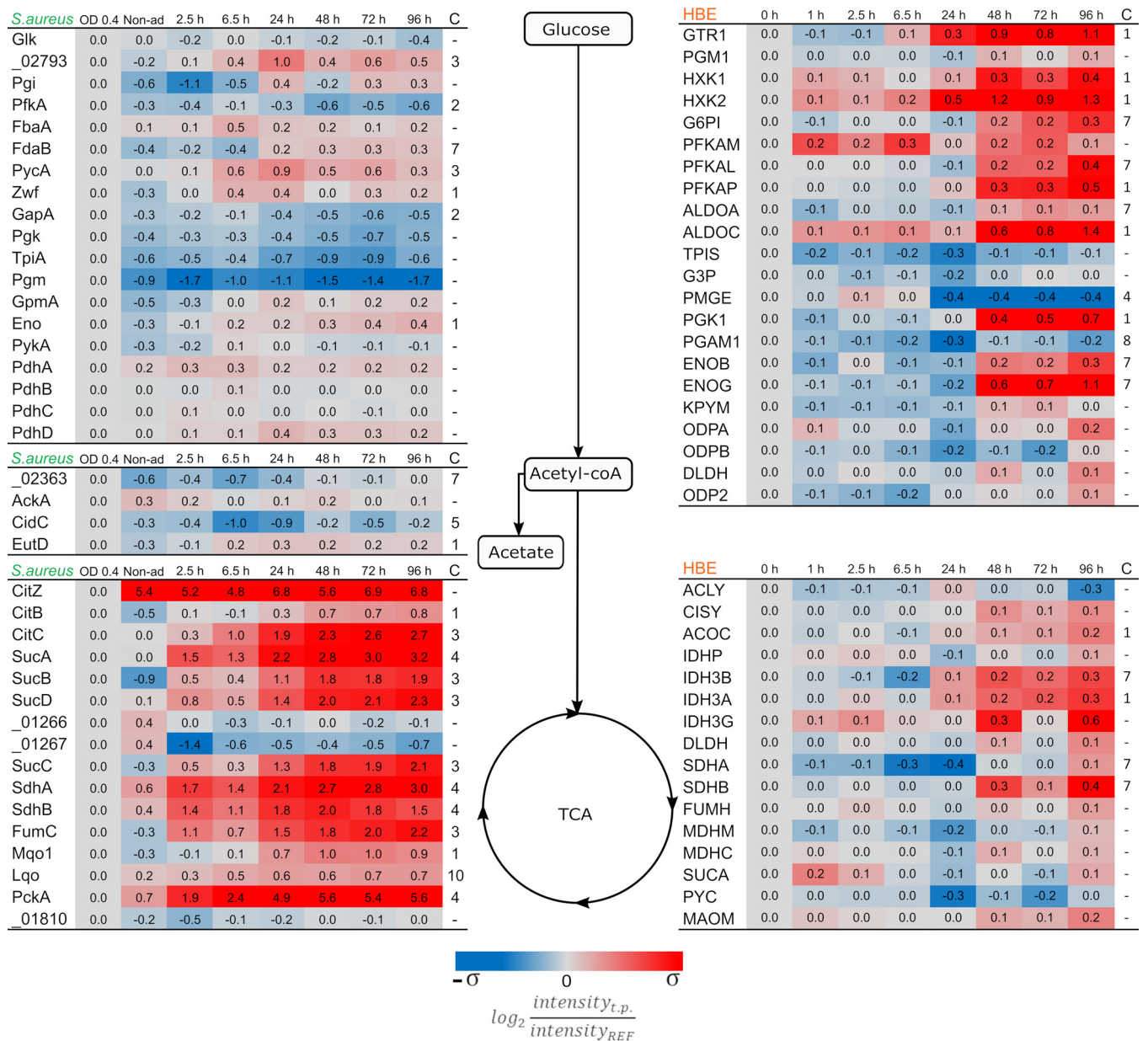


Fig. 2. Quantification of proteins related to the central carbon metabolism. Proteins were grouped depending on their main function in the different pathways, and boxes mark selected central metabolites. The proteome data from *S. aureus* are depicted on the left side of the diagram and the proteins without assigned gene symbol are labeled according to their locus tag with the “SAOUHSC_” identifier. The host data are shown on the right side of the diagram. Protein trends that deviate (p value < 0.01) from constant linearity over time were fitted to clusters according to their behavior (supplemental Fig. S2). Clustering assignment is shown in the last column “C.” The color coding is based on the standard deviation of each set of data. $\sigma_{S. aureus} = 2.34$ and $\sigma_{HBE} = 0.33$. The ratios were obtained from the measurements of four biological replicates.

sary for the glycolysis process (40). Our analysis shows that, after internalization, the bacteria repress some of the respective enzymes, but this regulation is reversed 48 h p.i., suggesting that the dormant bacteria restart the production of acetate. Proteins involved in the TCA cycle do not mirror this behavior, displaying increased levels after internalization and remaining at a constant increased level over the whole duration of the experiment. This strategy would supply the bacteria with enough

metabolic energy under conditions of low glucose availability. The strongest level increase takes place around 24 h p.i., which corresponds to the transition in subpopulation predominance from the replicating to the dormant phenotype of the bacteria. The two proteins of this pathway that do not display increased levels (SAOUHSC_01266, and SAOUHSC_01267) are 2-oxoglutarate oxidoreductases, which participate in the reverse TCA cycle and synthesize 2-ketoglutarate. In contrast to the bacteria,

most TCA cycle proteins from the human host cells show no changes in abundance, except for those involved in the metabolism of fumarate and α -ketoglutarate. Of note, these compounds are key intermediates in amino acid metabolism.

After the first 2.5 h p.i., the bacteria activate pathways related to energy acquisition from alternative sources, like fatty acids, as can be inferred from the elevated levels of *S. aureus* proteins involved in the catabolism of glycerol and fatty acids (Fig. 3). This increased level of fatty acid degrading enzymes becomes even stronger after the bacteria are predominantly in a dormant state, suggesting that the persistent bacteria prefer fatty acids over glycerol. This effect is not observed in the host proteome, where the levels of most proteins related to fatty acid degradation do not show major modifications.

Oxygen is a key factor limiting bacterial metabolism during infection because of the microaerobic environment they are exposed to after internalization (13, 17). This is clearly reflected by the mild increase in *S. aureus* proteins related to fermentation after internalization (Fig. 3), but the major rise takes place after 24 h p.i. On the other hand, most proteins related to the electron transport chain present a peak of upregulation at 24 h p.i., after which their abundance decreases. Conversely, the host proteome does not show major changes regarding the possible competition for oxygen. Only four proteins related to fermentation and oxidative phosphorylation present some regulation (Fig. 3). In addition, the proteins Elongin-B (ELOB), Elongin-C (ELOC), and Cullin-5 (CUL2), which are involved in the degradation of the Hypoxia-Inducible Factor (HIF α) remain constant over time, indicating that there is no increase in the abundance of this factor and suggesting that the host cells do not perceive a reduction of the available oxygen.

Amino acids are major alternative sources for carbon and nitrogen, which may play a role in adaptations of a pathogen to the intracellular milieu. Arginine, asparagine, and tryptophan have gained interest because of their relevance in the host's and bacterial defense mechanisms (41, 42). The pathways for biosynthesis and degradation of amino acids are complex because they are interconnected. Nevertheless, the proteomics data purport that *S. aureus* employs most pathways needed for amino acid metabolism during the entire course of infection. Conversely, the host proteome is geared toward utilization of amino acids that are primarily linked to the TCA cycle (Fig. 4).

Bacterial enzymes linked to the degradation of histidine, serine, cysteine and threonine show increased levels right from the moment of internalization, even preceding the start of intracellular replication. These proteins are related directly with the generation of precursors of pyruvate and the one-carbon metabolism. Nevertheless, because there is no replication of bacteria after 48 h p.i., the degradation of histidine, serine, cysteine, and threonine is most likely used to access an alternative carbon source. On the contrary, the infected

host cells induce the GLYM protein at later stages of infection, which is directly related to the conversion of serine into glycine and 5,10-methylenetetrahydrofolate. The latter molecule is an indispensable building block for purine biosynthesis (43, 44). This finding is fully consistent with the continuing growth of the human cell layer once the internalized pathogen predominantly presents a state of dormancy after 48 h p.i.

Branched-chain amino acids (BCAAs; Valine, Leucine, Isoleucine) are essential for human cells and, thus, need to be ingested from the environment. Consequently, the human proteins handling these amino acids are all involved in their degradation. Only at the last time point p.i. the human proteins addressing BCAAs have slightly increased levels (Fig. 4). On the other hand, the BCAAs play a crucial role in the regulation of *S. aureus* metabolism, because they serve as cofactors for CodY. This regulator also represses the production of the IlvA1, IlvE, and IlvC proteins, which have all increased abundance to participate in the biosynthesis of BCAAs (45). In parallel, the observed upregulation of the Ald1 and Ald2 proteins involved in the degradation of alanine implies that also this amino acid is used as an alternative carbon source during infection.

The amino acid catabolic pathways related to the TCA cycle are mostly needed to generate the intermediate molecules α -ketoglutarate, oxaloacetate and fumarate. *S. aureus* increases levels of proteins from the urea cycle (RocADF; ArcAC) to degrade arginine and to produce α -ketoglutarate from glutamate (GudB). Therefore, these substrates are used to feed the TCA cycle. Of note, the degradation of arginine is carried out by two pathways. First, the arginase pathway will produce glutamate (46) and presents upregulation right from the beginning of the infection (RocADF). Similarly, the arginine deaminase pathway (ArcAC), converting arginine to ornithine (47), expresses higher levels from the beginning of the infection but its highest quantities are found in the dormant bacterial population. Further, upregulated proteins associated with aspartate metabolism are dedicated to the production of lysine, as was also observed by Michalik *et al.* and Surmann *et al.* (13, 24). In contrast to the internalized bacteria, the amino acid metabolism in the infected host cells is geared toward the biosynthesis of amino acids for protein production. Specifically, the glutamate and aspartate metabolic pathways are upregulated in relation to the formation of collagen (PLOC2, PLOC3, P4HA1, P4HA2), and the synthesis of asparagine (ASNS), L-proline (OAT, P5CR1) and spermidine (SPEE). With exception of ASNS and SPEE, all these regulatory events start to take place after 48 h p.i. with a constant increment.

Staphylococcus aureus Escapes Degradative Compartments, Leading to a Persistent Population Residing in the Cytosol—Because *S. aureus* can reside in different intracellular niches, we determined the subcellular location of the bacteria over time through confocal immunofluorescence microscopy (Fig. 5A; supplemental Fig. S3). Inspection of the colocalization of the GFP-positive bacteria with LAMP-1 and

Cross-talk of Bronchial Cells and *S. aureus* During Infection

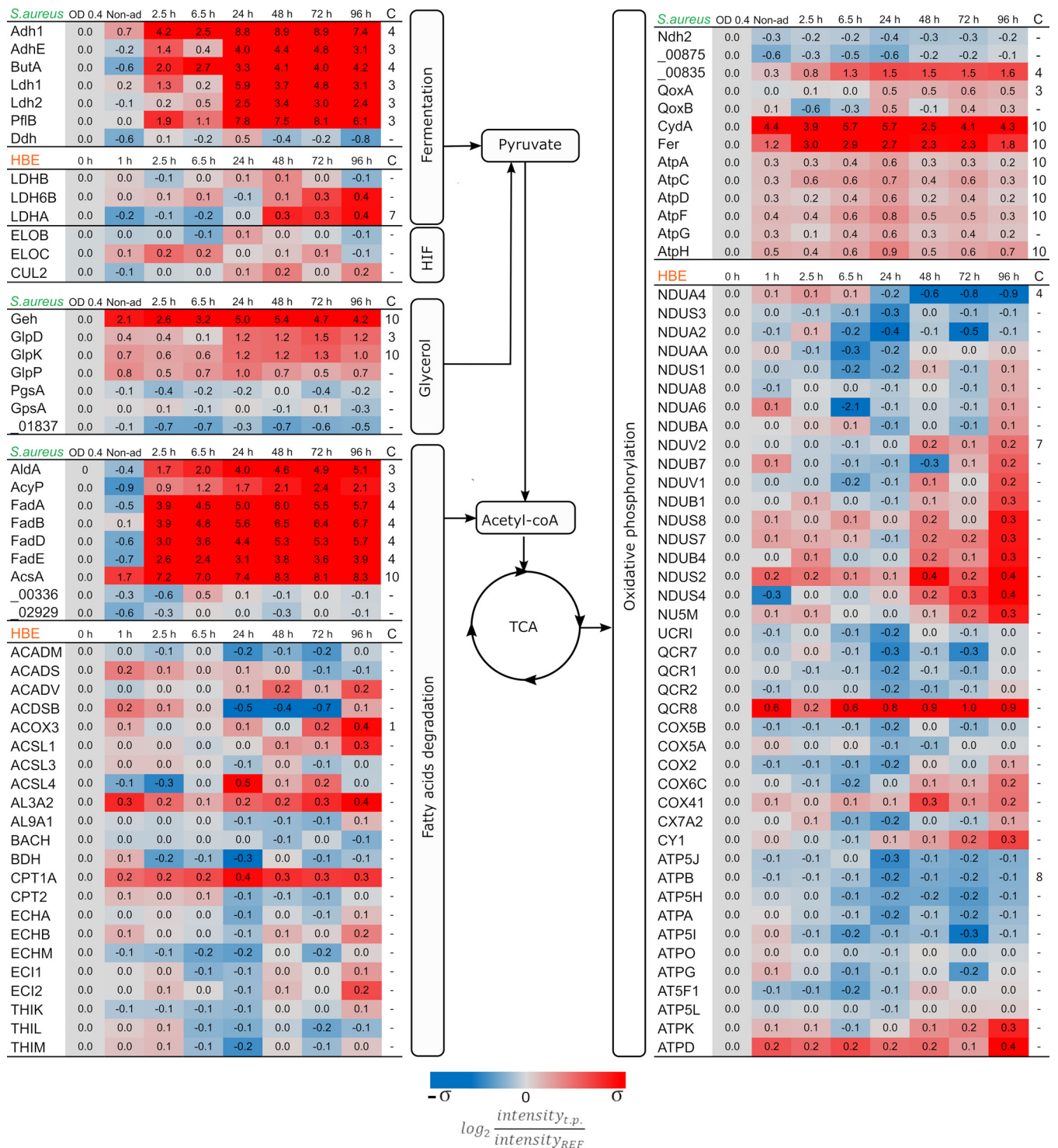
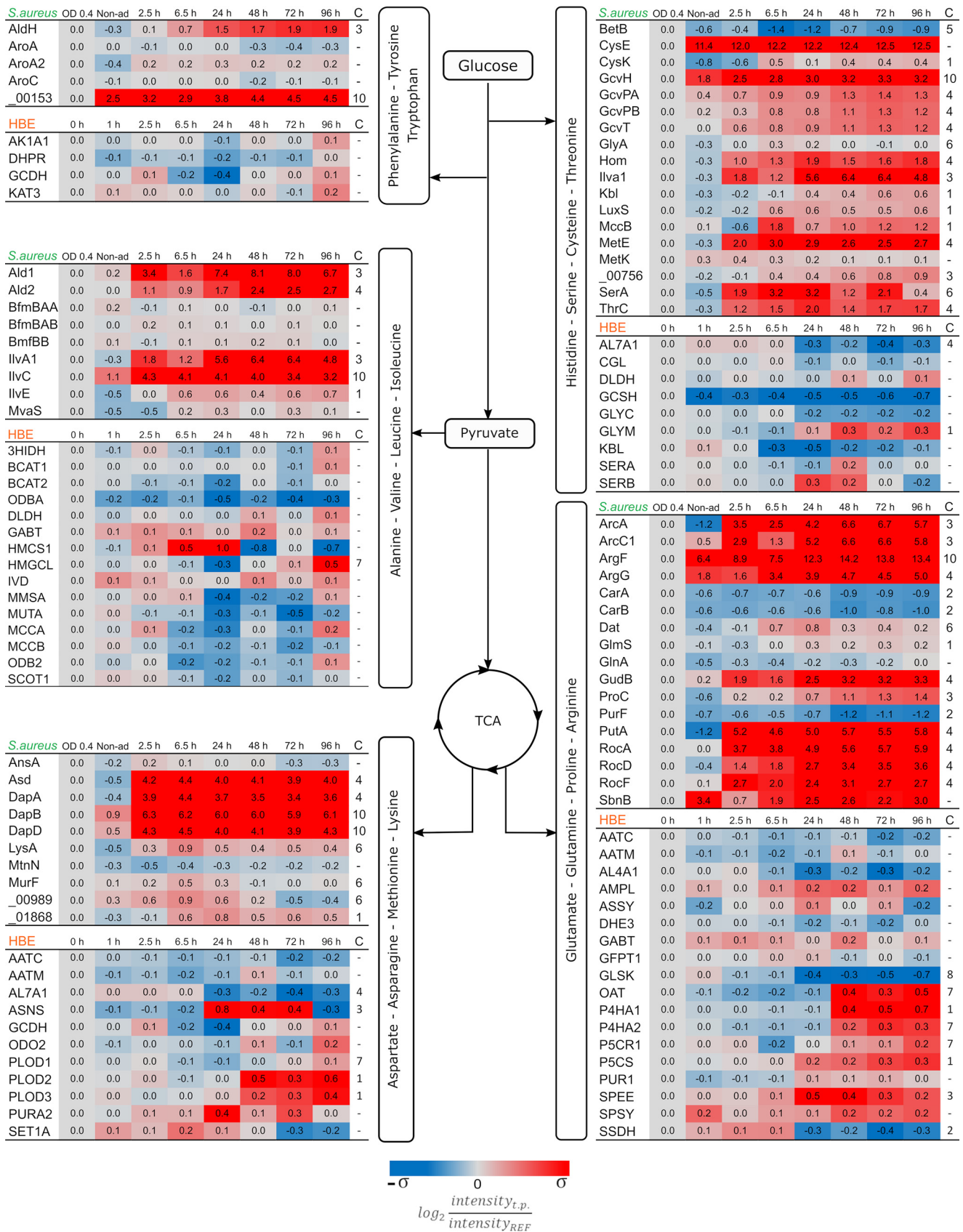


FIG. 3. Proteins related to alternative carbon sources and respiration. Assignment of proteins to pathogen (*S. aureus*) and host (HBE) is indicated on top of the respective protein groups. Because of limited carbon sources post infection, alternative pathways for energy consumption are upregulated. Substrates like glycerol and fatty acids are being consumed during intracellular conditions. Another major adjustment is related to oxygen availability which affects the pathways related to fermentation and oxidative phosphorylation. The protein quantities were derived from the mass spectrometry measurements of four biological replicates. Time trends with significant changes (p value < 0.01) were fitted into different clusters depending on their behavior (supplemental Fig. S2). The assigned cluster is presented in the column “C.” The color coding is based on the standard deviation of each set of data. $\sigma_{S. aureus} = 2.34$ and $\sigma_{HBE} = 0.33$. HIF: Hypoxia-Inducible Factor.

Cross-talk of Bronchial Cells and *S. aureus* During Infection



LC3 revealed that, during the first 24 h p.i., most of the bacteria are located inside membranes. Specifically, the replicating population resides inside LAMP-1- or LC3-associated compartments, whereas individual bacteria or small bacterial clusters are detectable in the cytosol from 2.5 h p.i. with increasing frequency over time. Ultimately, at the last time points, most observed bacteria do not colocalize with the membrane markers, suggesting a cytosolic localization of the persistent subpopulation. To substantiate these findings, we performed transmission electron microscopy (Fig. 5B; supplemental Fig. S4), highlighting that the replicating bacteria localize inside so-called “light phagosomal compartments” or “dark degradative compartments,” which are possibly lysosomes, during the first 48 h p.i. Further, at 72 h and 96 h p.i. all bacteria are no longer enclosed by a membrane, ultimately showing that the persister population resides in the cytosol. Of note, all bacterial clusters found inside phagocytic membranes are in dying host cells (Fig. 5B, supplemental Fig. S4).

Considering that the escape of *S. aureus* from the vesicles relies on toxins (48), we inspected the possible presence of staphylococcal virulence factors. Consistent with the observed changes in the subcellular localization of bacteria, the toxins HlgB, HlgC, Hla, LukH, and LukG are upregulated after 24 h p.i. (Fig. 5C; supplemental Table S8). These pore-forming proteins have been correlated with escape and subsequent host cell lysis (48). Of note, all these toxins are regulated by the Agr and/or SaeSR systems that positively or negatively regulate many more proteins involved in pathogenesis (49). Accordingly, the surface-bound protein Spa, which is negatively regulated by Agr, is present in decreased amounts after internalization with a minimum at 6.5 h p.i. A similar but less pronounced trend is observed for the surface-bound proteins ClfA and ClfB, the expression of which relies also on other regulators, such as the sigma factor B (SigB) in case of ClfA.

As the physicochemical conditions within different subcellular compartments differ substantially, bacteria within these compartments need to respond appropriately to different stresses. This is particularly true for lysosomes and phagosomes, which represent acidic compartments. In addition, the phagosomes are known to produce massive amounts of reactive oxygen species (ROS). Such conditions have a strong propensity to damage bacterial macromolecules, such as DNA and proteins. Indeed, the levels of many bacterial proteins involved in the prevention of oxidative damage are altered over time. These include KatA, SodA, SodM, AhpC, GapB, Hmp, and Dps (Fig. 5D; supplemental Table S8). In particular, the latter three proteins present a strong upregulation. GapB increases the production of NADPH, which is

needed to keep antioxidant proteins in a reduced state to allow them to reduce ROS and keep cytoplasmic proteins in a reduced state. The nitric oxide dioxygenase Hmp is involved in NO detoxification (50), whereas the ferroxidase Dps prevents DNA damage by binding iron and thus inhibiting hydroxyl radical formation (51). Of note, the upregulation of these proteins is triggered at 24 h p.i., a time point at which the growing bacteria are escaping from the cells (Fig. 1A, 1B) and the shift toward dormant/latent cells is taking place. Interestingly, the levels of these proteins remain high at later time points p.i. Similarly, proteins involved in the response to DNA damage, XseA and AddA, are upregulated after internalization (Fig. 5E), whereas other DNA damage-inducible proteins such as UvrB and UvrB2 are upregulated after 24 h. These observations imply that the bacteria are probably exposed to ROS and reactive nitrogen species (RNS) until 96 h p.i. and, remarkably, that this upregulation is not specifically related to the compartment in which they reside.

Intracellular S. aureus Persists Induce a Nonapoptotic Reaction of the Host—During the live cell imaging, it was observed that host cells lyse as a consequence of the replication of intracellular bacteria. This raised the question whether any indicators related to the mode of cell death could be identified. Indeed, the host cell proteome showed strong regulation of particular proteins implicated in cell death (Fig. 6). Of note, most regulators involved in the so-called apoptotic and pyroptotic pathways are kinases, which require activation to perform their functions, therefore their abundance does not correlate with host responses. Consequently, we investigated the cell death mechanisms by looking at other modulators of these pathways not requiring activation. These include the BAX, BAD, DBLOH, UACA, APAF, and ACS proteins that promote the initiation of apoptosis by stimulating caspase production. However, most of these proteins present no regulation, likely leading to the observed mild abundance changes of pro-apoptotic caspases CASP3, CASP7, and CASP8. However, CASP3 and CASP8 still display moderately increased abundance at early time points until 24 h p.i., hinting at potential apoptotic events occurring during bacterial replication. CASP6, on the other hand, is upregulated during the entire course of infection. Another pathway that can trigger apoptosis of the host cells depends on activation of P53. Its respective activators (BAG6, DDX5) are upregulated at early time points p.i., but their abundance decreases at 48 h p.i. On the other hand, anti-apoptotic proteins, such as TFIP8, 2CL1, NOC2L and BNIP2, display upregulation after infection (Fig. 6). Consistent with the latter observation, the Calpain 1 and 2 proteins, which are regulated in response to apoptosis,

FIG. 4. Progression of the pathways related to amino acid metabolism. Assignment of proteins to pathogen (*S. aureus*) and host (HBE) is indicated on top of the respective protein groups. The catabolism of amino acids after infection would provide resources to compensate for the lack of carbon and nitrogen sources. Moreover, the anabolism of these molecules is likewise required for some defensive functions of either the bacteria or the host. The represented ratios are the average of four biological replicates. Proteins with significant changes (p value <0.01) were clustered in groups depending on their general behavior and their assigned cluster is showed in column “C.” The color coding is based on the standard deviation of each set of data. $\sigma_{S. aureus} = 2.34$ and $\sigma_{HBE} = 0.33$.

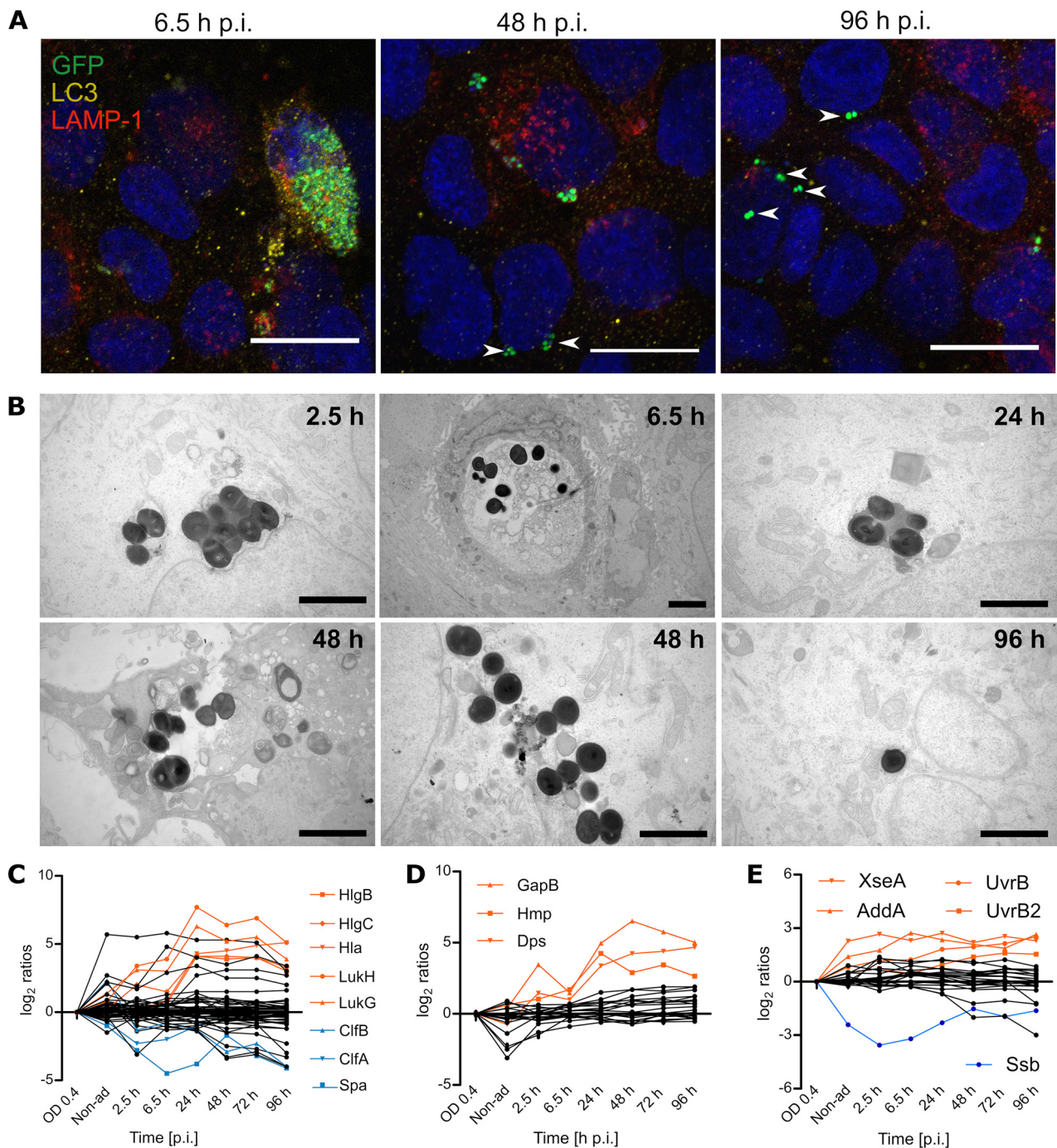


FIG. 5. The persistent subpopulation of *S. aureus* is mostly found in the cytosolic environment of the host. Intracellular localization of *S. aureus* was examined by colocalization with the protein markers LC3 for phagosomes, and LAMP-1 for lysosomes (A; scale bar: 20 μ m). After internalization of *S. aureus* by the HBE cells, most of the bacterial population is located inside closed compartments. Still, the percentage of bacteria that escapes the vesicles (white arrow heads) increases over time and most of the bacteria are found in a cytosolic environment by the end of the infection. These results were corroborated by electron microscopy (B; scale bar = 2 μ m). During the first hours of infection most bacteria are located inside degradative vesicles (dark compartments) or phagosomes (light compartments), but single bacteria escaped the closed compartments as early as 2.5 h p.i. (supplemental Fig. S4). By the end of the time of observation, at 72 h and 96 h p.i., all observed bacterial clusters are cytosolic. The displayed images are representative of different time points, additional images are provided in supplemental Figs. S3 and S4. The escape from the compartments could be induced by proteins related to pathogenesis, including toxins regulated by Agr and SaeRS (C). Moreover, this escape might have an impact on the production of stress proteins related to oxidative stress (D) and DNA repair (E). The list of proteins included in the line plots is available in the supplemental Table S8. A selection of significantly (p value < 0.01) regulated proteins is displayed in orange and blue. The mass spectrometry data represents the average of four biological replicates.

Cross-talk of Bronchial Cells and *S. aureus* During Infection

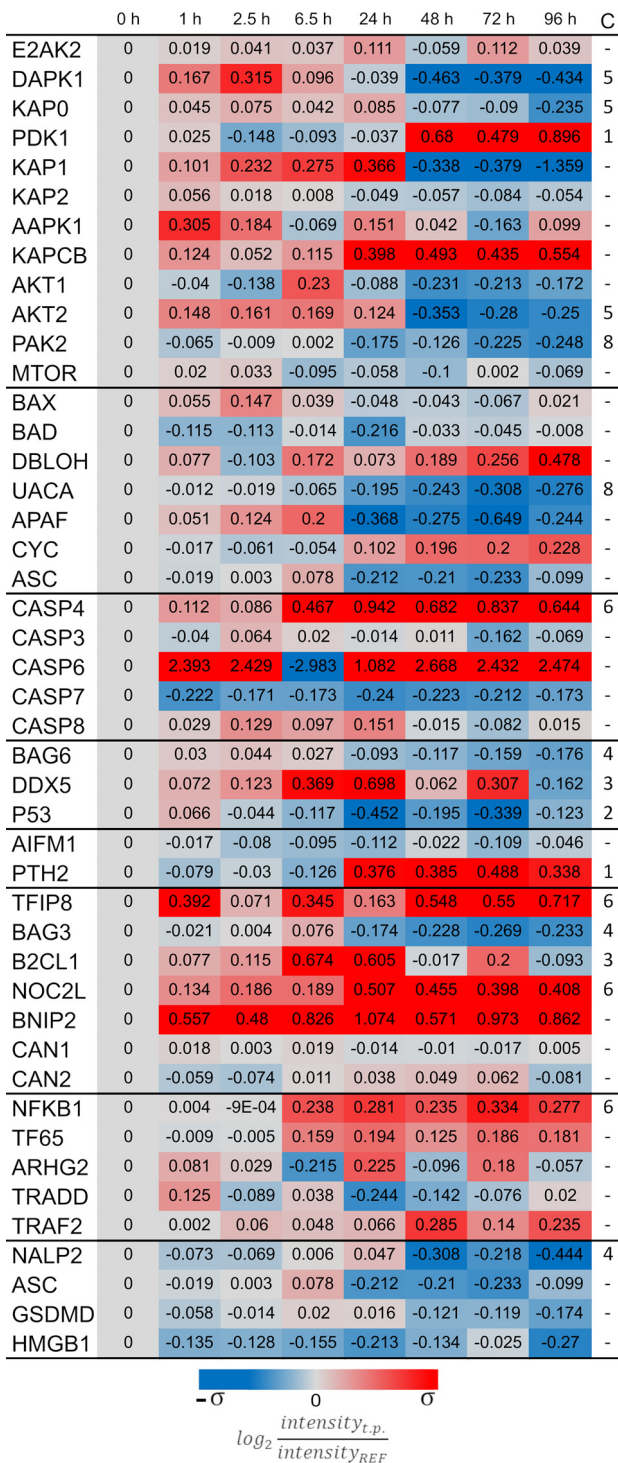


FIG. 6. Death of the host cells is not caused by induction of the classical apoptotic pathway. Mass spectrometry data of human proteins associated with different apoptosis-related pathways are presented in the figure. From top to bottom the presented groups of proteins include: kinases, regulators of caspases, the caspases, activators of P53, proteins related to caspase-independent activation of apoptosis, anti-apoptotic proteins, proinflammatory and pro-NF- κ B, and lastly pyroptosis. The color coding is based on the standard deviation $\sigma_{HBE} = 0.3$.

show no altered levels over the entire time course of infection. Lastly, it is noteworthy that CASP4 is the only caspase that shows consistent upregulation post internalization. This protein is related with the activation of the inflammasome or pyroptosis. Another protein involved in this pathway is NF- κ B, whose upregulation appears synchronized with CASP4. Nonetheless, so-called pro-NF- κ B proteins like TF65, ARHG2, TRADD and TRAF2 show no significant regulation. Lastly, proteins related to the activation of the inflammasome on the canonical or non-canonical pathways (52) were present at lower levels at the end of the infection, starting at 48 h p.i., indicating that these pathways are not activated. Altogether, these observations are consistent with apoptotic events that occur exclusively during the early stages of infection, when the bacteria still replicate.

DISCUSSION

During the infectious process, *S. aureus* has the option to evade the human immune defenses, or to invade nonprofessional phagocytic cells to hide from the host immune system and to evade antibiotic therapy. After internalization, the pathogen needs to adapt to the intracellular conditions so that it can survive, replicate, eventually leave the host cell and spread to other tissues. To optimize its fitness, the internalized pathogen displays population heterogeneity, where a fraction of the internalized bacterial population starts to replicate while another fraction displays low growth rates and reaches a state of dormancy (53). These two populations reflect the two main objectives of the internalized bacteria, where the replicating bacteria will ultimately disrupt epithelial cells to invade and infect the underlying tissue, whereas the dormant bacteria will persist intracellularly for an extended period.

Right from the moment the bacteria are internalized by a host cell, both host and pathogen need to adapt to the new situation, and then they will start to compete for resources. Importantly, such adaptations at the bacterial end are not limited to the production of virulence factors or mounting of defense mechanisms, but they also involve an optimized management of resources and the sensing of changes in the intracellular environment because of host cell adaptations (41). A clear example of this is the observed downregulation of bacterial ribosomal proteins, which relates to the formation of the alarmone ppGpp because of the induction of the stringent response after nutrient starvation and exposure to various stresses. The ppGpp molecule is synthesized from GTP, and the resulting decrease in GTP is sensed by the CodY regulatory system of *S. aureus* (50). Of note, CodY is a major modulator of *S. aureus* central carbon and amino acid metabolism and it also influences staphylococcal virulence (54). Consistent with the idea that the observed downregulation of ribosomal proteins is a consequence of ppGpp production, we observed an upregulation of the CodY-regulated proteins PycA, AcsA, ButA, Hom, metE, SerA, ThrC, Asd, DapABD,

and ArgG. In this context, the activation of the *ilv-leu* operon is particularly noteworthy as this will result in the synthesis of BCAAs, which are cofactors of CodY (50). Of note, the inactivation of CodY may also result from reduced amino acid supply, and induction of the biosynthetic operons will counteract this shortage. However, high levels of BCAAs would then lead to repression again, but it remains unclear whether intracellularly such high BCAA levels are reached.

Importantly, the biological pathways in the internalized bacteria and their host cells are not isolated from each other despite their physical separation by the bacterial cell envelope. Accordingly, changes in pathway regulation at either end will impact the whole biological system. Thus, host cell homeostasis will change in response to the presence of internalized bacteria. This is exemplified by the observed optimization of glucose uptake and catabolism in the host cells from 48 h p.i. onwards, once they predominantly carry dormant bacteria. Interestingly, this correlates with the observed shift of the bacterial intracellular localization from an encapsulated state in vesicles to a free state in the cytosol. Therefore, this relocalization event could very well be a trigger for direct competition for nutrients that has a significant impact on the host cells' energy metabolism. Similarly, under nutrient-deprived conditions, the regulation of metabolic processes plays a crucial role in the adaptation of *S. aureus* to the altered conditions intracellularly. This is underpinned by the observed upregulation of proteins that are crucial for survival of the bacteria, such as PckA, AckA, FumC, SdhA, SucC, SucA and GudB, or proteins that promote their proliferation, such as Ald1, Ald2, Pyc, AspA, Mqo1, GltA, AcnA, Icd, RocA, RocD, and PutA. In fact, this observation is reminiscent of an *in vitro* study by Halsey *et al.* (55), who reported that under glucose restriction several of the afore-mentioned proteins have an impact on the fitness of the pathogen.

A fermentative phenotype combined with inactivation of the electron transport chain has been associated with the development of SCVs that are often observed on intracellular persistence of *S. aureus* (40). Although we did not observe the emergence of such SCVs by plating intracellular bacteria during our experimental setup, the observed changes in the bacterial proteome from 24 h p.i. onwards are indicative of a metabolic shift toward fermentation. Some of the proteins involved in bacterial electron transport become less abundant at late stages p.i. compared with their levels during replication, whereas major proteins involved in the fermentation of pyruvate and acetate catabolism are constitutively upregulated throughout till the end of the infection. We observe a shift from terminal oxidases that require high oxygen concentrations to those that need lower concentrations, but do not pump H⁺ anymore, which is consistent with microaerobic intracellular conditions. Further, it is known that the formation of SCVs is linked to SigB, which is an important factor for intracellular persistence (11). Indeed, a clear role for SigB in reaching the dormancy state is indicated by changes in the abundance of

58 predominantly SigB-dependent proteins during the late stages of infection, among them Asp23, CidC, ClpL, FdhA, GapA, OpuBA, and SpoVG (supplemental Table S6). In this respect, it is noteworthy that *S. aureus* strains defective in SigB are unable to catabolize acetate (56).

Besides the changes in the central carbon metabolism of the host, the host's amino acid metabolism is also altered, altering the outcome of infection. Contrary to bacteria, which degrade amino acids mostly for the acquisition of carbon and nitrogen, the TCA-related amino acid production by the host cells is not related to an optimization of energy production but rather to the production of other molecules like purines, proline and collagen. However, some of the regulated proteins may also affect other pathways. For example, the observed asparagine synthase (ASNS) upregulation could be related to "infection stress." This idea would be consistent with a previous study on group A streptococcal (GAS) infection, where the pathogen was shown to induce pores in the host membranes, leading to Ca⁺² influx into the cytoplasm. In turn, this led to an upregulated synthesis of asparagine, which was then utilized by the pathogen (57). In our experimental setup, the ASNS abundance peaks at 24 h p.i., which correlates with the highest production of pore-forming toxins and escape of growing bacteria from the host cells, suggesting a similar regulatory mechanism as proposed for GAS.

Importantly, we observed that *S. aureus* induces its two pathways for arginine degradation right after internalization. It has been reported that arginine depletion by the pathogen induces death of the host cell (42). Moreover, competition for arginine with the host protein iNOS reduces the production of NO, which serves in the host's antibacterial defense. Lastly, the deaminase pathway produces NH₃ and ATP, which supports the intracellular survival of the bacteria. In particular, the ammonia produced increases the pH of the intracellular environment thereby preventing fusion of the endosome with the lysosome (41, 42), whereas the production of ATP generates a source of energy during hypoxic conditions or when the electron transport system is deficient (58). The degradation of arginine usually occurs in environments with a high proline concentration (46). Interestingly, proteins related to production of proline are upregulated in the host proteome after 48 h p.i., when the bacterial machinery for arginine catabolism becomes massively upregulated. The host cell death induced by arginine starvation is regulated by activation of AMPK, which suppresses the master regulator mTOR and could lead to autophagy (42). In this context, it is noteworthy that spermidine is implicated in autophagy as well (59), and the SPEE protein, responsible for spermidine synthesis, is part of the same host pathway that also leads to the synthesis of glutamate, glutamine, proline, and arginine. As for ASNS, the abundance of SPEE peaks at 24 h p.i., which correlates well with the localization of the bacteria in phagocytic membranes as observed by fluorescence microscopy and TEM. These results underpin a connection between host amino acid starva-

tion and autophagy, as suggested by a recent study (60). Although the roles of arginine and asparagine metabolism in infectious processes have been studied previously, the presently observed regulation of amino acid metabolism focuses additional attention on proline, glutamate, and alanine as potential modulators of infection.

Membrane-enclosed bacteria can be observed within the host cells up until 48 h p.i. as evidenced by TEM. This time-span covers the bacterial replication phase in the infectious process (*i.e.* until 24 h p.i.), but also nonreplicating bacteria can escape from lysosomes and phagosomes which is most clearly evident from 48 h p.i. onwards. Consequently, once the bacteria have entered the dormant state, they will persist intracellularly in the cytosol. These findings contrast with what has been found in other models that investigated *S. aureus* infection of other human cell types, where the cytosolic bacteria proliferated after phagosomal escape (61).

The bacterial escape from vesicles is most likely because of the production of pore-forming toxins that are regulated by SaeRS and Agr (50). The latter regulator is itself regulated by CodY, which also regulates most of the carbon resource management as described above (62). Nevertheless, some bacteria remain in vesicles until the very end of the present experiment. Together, these observations suggest that the bacteria-containing vesicles are possibly not fully functional. This intriguing idea is supported by the observed bacterial behavior, where oxidative stress and DNA-damage-induced responses are observed right from the beginning till the end of the experiment without a detectable peak. In fact, the abundance of some proteins related to these stress responses increased more strongly from 24 h p.i. onwards, when most bacteria have already been liberated from vesicular captivity. A possible vesicle malfunction could be because of the aforementioned production of pore-forming toxins. The latter idea is supported by the fact that the production of these toxins seems to be reduced at the last time points of infection where most intracellular bacteria are no longer enclosed in a membrane.

Lastly, our present proteome analyses reveal that apoptosis of the host cell could occur from 6.5 h until 24 h p.i., suggesting that the replicating bacteria employ this mechanism to escape from their host cells as captured by live cell imaging. In addition, we observed upregulation of Caspase 4 and NF- κ B indicating activation of the inflammasome from 6.5 h p.i. Nevertheless, the abundance profile of other proteins implicated in the so-called pyroptosis does not support the activation of this self-destructive pathway. This suggests that the bacterial presence, even in dormancy, somehow leads to suppression of pyroptosis, which would lead to host cell lysis. Whether this is directly related to the bacterial actions, or indirectly to the response of the host to the bacterial presence is unclear.

Taken altogether, the present proteomics dissection of interactions between human bronchial epithelial cells and internalized *S. aureus* highlights the dynamic adaptive changes in

the two interacting systems over an unprecedented period. This was necessary to obtain a proper understanding of the levels at which the two systems collide and eventually reach an equilibrium. Importantly, although in the past years host-pathogen interaction studies have focused attention on the roles of bacterial virulence factors and immune evasion, our present observations place dynamic host-pathogen interactions at the metabolic level in the limelight. Clearly, pore-forming toxins have a crucial role in giving the invading bacteria access to the resources that are hidden within the host cells, but it is the way in which these resources are used by the bacteria and how the host and the bacteria subsequently adapt to each other that determines the ultimate outcome of the infectious process.

Acknowledgments—We thank Jan Pané-Farré for providing the HG001 Δ spa, Jeroen Kuipers for his assistance during the electron microscopy imaging, Dr. Muriel Mari for the support in identifying the vesicles in the microscopy images, and Giorgio Gabarrini for his comments that greatly improved the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY

The identified peptide ions and their Q-values are included in [supplemental Tables S4 and S5](#). All protein annotations and median abundances are included in [supplemental Tables S6 and S7](#). The raw files from mass spectrometry have been deposited in MassIVE (<https://massive.ucsd.edu>) with the access codes MSV000082920 (*S. aureus*) and MSV000082923 (16HBE14o).

* Funding for this project was received from the Graduate School of Medical Sciences of the University of Groningen [to L.M.P.M., S.A.M., and J.M.v.D.], the Deutsche Forschungsgemeinschaft Grants GRK1870 [to L.M.P.M., S.A.M. and U.V.] and SFBTRR34 [to U.V.], and CEC MSCITN grant 713482 [ALERT, to H.Y., E.J.M.R., A.M.S., and J.M.v.D.]. Part of this work has been performed at the UMCG Imaging and Microscopy Center (UMIC), which is sponsored by NWO-grants 40-00506-98-9021 (TissueFaxes) and 175-010-2009-023 (Zeiss 2p).

§ This article contains [supplemental Figures and Tables](#). The authors declare that they have no financial and non-financial competing interests in relation to the documented research.

** To whom correspondence may be addressed: University of Groningen (UMCG), Department of Medical Microbiology, Hanzeplein 1, P.O. Box 30001, 9700 RB Groningen, Groningen, the Netherlands. E-mail: j.m.van.dijl01@umcg.nl.

‡‡ To whom correspondence may be addressed: Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Felix-Hausdorff-Str. 8, 17475 Greifswald, Mecklenburg-Vorpommern, Germany. E-mail: voelker@uni-greifswald.de.

Author contributions: L.M.P.M., A.S., L.K., J.M.v.D., and U.V. designed research; L.M.P.M., H.Y., E.J.R., P.H., M.G.S., K.S., H.P., and S.A.M. performed research; L.M.P.M., A.-K.B., S.M., L.K., J.M.v.D., and U.V. analyzed data; L.M.P.M., J.M.v.D., and U.V. wrote the paper.

REFERENCES

1. Wertheim, H. F., Melles, D. C., Vos, M. C., van Leeuwen, W., van Belkum, A., Verbrugh, H. A., and Nouwen, J. L. (2005) The role of nasal

- carriage in *Staphylococcus aureus* infections. *Lancet Infect. Dis.* **5**, 751–762
2. Lowy, F. D. (1998) *Staphylococcus aureus* Infections. *N. Engl. J. Med.* **339**, 520–532
 3. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L., and Fowler, V. G. (2015) *Staphylococcus aureus* Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management. *Clin. Microbiol. Rev.* **28**, 603–661
 4. Breathnach, A. S. (2013) Nosocomial infections and infection control. *Medicine* **41**, 649–653
 5. Appelbaum, P. C. (2006) MRSA—the tip of the iceberg. *Clin. Microbiol. Infect.* **12**, 3–10
 6. Fraunholz, M., and Sinha, B. (2012) Intracellular *Staphylococcus aureus*: Live-in and let die. *Front. Cell. Infect. Microbiol.* **2**, 43
 7. Garzoni, C., and Kelley, W. L. (2009) *Staphylococcus aureus*: new evidence for intracellular persistence. *Trends Microbiol.* **17**, 59–65
 8. Lehar, S. M., Pillow, T., Xu, M., Staben, L., Kajihara, K. K., Vandlen, R., DePalatis, L., Raab, H., Hazenbos, W. L., Hiroshi Morisaki, J., Kim, J., Park, S., Darwish, M., Lee, B.-C., Hernandez, H., Loyet, K. M., Lupardus, P., Fong, R., Yan, D., Chalouni, C., Luis, E., Khalif, Y., Plise, E., Cheong, J., Lyssikatos, J. P., Strandh, M., Koefoed, K., Andersen, P. S., Flygare, J. A., Wah Tan, M., Brown, E. J., and Mariathasan, S. (2015) Novel antibody–antibiotic conjugate eliminates intracellular *S. aureus*. *Nature* **527**, 323–328
 9. Proctor, R. A., van Langevelde, P., Kristjansson, M., Maslow, J. N., and Arbeit, R. D. (1995) Persistent and relapsing infections associated with small-colony variants of *Staphylococcus aureus*. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **20**, 95–102
 10. Seifert, H., Wisplinghoff, H., Schnabel, P., and von Eiff, C. (2003) Small colony variants of *Staphylococcus aureus* and pacemaker-related infection. *Emerg. Infect. Dis.* **9**, 1316–1318
 11. Tuchscherer, L., Bischoff, M., Lattar, S. M., Noto Llana, M., Pförtner, H., Niemann, S., Geraci, J., Van de Vyver, H., Fraunholz, M. J., Cheung, A. L., Herrmann, M., Völker, U., Sordelli, D. O., Peters, G., and Löffler, B. (2015) Sigma factor SigB is crucial to mediate *Staphylococcus aureus* adaptation during chronic infections. *PLoS Pathog.* **11**, e1004870
 12. Strobel, M., Pförtner, H., Tuchscherer, L., Völker, U., Schmidt, F., Kramko, N., Schnittler, H.-J., Fraunholz, M. J., Löffler, B., Peters, G., and Niemann, S. (2016) Post-invasion events after infection with *Staphylococcus aureus* are strongly dependent on both the host cell type and the infecting *S. aureus* strain. *Clin. Microbiol. Infect.* **22**, 799–809
 13. Surmann, K., Michalik, S., Hildebrandt, P., Gierok, P., Depke, M., Brinkmann, L., Bernhardt, J., Gesell Salazar, M., Sun, Z., Shteynberg, D., Kusebauch, U., Moritz, R. L., Wollscheid, B., Lalk, M., Völker, U., and Schmidt, F. (2014) Comparative proteome analysis reveals conserved and specific adaptation patterns of *Staphylococcus aureus* after internalization by different types of human non-professional phagocytic host cells. *Front. Microbiol.* **5**, 392
 14. Hecker, M., Mäder, U., and Völker, U. (2018) From the genome sequence via the proteome to cell physiology – Pathoproteomics and pathophysiology of *Staphylococcus aureus*. *Int. J. Med. Microbiol.* **308**, 545–557
 15. Kiedrowski, M. R., Paharik, A. E., Ackermann, L. W., Shelton, A. U., Singh, S. B., Starner, T. D., and Horswill, A. R. (2016) Development of an in vitro colonization model to investigate *Staphylococcus aureus* interactions with airway epithelia. *Cell. Microbiol.* **18**, 720–732
 16. Richter, E., Harms, M., Ventz, K., Nölker, R., Fraunholz, M. J., Mostertz, J., and Hochgräfe, F. (2016) Quantitative proteomics reveals the dynamics of protein phosphorylation in human bronchial epithelial cells during internalization, phagosomal escape, and intracellular replication of *Staphylococcus aureus*. *J. Proteome Res.* **15**, 4369–4386
 17. Surmann, K., Simon, M., Hildebrandt, P., Pförtner, H., Michalik, S., Stentzel, S., Steil, L., Dhople, V. M., Bernhardt, J., Schlüter, R., Depke, M., Gierok, P., Lalk, M., Bröker, B. M., Schmidt, F., and Völker, U. (2015) A proteomic perspective of the interplay of *Staphylococcus aureus* and human alveolar epithelial cells during infection. *J. Proteomics* **128**, 203–217
 18. Herbert, S., Ziebandt, A.-K., Ohlsen, K., Schäfer, T., Hecker, M., Albrecht, D., Novick, R., and Götz, F. (2010) Repair of global regulators in *Staphylococcus aureus* 8325 and comparative analysis with other clinical isolates. *Infect. Immun.* **78**, 2877–2889
 19. Liese, J., Rooijackers, S. H. M., van Strijp, J. A. G., Novick, R. P., and Dustin, M. L. (2013) Intravital two-photon microscopy of host–pathogen interactions in a mouse model of *Staphylococcus aureus* skin abscess formation. *Cell. Microbiol.* **15**, 891–909
 20. Cozens, A. L., Yezzi, M. J., Kunzelmann, K., Ohri, T., Chin, L., Eng, K., Finkbeiner, W. E., Widdicombe, J. H., and Gruenert, D. C. (1994) CFTR expression and chloride secretion in polarized immortal human bronchial epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **10**, 38–47
 21. R Core Team. (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria
 22. Pförtner, H., Wagner, J., Surmann, K., Hildebrandt, P., Ernst, S., Bernhardt, J., Schurmann, C., Gutjahr, M., Depke, M., Jehmlich, U., Dhople, V., Hammer, E., Steil, L., Völker, U., and Schmidt, F. (2013) A proteomics workflow for quantitative and time-resolved analysis of adaptation reactions of internalized bacteria. *Methods* **61**, 244–250
 23. Depke, M., Michalik, S., Rabe, A., Surmann, K., Brinkmann, L., Jehmlich, N., Bernhardt, J., Hecker, M., Wollscheid, B., Sun, Z., Moritz, R. L., Völker, U., and Schmidt, F. (2015) A peptide resource for the analysis of *Staphylococcus aureus* in host pathogen interaction studies. *Proteomics* **15**, 3648–3661
 24. Michalik, S., Depke, M., Murr, A., Gesell Salazar, M., Kusebauch, U., Sun, Z., Meyer, T. C., Surmann, K., Pförtner, H., Hildebrandt, P., Weiss, S., Palma Medina, L. M., Gutjahr, M., Hammer, E., Becher, D., Pribyl, T., Hammerschmidt, S., Deutsch, E. W., Bader, S. L., Hecker, M., Moritz, R. L., Mäder, U., Völker, U., and Schmidt, F. (2017) A global *Staphylococcus aureus* proteome resource applied to the in vivo characterization of host–pathogen interactions. *Sci. Rep.* **7**, 9718
 25. Zeitlin, P. L., Lu, L., Rhim, J., Cutting, G., Stetten, G., Kieffer, K. A., Craig, R., and Guggino, W. B. (1991) A cystic fibrosis bronchial epithelial cell line: immortalization by adeno-12-SV40 infection. *Am. J. Respir. Cell Mol. Biol.* **4**, 313–319
 26. Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O., and Reiter, L. (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410
 27. Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R. L. (2015) Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* **9**, 745–754
 28. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A., and Aebersold, R. (2010) A guided tour of the trans-proteomic pipeline. *Proteomics* **10**, 1150–1159
 29. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
 30. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics MCP* **10**
 31. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8**, 2405–2417
 32. Nagel, A., Michalik, S., Debarbouille, M., Hertlein, T., Gesell Salazar, M., Rath, H., Msadek, T., Ohlsen, K., Diji, J. M. van, Völker, U., and Mäder, U. (2018) Inhibition of Rho activity increases expression of SaeRS-dependent virulence factor genes in *Staphylococcus aureus*, showing a link between transcription termination, antibiotic action, and virulence. *mBio* **9**, e01332-18
 33. Fuchs, S., Mehlan, H., Bernhardt, J., Hennig, A., Michalik, S., Surmann, K., Pané-Farré, J., Giese, A., Weiss, S., Backert, L., Herbig, A., Nieselt, K., Hecker, M., Völker, U., and Mäder, U. (2018) AureoWiki - The repository of the *Staphylococcus aureus* research and annotation community. *Int. J. Med. Microbiol.* **308**, 558–568
 34. 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-seq and microarray studies. *Nucleic Acids Res.* **43**, e47
 35. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., and Smyth, G. K.

- (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.* **10**, 946–963
36. Futschik, M. E., and Carlisle, B. (2005) Noise-robust soft clustering of gene expression time-course data. *J. Bioinform. Comput. Biol.* **3**, 965–988
 37. Kumar, L., and Futschik, E. M. (2007) Mfuzz: A software package for soft clustering of microarray data. *Bioinformatics* **2**, 5–7
 38. Basu, A., and Yap, M.-N. F. (2016) Ribosome hibernation factor promotes *Staphylococcal* survival and differentially represses translation. *Nucleic Acids Res.* **44**, 4881–4893
 39. Ueta, M., Wada, C., and Wada, A. (2010) Formation of 100S ribosomes in *Staphylococcus aureus* by the hibernation promoting factor homolog SaHPF. *Genes Cells* **15**, 43–58
 40. Somerville, G. A., and Proctor, R. A. (2009) At the crossroads of bacterial metabolism and virulence factor synthesis in *Staphylococci*. *Microbiol. Mol. Biol. Rev.* **73**, 233–248
 41. Ren, W., Rajendran, R., Zhao, Y., Tan, B., Wu, G., Bazer, F. W., Zhu, G., Peng, Y., Huang, X., Deng, J., and Yin, Y. (2018) Amino acids as mediators of metabolic cross talk between host and pathogen. *Front. Immunol.* **9**, 319
 42. Xiong, L., Teng, J. L. L., Botelho, M. G., Lo, R. C., Lau, S. K. P., and Woo, P. C. Y. (2016) Arginine metabolism in bacterial pathogenesis and cancer therapy. *Int. J. Mol. Sci.* **17**, 363
 43. Giardina, G., Brunotti, P., Fiascarelli, A., Cicalini, A., Costa, M. G. S., Buckle, A. M., di Salvo, M. L., Giorgi, A., Marani, M., Paone, A., Rinaldo, S., Paiardini, A., Contestabile, R., and Cutruzzola, F. (2015) How pyridoxal 5'-phosphate differentially regulates human cytosolic and mitochondrial serine hydroxymethyltransferase oligomeric state. *FEBS J.* **282**, 1225–1241
 44. Morscher, R. J., Ducker, G. S., Li, S. H.-J., Mayer, J. A., Gitai, Z., Sperl, W., and Rabinowitz, J. D. (2018) Mitochondrial translation requires folate-dependent tRNA methylation. *Nature* **554**, 128–132
 45. Kaiser, J. C., King, A. N., Grigg, J. C., Sheldon, J. R., Edgell, D. R., Murphy, M. E. P., Brinsmade, S. R., and Heinrichs, D. E. (2018) Repression of branched-chain amino acid synthesis in *Staphylococcus aureus* is mediated by isoleucine via CodY, and by a leucine-rich attenuator peptide. *PLOS Genet.* **14**, e1007159
 46. Gardan, R., Rapoport, G., and Débarbouillé, M. (1997) Role of the transcriptional activator RocR in the arginine-degradation pathway of *Bacillus subtilis*. *Mol. Microbiol.* **24**, 825–837
 47. Ryan, S., Begley, M., Gahan, C. G. M., and Hill, C. (2009) Molecular characterization of the arginine deiminase system in *Listeria monocytogenes*: regulation and role in acid tolerance. *Environ. Microbiol.* **11**, 432–445
 48. Jarry, T. M., Memmi, G., and Cheung, A. L. (2008) The expression of alpha-haemolysin is required for *Staphylococcus aureus* phagosomal escape after internalization in CFT-1 cells. *Cell. Microbiol.* **10**, 1801–1814
 49. Geiger, T., Goerke, C., Mainiero, M., Kraus, D., and Wolz, C. (2008) The virulence regulator Sae of *Staphylococcus aureus*: promoter activities and response to phagocytosis-related signals. *J. Bacteriol.* **190**, 3419–3428
 50. Horn, J., Stelzner, K., Rudel, T., and Fraunholz, M. (2018) Inside job: *Staphylococcus aureus* host-pathogen interactions. *Int. J. Med. Microbiol.* **308**, 607–624
 51. Oogai, Y., Kawada-Matsuo, M., and Komatsuzawa, H. (2016) *Staphylococcus aureus* SrrAB affects susceptibility to hydrogen peroxide and co-existence with *Streptococcus sanguinis*. *PLOS ONE* **11**, e0159768
 52. Liu, X., and Lieberman, J. (2017) in *Advances in Immunology*, ed Alt FW (Academic Press), pp 81–117
 53. Fisher, R. A., Gollan, B., and Helaine, S. (2017) Persistent bacterial infections and persister cells. *Nat. Rev. Microbiol.* **15**, 453–464
 54. Pohl, K., Francois, P., Stenz, L., Schlink, F., Geiger, T., Herbert, S., Goerke, C., Schrenzel, J., and Wolz, C. (2009) CodY in *Staphylococcus aureus*: a regulatory link between metabolism and virulence gene expression. *J. Bacteriol.* **191**, 2953–2963
 55. Halsey, C. R., Lei, S., Wax, J. K., Lehman, M. K., Nuxoll, A. S., Steinke, L., Sadykov, M., Powers, R., and Fey, P. D. (2017) Amino acid catabolism in *Staphylococcus aureus* and the function of carbon catabolite repression. *mBio* **8**, e01434-16
 56. Somerville, G. A., Saïd-Salim, B., Wickman, J. M., Raffel, S. J., Kreiswirth, B. N., and Musser, J. M. (2003) Correlation of acetate catabolism and growth yield in *Staphylococcus aureus*: implications for host-pathogen interactions. *Infect. Immun.* **71**, 4724–4732
 57. Baruch, M., Belotserkovsky, I., Hertzog, B. B., Ravins, M., Dov, E., Mclver, K. S., Le Breton, Y. S., Zhou, Y., Youting, C. C., and Hanski, E. (2014) An extracellular bacterial pathogen modulates host metabolism to regulate its own sensing and proliferation. *Cell* **156**, 97–108
 58. Makhlin, J., Kofman, T., Borovok, I., Kohler, C., Engelmann, S., Cohen, G., and Aharonowitz, Y. (2007) *Staphylococcus aureus* ArcR controls expression of the arginine deiminase operon. *J. Bacteriol.* **189**, 5976–5986
 59. Minois, N. (2014) Molecular basis of the “anti-aging” effect of spermidine and other natural polyamines - a mini-review. *Gerontology* **60**, 319–326
 60. Bravo-Santano, N., Ellis, J. K., Mateos, L. M., Calle, Y., Keun, H. C., Behrends, V., and Letek, M. (2018) Intracellular *Staphylococcus aureus* modulates host central carbon metabolism to activate autophagy. *mSphere* **3**, e00374-18
 61. Grosz, M., Kolter, J., Paprotka, K., Winkler, A.-C., Schäfer, D., Chatterjee, S. S., Geiger, T., Wolz, C., Ohlsen, K., Otto, M., Rudel, T., Sinha, B., and Fraunholz, M. (2014) Cytoplasmic replication of *Staphylococcus aureus* on phagosomal escape triggered by phenol-soluble modulins. *Cell. Microbiol.* **16**, 451–465
 62. Roux, A., Todd, D. A., Velázquez, J. V., Cech, N. B., and Sonenshein, A. L. (2014) CodY-Mediated Regulation of the *Staphylococcus aureus* Agr System Integrates Nutritional and Population Density Signals. *J. Bacteriol.* **196**, 1184–1196

Part III

REFERENCES

BIBLIOGRAPHY

- Becker, A. K., M. Dörr, S. B. Felix, F. Frost, H. J. Grabe, M. M. Lerch, M. Nauck, U. Völker, H. Völzke, and L. Kaderali (2021). From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. *PLoS Computational Biology* 17(2).
- Becker, A.-K., H. Erfle, M. Gunkel, N. Beil, L. Kaderali, and V. Starkuviene (2018). Comparison of cell arrays and multi-well plates in microscopy-based screening. *High-throughput* 7(2), 13.
- Becker, A.-K. and H. Holzmann (2019). Nonparametric identification in the dynamic stochastic block model. *IEEE Transactions on Information Theory* 65(7), 4335–4344.
- Becker, A.-K., T. Ittermann, M. Dörr, S. B. Felix, M. Nauck, A. Teumer, U. Völker, H. Völzke, L. Kaderali, and N. Nath (2021). Discovering association patterns of individual serum Thyrotropin concentrations using machine learning: An example from the Study of Health in Pomerania (SHIP). *PLoS Computational Biology* (under review).
- Becker, A.-K. and L. Kaderali (2020). GroupBN: Inferring Group Bayesian Networks using Hierarchical Feature Clustering. *CRAN R package version 1.2.0*.
- Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.
- Bratko, I. (1997). Machine learning: Between accuracy and interpretability. In *Learning, networks and statistics*, pp. 163–177. Springer.
- Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5–32.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (2017). *Classification and regression trees*. Routledge.
- Bro, R. and A. K. Smilde (2014). Principal component analysis. *Analytical Methods* 6(9), 2812–2831.

- Burra, P., C. Becchetti, and G. Germani (2020). NAFLD and liver transplantation: Disease burden, current management and future challenges. *JHEP reports*, 100192.
- Buzzetti, E., M. Pinzani, and E. A. Tsochatzis (2016). The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism* 65(8), 1038–1048.
- Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168.
- Carvalho, D. V., E. M. Pereira, and J. S. Cardoso (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics* 8(8), 832.
- Caterini, A. L. and D. E. Chang (2018). *Deep Neural Networks in a Mathematical Framework*. Springer.
- Cazanave, S., A. Podtelezhnikov, K. Jensen, M. Seneshaw, D. P. Kumar, H. K. Min, P. K. Santhekadur, B. Banini, A. G. Mauro, A. M. Oseini, R. Vincent, K. Q. Tanis, A. L. Webber, L. Wang, P. Bedossa, F. Mirshahi, and A. J. Sanyal (2017). The Transcriptomic Signature of Disease Development and Progression of Nonalcoholic Fatty Liver Disease. *Scientific Reports*.
- Chavent, M., V. Kuentz, A. Labenne, B. Liquet, and J. Saracco (2017). PCAmixdata: Multivariate Analysis of Mixed Data. *CRAN R package version 3.1*.
- Chavent, M., V. Kuentz-Simonet, B. Liquet, and J. Saracco (2012). ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*.
- Chickering, D. M., D. Geiger, D. Heckerman, et al. (1994). Learning Bayesian networks is NP-hard. *Microsoft Research* (1999), 94–17.
- Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1), 266–298.
- Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence* 42(2-3), 393–405.
- Dagum, P. and M. Luby (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial intelligence* 60(1), 141–153.

- Davies, E. R. (2017). *Computer vision: principles, algorithms, applications, learning*. Academic Press.
- Drescher, H. K., S. Weiskirchen, and R. Weiskirchen (2019). Current status in testing for nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH). *Cells* 8(8), 845.
- Drton, M., B. Sturmfels, and S. Sullivant (2008). *Lectures on algebraic statistics*, Volume 39. Springer Science & Business Media.
- Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*.
- Erickson, B. J. (2021). Artificial intelligence in medicine: Technical basis and clinical applications. In *Artificial Intelligence in Medicine*, pp. 19–34. Elsevier.
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.
- Fisher, A., C. Rudin, and F. Dominici (2018). Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the “Rashomon” Perspective. *Journal of Machine Learning Research* 20.
- Franceschini, N., D. I. Chasman, R. M. Cooper-DeHoff, and D. K. Arnett (2014). Genetics, ancestry, and hypertension: implications for targeted antihypertensive therapies. *Current hypertension reports* 16(8), 1–9.
- Greenacre, M. and J. Blasius (2006). *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC.
- Gyftodimos, E. and P. A. Flach (2002). Hierarchical bayesian networks: A probabilistic reasoning model for structured domains. In *Proceedings of the ICML-2002 Workshop on Development of Representations*, pp. 23–30. Citeseer.
- Hamon, R., H. Junklewitz, and J. Sanchez Martin (2020). *Robustness and Explainability of Artificial Intelligence* (JRC119336 ed.). Luxembourg: Publications Office of the European Union.
- Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. Series c (applied statistics)* 28(1), 100–108.

- Haury, A.-C., P. Gestraud, and J.-P. Vert (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one* 6(12), e28210.
- Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3), 197–243.
- Hennessy, C., A. Bugarin, and E. Reiter (2020). Explaining Bayesian Networks in Natural Language: State of the Art and Challenges. *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 28–33.
- Hira, Z. M. and D. F. Gillies (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics* 2015.
- James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.
- Jović, A., K. Brkić, and N. Bogunović (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, 1200–1205.
- Kiers, H. A. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika* 56(2), 197–212.
- Kim, B., R. Khanna, and O. Koyejo (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*.
- Kingma, D. and M. Welling (2014). Efficient gradient-based inference through transformations between bayes nets and neural nets. In *International Conference on Machine Learning*, pp. 1782–1790. PMLR.
- Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
- Krakovna, V. (2016). *Building Interpretable Models: From Bayesian Networks to Neural Networks*. Ph. D. thesis, Harvard University.
- Kyrimi, E., S. Mossadegh, N. Tai, and W. Marsh (2020). An incremental explanation of inference in bayesian networks for increasing model trustworthiness and supporting clinical decision making. *Artificial intelligence in medicine* 103, 101812.

- Lacave, C. and F. J. Díez (2002). A review of explanation methods for Bayesian networks. *Knowledge Engineering Review* 17(2), 107–127.
- Landwehr, N., M. Hall, and E. Frank (2005). Logistic model trees. *Machine Learning* 59(1-2), 161–205.
- Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy* 23(1), 1–45.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3), 31–57.
- Liu, S., X. Wang, M. Liu, and J. Zhu (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*.
- Lundberg, S. M., G. G. Erion, and S.-I. Lee (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.
- Madiega, T. EU guidelines on ethics in artificial intelligence: Context and implementation. *European Parliamentary Research Service PE 640.163*.
- Madsen, H. and P. Thyregod (2010). *Introduction to general and generalized linear models*. CRC Press.
- Maier, A., C. Syben, T. Lasser, and C. Riess (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik* 29(2), 86–101.
- Masarone, M., V. Rosato, M. Dallio, A. G. Gravina, A. Aglitti, C. Loguercio, A. Federico, and M. Persico (2018). Role of oxidative stress in pathophysiology of nonalcoholic fatty liver disease. *Oxidative medicine and cellular longevity* 2018.
- Masís, S. (2021). *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*. Packt Publishing Ltd.
- Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 411–418. Morgan Kaufmann Publishers Inc.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267, 1–38.

- Mills, K. T., A. Stefanescu, and J. He (2020). The global epidemiology of hypertension. *Nature Reviews Nephrology* 16(4), 223–237.
- Moini, J., M. Samsam, and K. Pereira (2020). Epidemiology of thyroid disorders. *Epidemiology of Thyroid Disorders*, 1–336.
- Molnar, C. (2020). *Interpretable machine learning: A Guide for Making Black Box Models Explainable*.
- Montavon, G., W. Samek, and K.-R. Müller (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73, 1–15.
- Muchlinski, D., D. Siroky, J. He, and M. Kocher (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis* 24(1), 87–103.
- Müller, H., A.-K. Becker, G. J. Palm, L. Berndt, C. P. Badenhorst, S. P. Godehard, L. Reisky, M. Lammers, and U. T. Bornscheuer (2020). Sequence-based prediction of promiscuous acyltransferase activity in hydrolases. *Angewandte Chemie International Edition* 59(28), 11607–11612.
- Müller, H., S. P. Godehard, G. J. Palm, L. Berndt, C. P. Badenhorst, A.-K. Becker, M. Lammers, and U. T. Bornscheuer (2021). Discovery and design of family VIII carboxylesterases as highly efficient acyltransferases. *Angewandte Chemie International Edition* 60(4), 2013–2017.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Njah, H., S. Jamoussi, and W. Mahdi (2019). Deep Bayesian network architecture for Big Data mining. *Concurrency and Computation: Practice and Experience* 31(2), e4418.
- Palma Medina, L. M., A.-K. Becker, S. Michalik, K. Surmann, P. Hildebrandt, M. Gesell Salazar, S. A. Mekonnen, L. Kaderali, U. Völker, and J. M. van Dijk (2020). Interaction of staphylococcus aureus and host cells upon infection of bronchial epithelium during different stages of regeneration. *ACS infectious diseases* 6(8), 2279–2290.
- Palma Medina, L. M., A.-K. Becker, S. Michalik, H. Yedavally, E. J. Raineri, P. Hildebrandt, M. G. Salazar, K. Surmann, H. Pfortner, S. A. Mekonnen, et al. (2019). Metabolic cross-talk between human bronchial epithelial cells and internalized

- staphylococcus aureus as a driver for infection. *Molecular & Cellular Proteomics* 18(5), 892–908.
- Pearl, J. (1985). Bayesian Networks A Model of Self-Activated Memory for Evidential Reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society*, 329–334.
- Ranzato, M., Y. L. Boureau, and Y. Le Cun (2009). Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference*.
- Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why should i trust you? - explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
- Rish, I. et al. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Volume 3, pp. 41–46.
- Rubio, A. and J. A. Gámez (2011). Flexible learning of k-dependence Bayesian network classifiers. *Genetic and Evolutionary Computation Conference, GECCO'11*, 1219–1226.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), 206–215.
- Sakr, A., F. Brégeon, J.-L. Mège, J.-M. Rolain, and O. Blin (2018). Staphylococcus aureus nasal colonization: an update on mechanisms, epidemiology, risk factors, and subsequent infections. *Frontiers in microbiology* 9, 2419.
- Santhanam, P., T. Nath, F. K. Mohammad, and R. S. Ahima (2020). Artificial intelligence may offer insight into factors determining individual TSH level. *PloS one* 15(5), e0233336.
- Scutari, M., C. E. Graafland, and J. M. Gutiérrez (2019). Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning* 115, 235–253.
- Scutari, M. and R. Nagarajan (2013). Identifying significant edges in graphical models of molecular networks. *Artificial intelligence in medicine* 57(3), 207–217.

- Serre, T., L. Wolf, and T. Poggio (2005). Object recognition with features inspired by visual cortex. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*.
- Skanski, S. (2018). *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Springer.
- Taylor, P. N., D. Albrecht, A. Scholz, G. Gutierrez-Buey, J. H. Lazarus, C. M. Dayan, and O. E. Okosieme (2018). Global epidemiology of hyperthyroidism and hypothyroidism. *Nature Reviews Endocrinology* 14(5), 301–316.
- Timmer, S. T., J. J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij (2017). A two-phase method for extracting explanatory arguments from Bayesian networks. *International Journal of Approximate Reasoning* 80, 475–494.
- Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* 32(24), 18069–18083.
- Vigneau, E. and E. M. Qannari (2003). Clustering of Variables Around Latent Components. *Communications in Statistics Part B: Simulation and Computation* 32(4), 1131–1150.
- Völzke, H., D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, et al. (2011). Cohort profile: the Study of Health in Pomerania. *International journal of epidemiology* 40(2), 294–307.
- Völzke, H., G. Fung, T. Ittermann, S. Yu, S. E. Baumeister, M. Dörr, W. Lieb, U. Völker, A. Linneberg, T. Jørgensen, et al. (2013). A new, accurate predictive model for incident hypertension. *Journal of hypertension* 31(11), 2142–2150.
- Wehenkel, M., A. Sutura, C. Bastin, P. Geurts, and C. Phillips (2018). Random forests based group importance scores and their statistical interpretation: Application for Alzheimer’s disease. *Frontiers in neuroscience* 12, 411.
- Wu, Y., H. Tjelmeland, and M. West (2007). Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics* 16(1), 44–66.
- Yuan, C. and M. J. Druzdzel (2006). Importance sampling algorithms for Bayesian networks: Principles and performance. *Mathematical and Computer Modelling* 43(9-10), 1189–1207.

Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research* 5, 697–723.

Zhou, W. and L. Nakhleh (2012). Convergent evolution of modularity in metabolic networks through different community structures. *BMC evolutionary biology* 12(1), 1–14.

LIST OF ABBREVIATIONS AND SYMBOLS

\mathbb{P}	probability distribution
ε	residuals
Θ	set of parameters
G	graph
R^2	R-squared metric
X	random vector
Y	target variable
Z	latent random variable or vector
AI	artificial intelligence
BART	Bayesian additive regression trees
BIC	Bayesian information criterion
BN	Bayesian network
BNC	Bayesian network classifier
CPDAG	completed partially directed acyclic graph
CRAN	Comprehensive R Archive Network
DAG	directed acyclic graph
EM	expectation-maximization algorithm
FA	fatty acids
FE	feature extraction
FSS	feature subset selection
GLM	generalized linear model

GMM	Gaussian mixture model
HBE	human bronchial epithelial cells
HBN	hierarchical Bayesian network
LCM	latent class model
LM	linear model
MCA	multiple correspondence analysis
MCMC	markov chain monte carlo sampling
ML	machine learning
MLE	maximum likelihood estimate
MSE	mean squared error
NAFLD	non-alcoholic fatty liver disease
$\text{par}(i)$	set of parent nodes
PC	principal component
PCA	principal component analysis
PGM	probabilistic graphical model
RF	Random Forest
ROS	reactive oxygen species
S. aureus	Staphylococcus aureus
SHIP	Study of Health in Pomerania
SNP	single nucleotide polymorphism
T ₃	triiodothyronine
T ₄	thyroxine
TSH	thyrotropin

EIGENSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, dass diese Arbeit bisher von mir weder an der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Greifswald noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Promotion eingereicht wurde.

Ferner erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die darin angegebenen Hilfsmittel und Hilfen benutzt und keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen habe.

(Unterschrift des Promovenden)

LIST OF PUBLICATIONS

THESIS ARTICLES

- I Becker, A. K., M. Dörr, S. B. Felix, F. Frost, H. J. Grabe, M. M. Lerch, M. Nauck, U. Völker, H. Völzke, and L. Kaderali (2021). From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. *PLoS Computational Biology* 17(2)
- II Becker, A.-K., T. Ittermann, M. Dörr, S. B. Felix, M. Nauck, A. Teumer, U. Völker, H. Völzke, L. Kaderali, and N. Nath (2021). Discovering association patterns of individual serum Thyrotropin concentrations using machine learning: An example from the Study of Health in Pomerania (SHIP). *PLoS Computational Biology* (under review)
- III Palma Medina, L. M., A.-K. Becker, S. Michalik, H. Yedavally, E. J. Raineri, P. Hildebrandt, M. G. Salazar, K. Surmann, H. Pförtner, S. A. Mekonnen, et al. (2019). Metabolic cross-talk between human bronchial epithelial cells and internalized staphylococcus aureus as a driver for infection. *Molecular & Cellular Proteomics* 18(5), 892–908

FURTHER PUBLICATIONS

- Becker, A.-K., H. Erfle, M. Gunkel, N. Beil, L. Kaderali, and V. Starkuviene (2018). Comparison of cell arrays and multi-well plates in microscopy-based screening. *High-throughput* 7(2), 13
- Becker, A.-K. and H. Holzmann (2019). Nonparametric identification in the dynamic stochastic block model. *IEEE Transactions on Information Theory* 65(7), 4335–4344

- Müller, H., A.-K. Becker, G. J. Palm, L. Berndt, C. P. Badenhorst, S. P. Godehard, L. Reisky, M. Lammers, and U. T. Bornscheuer (2020). Sequence-based prediction of promiscuous acyltransferase activity in hydrolases. *Angewandte Chemie International Edition* 59(28), 11607–11612
- Müller, H., S. P. Godehard, G. J. Palm, L. Berndt, C. P. Badenhorst, A.-K. Becker, M. Lammers, and U. T. Bornscheuer (2021). Discovery and design of family VIII carboxylesterases as highly efficient acyltransferases. *Angewandte Chemie International Edition* 60(4), 2013–2017
- Palma Medina, L. M., A.-K. Becker, S. Michalik, K. Surmann, P. Hildebrandt, M. Gesell Salazar, S. A. Mekonnen, L. Kaderali, U. Völker, and J. M. van Dijk (2020). Interaction of staphylococcus aureus and host cells upon infection of bronchial epithelium during different stages of regeneration. *ACS infectious diseases* 6(8), 2279–2290

Part IV

APPENDIX



ADDITIONAL NOTES ON BAYESIAN NETWORKS

The concept of Bayesian networks lies at the intersection of statistics, graph theory and machine learning. In addition to the brief description that is part of the cumulative thesis, this chapter summarizes some of the fundamentals and basic concepts upon which the theory builds.

1 CONDITIONAL PROBABILITY AND BAYES THEOREM

We denote with $\mathbb{P}(A)$ the probability that an event A is true. The *conditional probability* of an event A given B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0.$$

Further, the combination of the definition of conditional probability with the *product and sum rules* of probability yields the following:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

which is known as *Bayes Theorem*.

For the following, we consider an index set $V = \{1, 2, \dots, n\}$ and a related random vector $X_V = (X_1, X_2, \dots, X_n)$. We denote by $\mathbb{P}(X_i)$ the *marginal distribution* of X_i , and by $\mathbb{P}(X_1, \dots, X_n)$ the *joint distribution* of X_V .

2 (CONDITIONAL) INDEPENDENCE

Two random variables X_i and X_j are *conditionally independent* given a set of random variables $X_S \subset X_V$, if their joint distribution factorizes into the product of the *conditional marginals*

$$\mathbb{P}(X_i, X_j | \mathcal{S}) = \mathbb{P}(X_i | \mathcal{S})\mathbb{P}(X_j | \mathcal{S}).$$

We write $X_i \perp\!\!\!\perp X_j | \mathcal{S}$. It means that the realization of X_i does not affect the distribution of X_j . Accordingly, more than two random variables are called *mutually conditionally independent*, if their joint distribution can be written as the product of all conditional marginals. Further, two random variables X_i and X_j are said to be *directly dependent* if they are *not* conditionally independent given any other subset of $\mathcal{V} \setminus \{X_i, X_j\}$.

These definitions generalize to subsets of random variables \mathcal{A}_1 and $\mathcal{A}_2 \subset \mathcal{V}$.

3 FACTORIZATION BY CHAIN RULE

According to to the *chain rule* of probability, the joint distribution of X_V factorizes to a product of conditionals:

$$\mathbb{P}(X_V) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1, X_2) \cdots \mathbb{P}(X_n|X_1, \dots, X_{n-1}).$$

4 GRAPH TERMINOLOGY

A *graph* is the formal representation of a network and consists of *nodes* and *edges*. Here, we denote a graph G as a pair $G = (V, E)$, where V is a set of nodes $V = \{1, \dots, n\}$, and E is a set of edges $E \subset V \times V$ between the nodes. A graph is called *directed* if the edges have an orientation.

A *path* in a graph G is a finite number of edges connecting a sequence of nodes. A *directed path* is a path consisting of directed edges with the property that the ending node of each edge in the sequence is the starting node of the next edge in the sequence; a directed path forms a *directed cycle* if the starting node of its first edge equals the ending node of its last edge.

If there is an edge between node i and j in E , we call i and j *neighbors* in the graph G . We denote by nbrs the map which assigns each node to its neighbors

$$\text{nbrs} : V \longrightarrow V; i \longmapsto \text{nbrs}(i) := \{j \in V \mid i \text{ is neighbour of } i \text{ in } G\}.$$

The number of neighbors is called *degree* of a node. In a directed graph, if E contains the directed edge from node i to node j , then we call i a *parent* of j and j a *child* of i . We denote by $\text{par}(i)$ the map which assigns each node to its parents. A node without any parent nodes is called *root* of the network. More general, if there is any directed path from i to j , we say that i is an *ancestor* of j , $i \in \text{an}(j)$, and j is a *descendant* of i , $j \in \text{de}(i)$. The non-descendants of a node i are defined as the set $\text{nd}(i) = V \setminus \{i \cup \text{de}(i)\}$.

A *directed acyclic graph* (DAG) is a directed graph G that has no directed cycles. Its key property is that there is an inherent topological ordering that places parents before children. The *skeleton* of a DAG is an undirected graph that is obtained by replacing all directed by undirected edges. The *moralization* of a DAG is its skeleton with additional edges between nodes that have a common child ('marrying all parents').

5 BAYESIAN NETWORKS

In *probabilistic graphical models* (PGM), random variables are referred to as nodes in a graph and edges encode (in)dependencies between these variables. For a set of nodes $V = \{1, \dots, n\}$, each node i is associated with a random variable X_i . A *Bayesian network* is a directed probabilistic graphical model on a directed acyclic graph. It makes use of the topological ordering and the Markov property to factorize the joint distribution simpler than according to the chain rule. Fig. 8 shows an example of a Bayesian network and its encoded independence statements.

6 V-STRUCTURES AND D-SEPARATION

In a DAG, two nodes can be connected via a third one via three patterns that are shown in Fig 9: Head-to-head, head-to-tail or tail-to-tail. Head-to-head nodes, also known as v-structures, are central structures in Bayesian networks, as they differ from the two other possibilities in terms of the implied conditional independencies.

We will see, that an unobserved tail-to-tail node or a head-to-tail node leaves a path *unblocked*. If it is observed, it *blocks* the path in terms of conditional independence

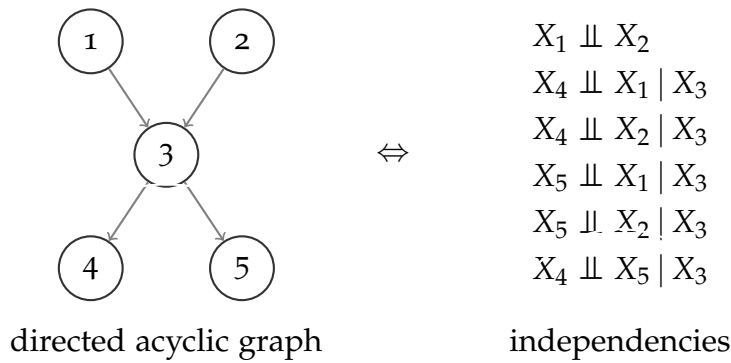


Figure 8: Example Bayesian network structure and encoded independence statements.

(and the flow of information). By contrast, a head-to-head node *blocks* a path if it is unobserved but once the node, and or at least one of its descendants, is observed the path becomes unblocked.

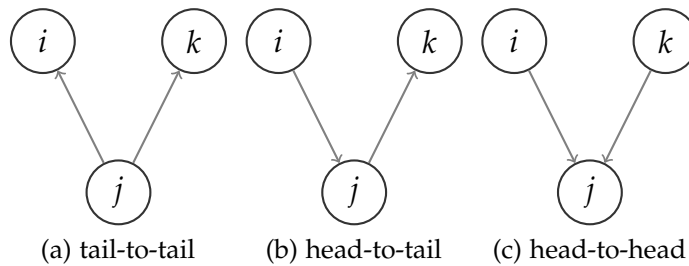


Figure 9: Three basic DAG structures consisting of three nodes and two edges.

In Figure 9(a) a DAG consisting of three nodes with a **tail-to-tail node** is shown. We show that X_i and X_k are not independent but conditionally independent given X_j .

$$\begin{aligned}
 \mathbb{P}(X_i, X_j, X_k) &= \mathbb{P}(X_i | X_j) \mathbb{P}(X_k | X_j) \mathbb{P}(X_j) \\
 \mathbb{P}(X_i, X_k) &= \sum_{X_j=x_j} \mathbb{P}(X_i, X_j = x_j, X_k) \neq \mathbb{P}(X_i) \mathbb{P}(X_k) && \Rightarrow X_i \not\perp X_k \\
 \mathbb{P}(X_i, X_k | X_j) &= \frac{\mathbb{P}(X_i, X_j, X_k)}{\mathbb{P}(X_j)} = \mathbb{P}(X_i | X_j) \mathbb{P}(X_k | X_j) && \Rightarrow X_i \perp X_k | X_j
 \end{aligned}$$

In Figure 9(b) a DAG consisting of three nodes with a **head-to-tail node** is shown. We show that X_i and X_k are not independent but conditionally independent given X_j .

$$\begin{aligned} \mathbb{P}(X_i, X_j, X_k) &= \mathbb{P}(X_i)\mathbb{P}(X_j | X_i)\mathbb{P}(X_k | X_j) \\ \mathbb{P}(X_i, X_k) &= \sum_{X_j=x_j} \mathbb{P}(X_i, X_j = x_j, X_k) \neq \mathbb{P}(X_i)\mathbb{P}(X_k) &\Rightarrow X_i \not\perp X_k \\ \mathbb{P}(X_i, X_k | X_j) &= \frac{\mathbb{P}(X_i, X_j, X_k)}{\mathbb{P}(X_j)} = \mathbb{P}(X_i | X_j)\mathbb{P}(X_k | X_j) &\Rightarrow X_i \perp X_k | X_j \end{aligned}$$

In Figure 9(c) a DAG consisting of three nodes with a **head-to-head node** is shown. We show that X_i and X_k are independent but not conditionally independent given X_j .

$$\begin{aligned} \mathbb{P}(X_i, X_j, X_k) &= \mathbb{P}(X_i)\mathbb{P}(X_k)\mathbb{P}(X_j | X_i, X_k) \\ \mathbb{P}(X_i, X_k) &= \sum_{X_j=x_j} \mathbb{P}(X_i, X_j = x_j, X_k) = \mathbb{P}(X_i)\mathbb{P}(X_k) \cdot 1 &\Rightarrow X_i \perp X_k \\ \mathbb{P}(X_i, X_k | X_j) &= \frac{\mathbb{P}(X_i, X_j, X_k)}{\mathbb{P}(X_j)} \neq \mathbb{P}(X_i | X_j)\mathbb{P}(X_k | X_j) &\Rightarrow X_i \not\perp X_k | X_j \end{aligned}$$

The concept of *d-separation* translates the concept of separation to directed graphs accordingly: Nodes i and j are *d-separated* by nodes \mathcal{S} if every path connecting node i and node j is blocked by at least one node in \mathcal{S} . Or more detailed: Nodes i and j are *d-connected* by nodes \mathcal{S} along a path from i to j if every head-to-head node along the path is in \mathcal{S} or has a descendant in \mathcal{S} and none of the other nodes (head-to-tail, tail-to-tail) is in \mathcal{S} . Nodes i and j are *d-separated* by nodes \mathcal{S} if they are not *d-connected* by \mathcal{S} along any path from i to j .

7 MARKOV PROPERTIES

A central assumption of PGMs is the *Markov property* that can be defined in increasing strength levels. We start with the discussion of undirected graphs and discuss the application to directed graphs afterwards.

With relation to an undirected graph $G = (V, E)$ an independence model satisfies the

- (i) *pairwise Markov-property*, if the absence of an edge implies conditional independence:

$$ij \notin E \Rightarrow X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}.$$

- (ii) *local Markov-property*, if all nodes are conditionally independent of non-neighbourhood nodes given their neighbours in the graph:

$$X_i \perp\!\!\!\perp X_{V \setminus (\text{nbrs}(i) \cup i)} \mid X_{\text{nbrs}(i)} \text{ for all } i \in V.$$

- (iii) *global Markov-property*, if separation in the graph implies conditional independence: For any A, B, C disjoint subsets of V ,

$$C \text{ separates } A \text{ and } B \text{ in } G \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_C.$$

The assumptions are arranged in order of increasing strength. If a distribution satisfies the global Markov property associated to the graph G , it necessarily satisfies the local Markov property on the graph G . And if a distribution satisfies the local Markov property on the graph G , it necessarily satisfies the pairwise Markov property. The reverse implication is not true in general. It can however be shown, that if the joint distribution \mathbb{P} of X is such that it satisfies the Intersection Axiom, then all three properties are equivalent. In particular, this is the case if \mathbb{P} is strictly positive.

The Markov-properties translate to the directed case as follows.

With relation to a directed graph $G = (V, E)$ an independence model satisfies the

- (i) *directed pairwise Markov-property*, if the absence of an edge implies conditional independence:

$$(i, j) \notin E \Rightarrow X_i \perp\!\!\!\perp X_j \mid X_{\text{nd}(i) \setminus j}.$$

- (ii) *directed local Markov-property*, if all nodes are conditionally independent of their non-descendants given their parents in the graph:

$$X_i \perp\!\!\!\perp X_{(\text{nd}(i) \setminus \Pi(i))} \mid X_{\Pi(i)} \text{ for all } i \in V.$$

- (iii) *directed global Markov-property*, if d-separation in the graph implies conditional independence: For any A, B, C disjoint subsets of V with

$$C \text{ d-separates } A \text{ and } B \text{ in } G \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_C.$$

Again, global implies local and local implies pairwise Markov property. Additionally, it can be shown, that local and global directed Markov-property are equivalent. That is why it is enough to require the fulfillment of the local Markov property in order to profit from the global condition (Drton et al., 2008; Koller and Friedman, 2009).

8 FAITHFULNESS

Due to the pairwise Markov condition the absence of an edge implies that there is no direct dependency between two random variables. However, the existence of an edge does not necessarily imply that there is a dependency, which is the reversal of the global Markov-property. Full equivalence between graph and distribution can be achieved by choosing faithful distributions.

An independence model \mathbb{P} is said to be *faithful* to a DAG G if G and \mathbb{P} imply exactly the same set of conditional independencies. A distribution \mathbb{P} *admits a faithful DAG representation* if \mathbb{P} is faithful to some DAG (Drton et al., 2008; Koller and Friedman, 2009).

Note that not every distribution admits a faithful DAG representation. However, Meek (1995) showed that for every DAG G it exists a multinomial or respectively a multivariate normal distribution that is faithful to G . Also, the set of unfaithful distributions is of Lebesque-measure zero.

If \mathbb{P} is faithful to G , then conditional independence and d-separation in the graph become equivalent

$$\begin{aligned} X_i \text{ and } X_j \text{ are conditionally independent given } \mathcal{S} \\ \Leftrightarrow i \text{ and } k \text{ are d-separated by } \mathcal{S} \text{ in } G. \end{aligned}$$

This generalizes to sets of variables. If \mathbb{P}_G is faithful to G , then

$$\begin{aligned} \mathcal{A}_1 \text{ and } \mathcal{A}_2 \text{ are conditionally independent given } \mathcal{S} \\ \Leftrightarrow i \text{ and } k \text{ are d-separated by } \mathcal{S} \text{ in } G \text{ for all } i \in \mathcal{A}_1 \text{ and } k \in \mathcal{A}_2. \end{aligned}$$

Under the assumption of Faithfulness and local Markov property, independence and d-separation become equivalent. Independence statements encoded by Bayesian networks can then be read from the graph using the criterion of d-separation. The easiest way to do this is to transform the according subgraph into an undirected graph, as d-Separation is equivalent to separation in the moralized *ancestral graph*, which consists of all nodes (involved in an independence statement) and its ancestors.

The following steps can be used to check if two nodes are d-separated by a given set of variables:

1. Determination of ancestral graph (remove all nodes that are not involved in the independence statement or are ancestors of such nodes)
2. Moralization of the ancestral graph (marrying all parents with a common child)
3. Disorientation of the ancestral graph (Remove all directions)
4. Deletion of the features which are evidence

If two variables are now separated, that means no path is left, they are conditionally independent.

10 MARKOV EQUIVALENCE

Two DAGs G and G' are said to be *Markov-equivalent* if they imply the same set of conditional independence statements, that is, for every Bayesian network $\mathcal{B} = (G, \Theta_G)$ there exists a Bayesian network $\mathcal{B}' = (G', \Theta_{G'})$ such that \mathcal{B} and \mathcal{B}' define the same probability distributions, and vice versa. This implies that G and G' have the same d-separations. Markov equivalence is an equivalence relation on the set of network structures. Two DAGs G_1 and G_2 are Markov equivalent iff they have the same skeleton and the same set of v-structures.

The BIC of a BN structure G is a common choice as scoring metric. It is defined as

$$\text{BIC}(G | \mathcal{D}) := \log \mathbb{P}(\mathcal{D} | G) + \frac{d}{2} \log(N),$$

where d the number of free parameters and N is the number of samples.

Whereas the first part increases linearly with N , the second part increases logarithmically. That means, as N grows large, the model fit is weighted more than the complexity. It decomposes to parts which are only dependent on one variable and its parents, allowing for efficient computations.

$$\begin{aligned} \text{BIC}(G | \mathcal{D}) &= -\log \mathbb{P}(\mathcal{D} | G) + \frac{d}{2} \log(n) \\ &= \sum_i \sum_j \sum_k -N_{ijk} \log \frac{N_{ijk}}{\sum_j N_{ij'k}} + \sum_i \frac{q_i(r_i - 1)}{2} \log(N) \\ &= \sum_i \text{BIC}(X_i, X_{\text{par}_G(i)} | \mathcal{D}), \end{aligned}$$

where N_{ijk} is the number of observations in which $X_i = k$ and $\Pi_G(X_i) = j$, q_i is the number of possible states of the parents $X_{\text{par}_G(i)}$ and r_i the number of possible states of X_i itself. The family score $\text{BIC}(X_i, \Pi_G(X_i) | \mathcal{D})$ can be calculated as

$$\text{BIC}(X_i, X_{\text{par}_G(i)} | \mathcal{D}) := -\sum_j \sum_k N_{ijk} \log \frac{N_{ijk}}{\sum_j N_{ij'k}} - \frac{q_i(r_i - 1)}{2} \log(N).$$

The BIC is globally and locally consistent. If a score is *globally consistent*, it means that in the limit as the number of observations grows large, the model with the lowest complexity, that represents the generative distribution exactly, is preferred. Conversely, *local consistency* refers to consistency in case of local perturbations.

Consider a DAG G and denote by G' the DAG which results from adding an arc $X_i \rightarrow X_j$ to G . A scoring metric S is locally consistent if it holds that

- i) $X_i \not\perp\!\!\!\perp X_j | X_{\text{par}_G(i)} \Rightarrow S(G'|D) > S(G|D)$
- ii) $X_i \perp\!\!\!\perp X_j | X_{\text{par}_G(i)} \Rightarrow S(G'|D) < S(G|D)$

Note that also the following holds: if G and G' are DAGs from the same Markov equivalence class, they are also score-equivalent.

Thus, if we consider a DAG G and G' , where G' results from G by adding one arc $j \rightarrow i$, then it holds

$$\text{BIC}(G' \mid \mathcal{D}) - \text{BIC}(G \mid \mathcal{D}) = \text{BIC}(X_i, X_{\text{par}_{G'}(i)} \mid \mathcal{D}) - \text{BIC}(X_i, X_{\text{par}_G(i)} \mid \mathcal{D}).$$

So that the score of a new network resulting from perturbing one arc, can be calculated using local calculations and the known score of the old network, only. This property is used in greedy search algorithms.

ADDITIONAL NOTES ON R-PACKAGE *GROUPBN*

GroupBN CHEAT SHEET

an R Package for fitting and automatic adaptive refinement of group Bayesian network models based on hierarchical variable clustering

Syntax builds on two external packages:

Clustering of heterogeneous data with ClustOfVar:

an R package for feature clustering of heterogeneous data

```
ClustOfVar::hclustvar(X.quant, X.quali)
plot(hierarchy)
plot(hierarchy, type="index")
```

see: <https://CRAN.R-project.org/package=ClustOfVar>

Handling Bayesian networks in R with bnlearn:

an R package for Bayesian network learning and inference

Central S3 classes:

Class *bn*
An object of class *bn* is a list containing at least the components:

- learning (a list with information about the algorithm)
- nodes (list)
- arcs (two-column matrix)

Class *bn.fit*
An object of class *bn.fit* is a list whose elements correspond to the nodes of the Bayesian network. Each element contains information about parents, children and parameters (format dependent on network type)

see: <https://www.bnlearn.com/>

GroupBN Learning

GroupBN: `groupbn(hierarchy, k, target, separate, X.quali, X.quant)`

takes a hierarchy (output of `hclustvar`), a number *k* of initial clusters, the name of target variables and potential additional special variables (like sex, age) and data split to quantitative and qualitative variables and learns a group Bayesian network

the output is an object of S3 class *groupbn* (contains the grouping and grouping parameters, Bayesian network structure and parameters, predictive metrics and additional information)

GroupBN Adaptive Refinement

GroupBN: `groupbn_refinement(res, hierarchy, refinement.part="mb")`

takes a *groupbn* object and a *hclustvar* hierarchy and refines the grouping adaptively in order to optimize the network's predictive performance. A refinement part can be chosen from the options Markov blanket (*mb*), second order Markov blanket (*mb2*) or all nodes (*all*).

GroupBN: `groupbn_refine_manually(res, hierarchy, refine="cl2")`

manually choose a group of variables to be refined

S3 methods for class *groupbn*

GroupBN: <code>predict</code> Predictions for new data	GroupBN: <code>print</code> prints a summary of the <i>groupbn</i> object
GroupBN: <code>plot</code> basic network plot, use <code>bnlearn::graphviz.plot</code> for nicer visualizations	GroupBN: <code>is.groupbn</code> test if an object is of class <i>groupbn</i>


GroupBN evaluation

GroupBN: `groupbn.output.table`



creates a table with one column per node, sorts the features by importance, importance scores can be added

GroupBN: `groupbn.vis.html.plot`



Creates an interactive html network object with `visNet`

additional functions

GroupBN: `cross.en`
Calculates the weighted cross-entropy / log-loss for a vector of observations and predicted target probabilities

GroupBN: `discretize.dens`
Density approximative data discretization

Figure 10: “Cheatsheet” for R-package *GroupBN*