HELMHOLTZ
MUNICH

Statistical Analyses of Combinatorial Effects in High-Throughput Biological Data

Mara Stefanie Stadler

München 2024

Statistical Analyses of Combinatorial Effects in High-Throughput Biological Data

Mara Stefanie Stadler

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Mara Stefanie Stadler

München, den 30. April 2024

Erster Berichterstatter: Prof. Dr. Christian L. Müller (LMU Munich)
Zweiter Berichterstatter: Prof. Dr. Jacob Bien (University of Southern California)
Dritter Berichterstatter: PD. Dr. Fabian Scheipl (LMU Munich)

Tag der Disputation: 30.07.2024

"The best thing about being a statistician is you get to play in everybody else's backyard"

John Tukey

Zusammenfassung

Das Aufkommen von großen Mengen an biologischen Zähldaten durch Hochdurchsatz-Technologien hat die Entwicklung geeigneter statistischer Methoden zu einer wichtigen Herausforderung moderner interdisziplinärer Forschung gemacht. Diese Daten weisen oft eine Vielzahl von Kovariablen auf, sind jedoch durch geringe Beobachtungsgrößen und experimentelles Rauschen limitiert. Eine zentrale Forschungsfrage in datengetriebenen Untersuchungen ist, wie biologische Kovariablen eine relevante Zielvariable beeinflussen. Meist sind nur einige der Kovariablen von Bedeutung. Diese können jedoch auf komplexe Art und Weise miteinander interagieren. Eine Hauptaufgabe besteht daher darin, die relevanten Effekte aus einer Vielzahl an möglichen Kombinationen zu identifizieren.

In dieser Arbeit habe ich Methoden entwickelt, die robuste Schätzungen von Interaktionseffekten durch quadratische Regressionsmodelle ermöglichen. Diese Methoden sind sowohl für Beobachtungs- als auch für experimentelle Daten geeignet, unabhängig davon, ob die experimentellen Designs vollständig sind. Die entwickelten Modelle berücksichtigen verschiedene Arten von biologischen Zähldaten: (i) quantitative Zähldaten, (ii) binäre Daten und (iii) relative Zähldaten, auch bekannt als kompositionelle Daten. Um in Szenarien mit mehr Kovariablen als Beobachtungen sowie in niedrigdimensionalen Szenarien interpretierbare Modelle zu entwickeln, habe ich in meinen Ansätzen Penalisierung verwendet. Durch die Integration von Konzepten der hierarchisch Interaktionsmodellierung und der stabilitätsbasierten Modellselektion wird die Interpretierbarkeit gewährleistet. Zur Reduktion ungewollter, auf technisches und biologisches Rauschen zurückzuführender Effekte sind Ansätze entwickelt worden, die weniger anfällig für Ausreißer sind. Dies ist von besonderer Bedeutung, wenn nur wenige und inkonsistente Replikate vorliegen.

In meinem ersten Projekt habe ich Daten der Affinitätsreinigung von Nukleosomen mit quantitativer Proteomik und hierarchischer Interaktionsmodellierung kombiniert. Ziel war es, die kombinatorischen Effekte bestimmter Chromatinmodifikationen auf die Proteinrekrutierung in einem unvollständigen experimentellen Design zu schätzen. Der hierfür entwickelte Workflow, `asteRIa`, ermöglicht eine stabile Schätzung robuster Interaktionen zwischen Chromatinmodifikationen und hat mehrere Proteine als epigenetische “Leser”-Kandidaten identifiziert.

In meinem zweiten Projekt habe ich ein generisches quadratisches Interaktionsmodell entwickelt, um Umwelt- oder Wirtsbedingungen aus Daten über die mikrobielle Abundanz vorherzusagen. Dieses Modell unterstützt verschiedene Datenmodalitäten und hat einen breiten Anwendbarkeitsbereich. Diesen habe ich auf unterschiedlichen Daten demonstriert und robuste Interaktionseffekte zwischen mikrobiellen Taxa aufgedeckt.

In meinem dritten Projekt habe ich Wechselwirkungen von Medikamenten in Hochdurchsatz-Screening-Verfahren für einzelne Zellen analysiert. Dabei habe ich hierarchische Interaktionsmodellierung mit einer Optimierungstechnik kombiniert, die robust gegenüber Ausreißern ist, und somit einen generischen und reproduzierbaren Workflow erstellt.

Insgesamt habe ich statistische Methoden zur Schätzung robuster Interaktionseffekte in biologischen Daten entwickelt. Die Modelle ermöglichen präzise Analysen verschiedener Datentypen und identifizieren Interaktionseffekte, die Hypothesen für weiterführende funktionelle Untersuchungen darstellen.

Summary

The advent of large-scale biological count data from high-throughput technologies has made the development of suitable statistical techniques a cornerstone of modern interdisciplinary research. These data often contain many features but limited sample size, and are accompanied by experimental noise. A common research question in data-driven observational studies is to determine how such biological features impact a readout of interest. Typically, only a subset of features is relevant, and they may interact in a concerted fashion. Thus, a major concern is to identify these relevant effects from a large number of possible combinations of features.

In this thesis, I developed and evaluated ways to estimate stable main and interaction effects via quadratic regression models in both observational and experimental data with complete or incomplete designs. The models developed are applicable to different data modalities in which biological count data typically appear: (i) quantitative count data, (ii) presence-absence data, and (iii) relative count data, also known as compositional data. To derive parsimonious models in underdetermined regimes, as well as in low- and moderate-dimensional settings, I implemented the models under penalization. To facilitate interpretability, I included the concept of hierarchy in interaction modeling and stability-based model selection. In order to account for technical and biological noise in the data, I introduced ways to be less sensitive towards outliers, especially when few and inconsistent replicates are available.

In my first project, I integrated nucleosome affinity purification data with high-throughput quantitative proteomics and hierarchical interaction modeling to estimate combinatorial effects of the presence or absence of certain chromatin modifications on protein recruitment within an incomplete experimental design study. This is facilitated by the computational workflow `asteRIa` which combines hierarchical interaction modeling, stability-based model selection, and replicate consistency checks for a stable estimation of robust interactions among chromatin modifications. `asteRIa` identifies several epigenetic "reader" candidate proteins responding to specific interactions between chromatin modifications.

In my second project, I developed a generic quadratic interaction model for the prediction of environmental or host-related conditions from observational and experimental microbial abundance data. The interaction model covers common data modalities of microbial data, ranging from quantitative microbiome and presence-absence information to compositional microbiome data. I demonstrated the broad applicability of our framework across various ecosystems and showcased how quadratic models improve predictive accuracy while uncovering stable interaction effects between microbial taxa when integrated with hierarchical interaction modeling and stability-based model selection.

In my third project, I analyzed drug interaction effects in high-content screening (HCS) cell studies. Here, I combined hierarchical interaction modeling with an optimization that is less sensitive to outliers within a generally applicable and reproducible computational workflow to analyze combinatorial effects in HCS data.

In summary, I have developed statistical approaches for the stable estimation of interaction effects, with a particular emphasis on high-throughput biological data. The workflows and statistical models I developed enable the precise analysis of various data types to reveal highly stable interaction effects, facilitating further functional analyses.

Acknowledgements

First and foremost, I want to thank my supervisor, Christian Müller, for giving me both the freedom and the guidance I needed during my work as a Ph.D. student. Thank you for believing in me, encouraging me to present my work at so many conferences, and supporting my research stay—all within an open and friendly research environment. Your support exceeded anything I could have hoped for. Additionally, I would like to thank Till Bartke for giving me the opportunity to learn far more about epigenetics than I ever imagined and for patiently answering all the questions I had. I feel very lucky to have had you as a collaborator and my domain-expert PI. Further, I would like to express my gratitude to Jacob Bien for allowing me to conduct a research stay in his lab at USC, providing invaluable scientific input, and agreeing to be my external reviewer. It is truly a pleasure having you as a collaborator. Thank you to Fabian Scheipl for not only agreeing to be on my advisory committee and a reviewer of my dissertation but also for showing great interest in my projects and providing valuable feedback over the years. A huge thank you to my advisory committee member, Maria Colomé-Tatché, for your support and input. I would like to thank all my collaborators, especially Saulius Lukauskas, Alisa Dietl, Erwin Kupczyk, and Lance Buckett as well as the COVID-19 data analysis (CODAG) group at LMU. Working on such exciting interdisciplinary projects with you has been a blast. Many thanks to the most amazing lab members. Without all of you, this would not have been the same. Thanks to Roberto Olayo Alarcon for being a great collaborator, for your support when my biological knowledge reached its limits, and for many fun moments. I am grateful to Viet Tran and Johannes Ostner for many enlightening discussions on statistics and so many other topics. Thanks to Stefanie Peschel, Oleg Vlasovets, and Daniele Pugno for many conversations about almost everything, including science, and all the fun activities—especially the via ferrata! I would further like to thank Luise Rauer, Fabian Schaipp, Jinlong Ru, Tong Wu, and Aditya Mishra. A special thank you to the best office mates, Yiqiu Shen, Katerina Giannoutsou, and Rashmi Ranjan Bhuyan at USC. Thank you for being so welcoming and supportive during my research stay. A big thank you to the Munich School for Data Science for both scientific and non-scientific support—and especially for the nice retreats and other events. I would also like to thank the DAAD for the financial support. Thanks to the community at the Institute for Functional Epigenetics, the Computational Health Center at Helmholtz Munich, and the Statistics Department at LMU. Moreover, I want to thank the administration of the Computational Health Department Anna Sacher, Daniela Herrmann, Julia Schlehe, and Mara Kieke. Thank you to my friends who have supported me over the past years. I want to thank my grandmother, Eva Marie Stadler, my parents, Pia and Hans Christian Stadler, and my sister, Moana Stadler, for their unlimited support. Finally, I want to thank Henning Herbers. Where to even begin? Without you, this would not have been possible.

List of contributed publications

This thesis is based on the following publications and manuscripts:

Contributions as first author

1. **Stadler, M.**, Lukauskas, S., Bartke, T., and Müller, C.L. (2024). asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers. *Nucleic Acids Research, 6129–6144.* doi: https://doi.org/10.1093/nar/gkae361 (See also publication [1] in the bibliography.)

2. **Stadler, M.**, Müller, C. L., Bien, J. (2024). Predictive modeling of microbial data with interaction effects. *bioRxiv, 2024-04.* doi: https://doi.org/10.1101/2024.04.29.591596
(See also manuscript [2] in the bibliography.)

Draft manuscript as joint first author

3. **Stadler, M.***, Kupczyk, E.*, Buckett, L., Zhang, X., Müller, C.L. (2024). A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach. *Draft manuscript.* *joint first co-authorship
(See also manuscript [3] in the bibliography.)

Contribution as co-author

4. Lukauskas, S., Tvardovskiy, A., Nguyen, N. V., **Stadler, M.**, Faull, P., Ravnsborg, T.,..., Bartke, T. (2024). Decoding chromatin states by proteomic profiling of nucleosome readers. *Nature, 1-9.* doi: https://doi.org/10.1038/s41586-024-07141-5
(See also publication [4] in the bibliography.)

Other publications not included in this thesis:

5. **Stadler, M.**, Doebler, P., Mertins, B., Delucchi Danhier, R. (2021). Statistical modeling of dynamic eye-tracking experiments: Relative importance of visual stimulus elements for gaze behavior in the multi-group case. *Behavior Research Methods*, 53, 2650-2667. doi: https://doi.org/10.3758/s13428-021-01576-8

6. Dietl, A., Ralser, A., Taxauer, K., Dregelies, T., Sterlacci, W., **Stadler, M.**,...,Mejias Luque, R. (2024). RNF43 is a gatekeeper for colitis-associated cancer. *In revision.* doi: https://doi.org/10.1101/2024.01.30.577936

Contents

1 Introduction

Novel high-throughput technologies have revolutionized the generation of large-scale biological data, enabling the parallel sequencing of millions of DNA molecules [5], which can be further processed into count data [6, 7], or allowing the simultaneous analysis of multiple cellular features through high-content screening [8]. The development of statistical tools for analyzing these data has become a key task in modern interdisciplinary research [9].

A typical research question in data-driven observational studies involves how a set of features affects some outcome of interest. However, if the question involves a multitude of features, it is likely that only a subset is relevant and that these features interact with one another. Possible biological questions include examining how certain drugs within an unbalanced combinatorial design interact to affect specific cellular components or how interactions between microbial species influence multiple community functions, such as the production of metabolites (see Fig. 1 for illustrations of the dimensions of the underlying datasets that are denoted by $X_{n\times p_1}$ and $Y_{n\times p_2}$).

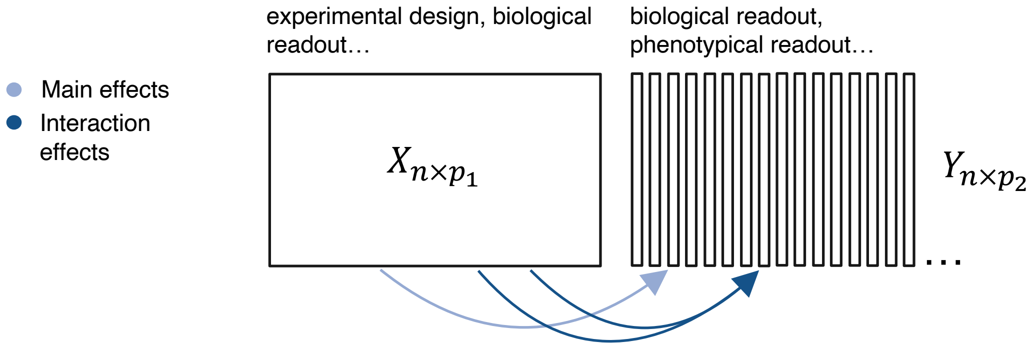

experimental design, biological readout...
biological readout, phenotypical readout...
● Main effects
● Interaction effects
$X_{n\times p_1}$
$Y_{n\times p_2}$
...

Figure 1: An illustration of a dataset $X_{n\times p_1}$ where it is assumed that a subset of the p_1 features exhibit both individual and combinatorial effects on certain features in $Y_{n\times p_2}$. The number of samples n may be smaller than the number of features p_1 and p_2. Each of the p_2 features in Y will be considered individually.

Statistically identifying such sparse sets of relevant features and interactions cannot be effectively approached by classical "hypothesis-driven" studies, as outlined by R.A. Fisher [10] and Neyman and Pearson [11] in the 1930s. Instead, more "data-driven" approaches for exploratory data analysis, as proposed by Tukey [12] in the late 1990s, are necessary. Specifically, in this thesis, I focus on deriving parsimonious quadratic regression models using regularization techniques [13]. While the models I define include quadratic interactions, they can naturally be extended to accommodate higher-order interaction models. In particular, the quadratic interaction model for each of the p_2 outcome variables

$y := Y_i \in \mathbb{R}^n$, $i = 1, \dots, p_2$ and all predictors $X = (X_1, \dots, X_{p_1}) \in \mathbb{R}^{n \times p_1}$ is given by

$$y = \beta_0 + \sum_{j=1}^{p_1} \beta_j X_j + \frac{1}{2} \sum_{j,k} \Theta_{jk} X_j X_k + \epsilon, \tag{1}$$

where β_0 is the intercept term, β_j represents the effect of feature X_j on y, $\Theta = \Theta^T \in \mathbb{R}^{p_1 \times p_1}$ is a matrix of interaction effects, and ϵ models the technical and biological noise.
Based on the sign of the interaction between two features X_j and X_k (Θ_{jk}) and the corresponding main effects β_j and β_k, I define six modes of combinatorial behavior (see Fig. 2). These modes describe potential combinatorial interactions that can arise in various biological research contexts. Depending on the context, they can be further extended to subcategories where a main effect is exactly zero, $\beta_j = 0$.

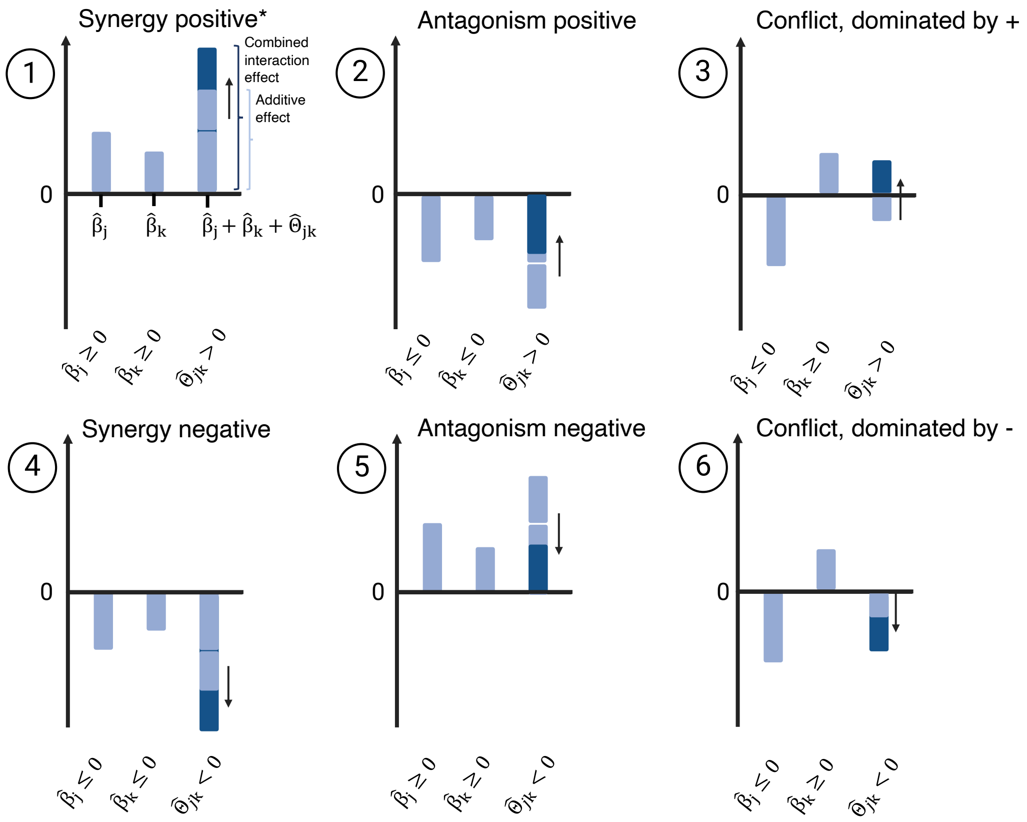


Figure 2: Modes of combinatorial behavior derived from the interaction model in Eq. 1 (created with BioRender.com). The first two bars in each mode display the individual effects of X_j and X_k, represented as $\hat{\beta}_j$ and $\hat{\beta}_k$ respectively. The third bar (dark blue) in each mode represents the overall combinatorial effect, which consists of the individual effects plus the additional combinatorial effect $\hat{\Theta}_{jk}$, calculated as $\hat{\beta}_j + \hat{\beta}_k + \hat{\Theta}_{jk}$. The light blue bar in the same column indicates the expected results under additivity (independence) between X_j and X_k, calculated as $\hat{\beta}_j + \hat{\beta}_k$.

1.1 Interaction effects in biology

Biological systems comprise a large number of different components (e.g., genes, microbial species, chemical compounds), that can all exhibit various modes and interact with one another. These interactions can change dynamically and are context-dependent [14].
The combinatorial modes, as depicted in Fig. 2 and encompassed by the quadratic interaction model from Eq. 1, are crucial in addressing a wide range of biological research questions. These include areas such as epigenetics, biological fitness, microbial ecology, and pharmacology.

Combinatorial Histone Code Chromatin, which is the nucleoprotein complex made up of DNA and histone proteins, controls the access to DNA and therefore plays a critical part in regulating gene expression [15]. Chromatin modifications are important contributors to chromatin regulation, playing a crucial role in orchestrating processes such as DNA transcription, replication, and repair. These modifications are known to recruit epigenetic reader proteins and often occur in specific combinations [16, 17, 18, 19, 20, 21, 22], suggesting that combinatorial chromatin modifications can encode epigenetic information, by generating synergistic or antagonistic interaction affinities for chromatin-associated proteins. This idea is known as the "histone code" hypothesis and has been proposed more than two decades ago [23, 24, 25]. While functions and readers of many individual chromatin modifications and few combinations have been described, a comprehensive analysis of this combinatorial behavior has not been feasible for many years due to technological limitations (see [1] for a more detailed review). Only recently has this changed with the publication of the modification atlas of regulation by chromatin states (MARCS) [4]. This novel data resource enables detailed studies of the fundamental principles of genome regulation by chromatin states. In [1], the MARCS data are integrated within a statistical interaction modeling framework to uncover previously unknown combinatorial chromatin modifications and epigenetic reader protein candidates. Specifically, these interactions are captured by the model in Eq. 1, where X represents a binary design matrix encoding the presence or absence of various combinations of chromatin modifications, and $y = Y_i$ denotes the binding affinity of the i-th protein out of p_2 proteins in an experiment. For example, a model coefficient of $\Theta_{jk} < 0$ but $\beta_j > 0$ and $\beta_k > 0$ would indicate that while the i-th protein binds both chromatin modifications X_j and X_k individually, the combined presence of these modifications results in weaker binding than would be expected under additivity, i.e., $\beta_j + \beta_k > \beta_j + \beta_k + \Theta_{jk}$. If, in this example, $|\Theta_{jk}| > \beta_j + \beta_k$, the combined effect could even completely inhibit binding or turn it into repulsion.

Epistatic Fitness Landscapes Epistatic fitness landscapes illustrate the relationship between the functional interplay of genes and their combined effect on fitness, specifically referring to the deviation from the expected additive effects of individual genes [26, 27, 28]. This deviation, known as epistasis, characterizes the non-linear interactions between gene pairs that either enhance (*synergistic epistasis*) or diminish (*antagonistic epistasis*) the organism's fitness beyond what would be predicted from the sum of their individual effects. A common approach to analyze these interactions is through linear regression, incorporating quadratic (and higher-order) terms to account for these complex gene-gene interactions [29, 30, 31]. Following the forward model definition in Eq. 1, with $y \in \mathbb{R}^n$ being the fitness and $X \in \{0,1\}^{n \times p_1}$ the binary information of gene mutations with p_1 genes, Θ_{jk} captures the epistatic effect between the j-th and k-th gene, where $\Theta_{jk} > 0$ indicates synergistic epistasis, $\Theta_{jk} < 0$ indicates antagonistic epistasis, and $\Theta_{jk} = 0$ signifies the absence of epistasis between the genes under consideration (see Fig. 3).

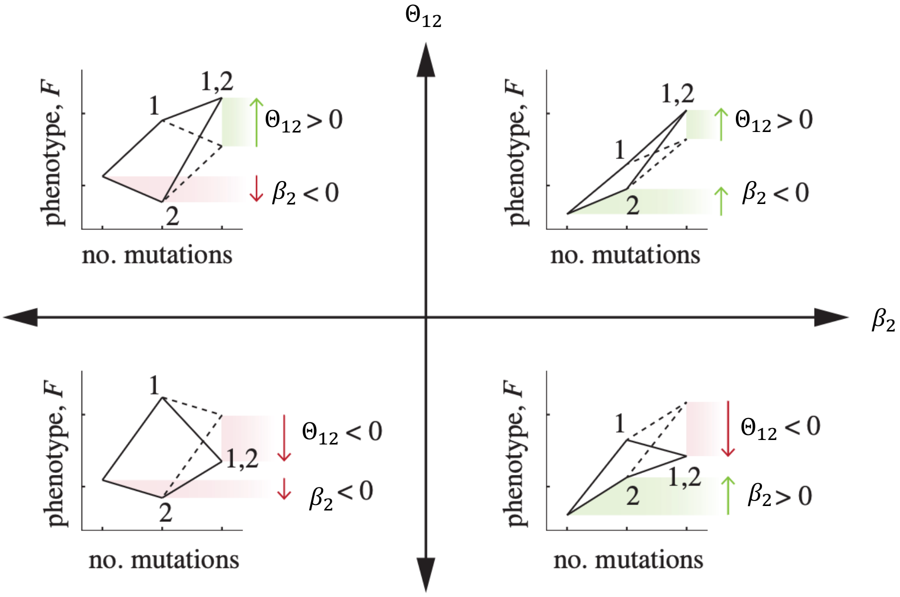


Figure 3: Illustration of different modes of epistasis according to the model in Eq. 1. In this example, the sign of the main effect β_2 indicates whether mutation 2 has a beneficial or deleterious effect in the absence of mutation 1. The main effect of mutation 1, β_1, is not varied here. Similarly, the sign of the interaction effect Θ_{12} determines the direction of fitness change when both mutations 1 and 2 are present, relative to their additive effects. Adapted from [30].

In the case of binary (presence-absence) input data, the interaction model in Eq. 1 can be associated with Taylor expansions for a 0/1 encoding and with Fourier expansions for a $-1/1$ encoding. The latter is particularly advantageous when y represents a fitness or phenotypical landscape, as the parameters of the Fourier expansions facilitate convenient descriptions of landscape properties, such as ruggedness [32, 31, 33].

Ecological function in microbial communities Microbial communities are characterized by a complex web of context-dependent inter- and intra-species interactions [14]. A major goal in microbial ecology is to understand how the composition of a community and the interplay within communities determine its function. In the context of the sharing of metabolites, interactions between microbes are also known as cross-feeding [34]. Such interactions between microbial populations are described with specific terms, including mutualism, parasitism, competition, or commensalism [35, 36, 34]. These descriptions focus on how species affect each other—for example, in parasitism, one species is harmed while the other benefits—rather than on how they affect a common outcome. Mathematically, community interactions are often studied using network approaches [37, 38, 39, 40] or within time-dynamic models [41, 42]. While these approaches do not directly connect interactions between microbial species (or broader taxonomic groups) to a function or a host-related or environmental outcome, a recent study introduced by [32] translates the interplay between communities and their function into a landscape concept. This concept is inspired by fitness landscapes from genetic epistasis and defines the community function (e.g., butyrate production) as a landscape that can be described by a quadratic regression model based on the presence or absence information of microbes within a community, without considering time dynamics. For instance, two species may exhibit a synergistic combinatorial behavior in producing a metabolite or compete for a resource. Further research [14] has connected community interactions with host fitness using quadratic regression modeling, revealing that bacterial interactions are not only context-dependent but also crucial for understanding host lifespan.

Mathematically, these interactions can be described by the interaction model in Eq. 1, where $y \in \mathbb{R}^n$ represents, for instance, a community function, species- or host fitness, and $X \in \mathbb{R}^{n \times p_1}$ represents the abundance information of a set of p_1 microbial species. When $\beta_j > 0$, $\beta_k > 0$, and $\Theta_{jk} > 0$, this is referred to as synergistic behavior, which may also be termed mutualism or cooperation depending on the context, indicating that both the j-th and the k-th microbial species "benefit" from the association. To give another more specific example, assume that the j-th species is a known butyrate producer ($\beta_j > 0$), while the k-th species is not ($\beta_k = 0$). However, if in co-culture the butyrate production is stopped or inhibited ($\Theta_{jk} < 0$), this suggests an inhibitory effect of the k-th species on the j-th species in the context of butyrate production.

Drug combination effects The statistical analysis of drug combination effects is pivotal in the identification of undesirable interactions and in the screening process for potential drug combinations in pharmacological research and clinical applications [43].

In pharmacology, the additive effect refers to a scenario where the combined effect of two drugs is equivalent to the sum of their individual effects. Deviations from this additive effect are commonly described through the concepts of synergistic and antagonistic effects. One popular approach in the analysis of drug combination data is the Bliss independence model [44], which focuses on the enhancement of treatment effects. The Bliss independence model, derived from the complete additivity of probability theory, serves as a robust reference model [45, 46]. Assuming two drugs, A and B, operate through distinct pathways with no mechanistic connection other than the response outcome under treatment t, y_t, $t \in \{A, B, AB\}$, the Bliss independence principle is given by the predicted combination response

$$\hat{y}_{AB} = y_A + y_B - y_A y_B,$$

that is compared to the observed combination response y_{AB}. This comparison can be summarized by three scenarios, namely

$$y_{AB} = \begin{cases} > \hat{y}_{AB} & \text{synergy} \\ = \hat{y}_{AB} & \text{independence} \\ < \hat{y}_{AB} & \text{antagonism.} \end{cases}$$

This comparison is generally conducted across all possible dose combinations. When replicates are available, the average percentage of inhibition is typically reported [45]. Another fundamental approach in drug interaction studies is the Loewe additivity model [47]. It assumes that two drugs acting through the same mechanism should show dose-additive effects. This model is particularly useful when studying drug combinations where the drugs are known to act through the same biological pathway.
The quadratic interaction model is a more recent development in drug-interaction modeling [48, 49]. It allows for the modeling of complex interactions, including synergistic and antagonistic effects, and is particularly useful when studying drug combinations with unknown or complex mechanisms of action or in large-scale observational data studies. Such interactions can be described by the model in Eq. 1, with $X \in \mathbb{R}^{n \times p_1}$ representing an (incomplete) drug design encompassing p_1 different drugs, and $y = Y_i$, with $Y \in \mathbb{R}^{n \times p_2}$, representing a single or multiple outcome comprising p_2 features (e.g., a patient's health status, but also a large set of gene expressions or cellular features). Here, the effects β_j and β_k indicate the individual effects, and Θ_{jk} represents the additional combinatorial effect of the j-th and the k-th drug. For instance, if $\beta_j > 0$ and $\beta_k > 0$, but $\Theta_{jk} < 0$, both drugs show a positive effect on the outcome y, but in combination, they build an antagonistic effect.

1.2 Research question

Although quadratic interaction modeling is well established [50, 51, 52], estimating interaction effects remains notoriously challenging in the presence of noisy, scarce data, or incomplete experimental designs, and these effects are prone to misinterpretation [53, 54]. The primary objective of this thesis is to develop statistical concepts for estimating parsimonious quadratic models with stable main and interaction effects, specifically within the context of biological high-throughput data. The methods developed in this thesis aim to address the challenges of noise and data scarcity while being applicable to various data modalities in which biological count data typically appear. In particular, I address specific questions in three biological contexts:

I) How to quantitatively estimate and validate novel chromatin modification interaction effects on protein recruitment using a unique and novel large-scale proteomics dataset?

II) How to create a generic statistical workflow to model microbial interaction effects on community function, or host- and environment-related outcomes?

III) How to develop a statistical workflow to estimate robust drug interaction effects on the morphological features of single cells derived from high-content screening?

1.3 Summary of results

In this thesis, I introduce statistical concepts tailored for deriving stable interaction effects in quadratic regression models, with a particular focus on biological count data from high-throughput experiments. This work bridges the gap between existing modeling approaches and the specific properties of the biological data by integrating solutions that account for data scarcity, incomplete designs, and experimental noise.

Specifically, the methods I propose address all data modalities commonly encountered in biological count data: (i) quantitative count data, (ii) presence-absence data, and (iii) relative count data, also known as compositional data. Additionally, I outline strategies for generating simulated data, which enable the evaluation of the model's accuracy in detecting interactions and provide methods for summarizing and visualizing the findings effectively.

By applying the methods I developed to specific biological contexts, I demonstrate their versatility and their capability to generate novel biological insights.

This thesis comprises one publication, one accepted manuscript, one submitted manuscript, and one draft manuscript, each contributing to the development of statistical approaches for the detection of stable interaction effects across various data modalities.

In [4] and [1], I explored research question I): Chromatin modifications are key players in regulating gene expression. A detailed understanding of the interplay of chromatin modifications in recruiting epigenetic "reader" proteins remained largely elusive. Publication [4] presents a novel nucleosome affinity purification dataset with high-throughput quantitative proteomics, as provided in the modification atlas of regulation by chromatin states (MARCS) enabling the discovery of fundamental principles of genome regulation by chromatin states. In [1], I developed a statistical workflow, termed `asteRIa`, for the detection and validation of stable interaction effects in the data-scarce regime when few and inconsistent replicates are available. Integrating the MARCS data with the `asteRIa` workflow provides the first quantitative framework to estimate combinatorial effects of chromatin modifications on protein recruitment. In [2], I explored research question II): Microbial interactions play a pivotal role in shaping microbial communities and their functions. Statistical tools that derive robust and interpretable interaction effects between microbial taxa on community function, or on host- and environment-related outcomes are crucial for understanding these complex relationships. By defining and evaluating novel ways of modeling interactions in compositional data and combining this with interaction modeling

strategies for absolute counts and presence-absence data, I developed a generic framework for interaction modeling in microbial data. In [3], research question III) was explored: High-content screening (HCS) is an important tool to study drug effects by offering a comprehensive cell-level view [55] and derives a multitude of data describing morphological features as summary statistics of single cells [56]. By combining robust interaction modeling with post-estimation data summaries, I developed a generally applicable framework for estimating drug interactions in HCS studies.
A comprehensive summary of the study findings is presented below.

- Contribution [4] in Appendix B.1: *Decoding chromatin states by proteomic profiling of nucleosome readers*
  Chromatin, the nucleoprotein complex consisting of DNA and histone proteins, plays a crucial role in regulating gene expression by controlling access to DNA. Chromatin modifications are key players in this regulation, as they help to orchestrate DNA transcription, replication, and repair. These modifications recruit epigenetic "reader" proteins, which mediate downstream events. While many reader proteins of individual modifications have been described [57, 58, 59], the interpretation of chromatin states comprising composite modification signatures, histone variants, and internucleosomal linker DNA remains a major open question. This study combines novel stable isotope labeling with amino acids in cell culture (SILAC) nucleosome affinity purification (SNAP) data [60] that probe the binding of proteins from HeLa S3 nuclear extracts to a library of semi-synthetic di-nucleosomes with high-throughput quantitative proteomics. The results of this study are presented as online resource, the modification atlas of regulation by chromatin states (MARCS), available at https://marcs.helmholtz-muenchen.de. The library of semi-synthetic di-nucleosomes incorporates biologically meaningful combinations of chromatin modifications representing promoter, enhancer, and heterochromatin modification states. Each affinity purification measures the relative abundances of nuclear proteins on a modified nucleosome in relation to an unmodified control nucleosome using the SILAC labeling and quantitative proteomics as a readout. This allows the high-throughput identification of proteins that are either recruited or excluded by the modification(s) and also indicates the extent of the recruitment or exclusion. Collectively, the MARCS data set catalogs the individual binding responses of 1915 nuclear proteins to nucleosomes carrying 55 different modification signatures. The study uses computational analysis methods to understand how chromatin states are read and interpreted by nuclear machineries.

- Contribution [1] in Appendix A.1: *asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers*
  The genetic material of eukaryotic cells is stored in the nucleus in the form of chromatin, a nucleo-protein complex consisting primarily of DNA and histone proteins.

DNA and histones carry chemical modifications that regulate chromatin function. These chromatin modifications recruit epigenetic "reader" proteins, which mediate processes such as DNA transcription, replication, and repair [15]. Most chromatin modifications occur in distinctive combinations within a nucleosome, suggesting that epigenetic information can be encoded in combinatorial chromatin modifications. The idea that combinations of histone modifications may form a "histone code" that together with DNA modifications could store epigenetic information in the chromatin template, thereby expanding the genetic information encoded in the DNA sequence, has been around for over two decades [23, 24, 25]. A detailed understanding of how multiple modifications cooperate in recruiting such proteins has, however, remained largely elusive. This study combines nucleosome affinity purification data with high-throughput quantitative proteomics, the modification atlas of regulation by chromatin states dataset (MARCS) data [4], and hierarchical interaction modeling to estimate combinatorial effects of chromatin modifications on protein recruitment. Specifically, this is achieved by the computational workflow `asteRIa` which combines ideas from hierarchical interaction modeling, stability-based model selection, and replicate consistency checks to provide stable estimation of robust interactions among chromatin modifications. On the MARCS dataset, `asteRIa` identifies several candidate proteins as epigenetic reader candidates that respond to specific interactions between histone modifications beyond mere additivity. The analysis suggests that proteins within the same protein complex tend to exhibit similar binding patterns not only to individual chromatin modifications but also with regards to interaction effects of modifications. The generalizability of the findings beyond a specific cell type or experimental setup is demonstrated by comparing the interaction effect of H3K27me3 and methylated DNA on the protein CBX8, as identified by `asteRIa`, using publicly available chromatin immunoprecipitation sequencing (ChIP-seq) and whole genome bisulfite sequencing (WGBS) data from K562, A549, H1, and mES cells. This study offers the first quantitative framework for identifying cooperative effects of chromatin modifications on protein binding. Furthermore, the `asteRIa` workflow is of general interest for estimating biological combinatorial interactions in contexts characterized by data scarcity and unbalanced design regimes.

- Contribution [2] in Appendix A.2: *Predictive modeling of microbial data with interaction effects*
  Microbial interactions are crucial in determining the structure and function of microbial communities [61]. These interactions change dynamically in response to community functions and environmental or host-related conditions [14]. Statistical tools that can derive robust and interpretable interaction effects between microbial taxa are crucial for unraveling microbial interactions. Recent studies have shown that quadratic interaction modeling of microbial data can accurately describe community functions [32] and host fitness [14]. However, the focus of these modeling ap-

proaches has primarily been on predictive accuracy rather than the stable detection of interaction effects. Furthermore, interaction modeling for compositional input data, which is often encountered in microbial studies, is not yet well-established. This study introduces a generic sparse quadratic interaction model designed for predicting environmental or host-related conditions. The model accommodates distinct data types commonly found in microbial abundance information ranging from quantitative microbiome, also known as absolute counts, over presence-absence information to compositional data. To achieve stable and interpretable interaction estimation, the interaction modeling framework comes with extensions to hierarchical interactions and stability-based model selection. The framework's versatility is demonstrated by its application to microbial datasets across various ecosystems encompassing all data modalities. Both simulated and real data show how quadratic models enhance prediction, identify known effects, and reveal previously unknown interactions. Notably, the study identifies sparse interaction models that accurately predict the abundance of antimicrobial resistance genes, enabling the formulation of novel biological hypotheses about microbial community composition and antimicrobial resistance.

- Contribution [3] in Appendix A.3: *A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach*
  High-content screening (HCS) generates large quantities of cell morphological data (e.g., nucleus size or cell shape) derived from microscopy images under various chemical conditions or drug combinations [62]. The derived morphological features are typically presented as summary statistics across multiple single cells [63]. A key question includes how certain drugs and combinations of drugs influence the morphology of cells. Not all morphological outcome features may carry relevant information, and some might be redundant. In this study, a combinatorial drug design involving 20 distinct compounds across 408 experiments was employed to analyze cell morphological features within HCS experiments. From this data, a generally applicable computational framework was developed to examine stable drug interaction effects. This framework incorporates a robustified version of the hierarchical interaction modeling workflow developed in [1]. Instead of reducing the space of morphological features before conducting further statistical analyses, as suggested in previous studies [64, 65, 56], the framework presented in this study examines both main and interaction effects across all outcome features and subsequently employs a post-estimation clustering approach. For each of the inferred clusters, prototypical morphological features were statistically determined providing a condensed view of the results. In summary, this contribution introduces a generally applicable computational tool that uncovers combinatorial drug effects on a reduced set of morphological features in HCS studies.

1.4 Summary of individual contributions

- Contribution [4] in Appendix B.1: *Decoding chromatin states by proteomic profiling of nucleosome readers*
  This project is a large effort over multiple years that combines a multidimensional proteomics strategy to systematically examine the interaction of nuclear proteins with modified dinucleosomes, along with computational tools to analyze and visualize the nucleosome-binding data. I joined this project in May 2020 to perform more sophisticated statistical analyses on the data, particularly focusing on the combinatorial effects between chromatin modifications on the binding behavior of proteins. Initially, I analyzed the data with state-of-the-art statistical tools, thereby confirming the already existing results. My core contribution to this publication was performing a statistical post-estimation clustering analysis on the identified feature effects and computationally defining a condensed version of the data by assigning "prototypical" proteins that describe the clusters. I was responsible for creating the figures for this analysis and contributed to the methods part of this publication. As the combinatorial analyses required the development of an entirely novel statistical workflow, they were analyzed and presented within a separate publication (see manuscript [1]).

- Contribution [1] in Appendix A.1: *asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers*
  As part of an interdisciplinary doctoral project, Dr. Till Bartke and Prof. Dr. Christian L. Müller envisioned combining existing and developing novel computational tools to uncover combinatorial chromatin modifications that affect the binding behavior of proteins by using the MARCS dataset. Notably, the MARCS dataset is unique and represents a major leap forward in the field of epigenetics, which justifies the development of a specialized statistical framework to support its analysis. With methodological input from Prof. Dr. Christian L. Müller and domain-expert knowledge in the results provided by Dr. Till Bartke, I developed a computational framework for the stable detection of robust interactions between chromatin modifications, named `asteRIa`. Additionally, I generated simulated data to evaluate the performance of the statistical models. To validate the interaction effects I detected with `asteRIa`, I collected available chromatin immunoprecipitation sequencing (ChIP-seq) and whole genome bisulfite sequencing (WGBS) data and came up with the idea of how to analyze the data such that it serves as a valid confirmation of my results. I also developed ways to visually present the results in an accessible and comprehensive manner and created all the figures myself. I was responsible for the entire computational analysis and for writing the manuscript. Many of the fundamental statistical ideas for my doctoral project were developed during this project and have formed the basis for subsequent work.

- Contribution [2] in Appendix A.2: *Predictive modeling of microbial data with inter-*

action effects
This project combines statistical ideas from [1] and includes a methodological extension of interaction models to compositional data. I was responsible for the entire implementation of the project. With input from Prof. Dr. Christian L. Müller and Prof. Dr. Jacob Bien, I developed the formulation of the statistical model. I implemented and tested the model on both real and simulated data and conducted a feasibility study on quadratic models for compositions. Moreover, I collected data from various fields to demonstrate the versatility of the framework. I conducted the entire computational methodology and formal analysis, wrote the manuscript, and created all the figures for the project.

- Contribution [3] in Appendix A.3: *A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach*
In this project I used modeling ideas from [1] and extended them to enhance the general robustness and with this account for technical and biological noise in the data. I was responsible for the statistical analysis of the data. I created the figures representing the data and the results and I wrote the methods part of the draft manuscript.

1.5 Outline

In the subsequent chapters, instances of the statistical models and workflows used in each individual project are presented in a unified manner and integrated into existing literature. Although I refer to specific biological applications when illustrating concepts in the following sections, the details and results of the biological questions are described in the corresponding publications and manuscripts. In Chapter 2, I provide an overview of existing techniques for interaction modeling. Moreover, I define a generic interaction model that I further instantiate to various data modalities and I describe how to perform penalized model estimation. In the last section of this chapter, I describe how the concept of hierarchical interactions can be integrated with the interaction models presented. In Chapter 3, I discuss and compare model selection techniques in the penalized interaction model. In Chapter 4, I introduce ways of accounting for inconsistent measurements among replicates to mitigate biological and experimental noise. In Chapter 5, I present examples of how to generate realistic synthetic data scenarios that allow assessing the accuracy of the effects derived by the statistical approaches. In Chapter 6, I discuss concepts of visually presenting the results of the statistical interaction analyses in a condensed and informative way.
The contributions [1] and [4] are included in Appendix A.1 and B.1. The manuscript [2] is included in Appendix A.2. Appendix A.3 contains manuscript [3].

2 Statistical interaction modeling

Quadratic interaction models of the form in Eq. 1 have a long tradition in statistics and experimental design [50, 51, 52]. Depending on the structure of the input data, specific considerations and adaptations are necessary. In [2], I introduced a generic model formulation that accommodates different data modalities. Here, these ideas are contextualized more broadly.

In this section, I define the input data comprising p features as a matrix $X \in \mathbb{R}^{n \times p}$ and the outcome as a vector $y \in \mathbb{R}^n$.

2.1 Background

While ordinary least squares (OLS) methods for statistical interaction modeling are applicable in settings with low to moderate numbers of dimensions, they become inadequate in high-dimensional settings where the number of features exceeds the number of samples ($n < p(p-1)/2$). Additionally, even when enough samples are available, OLS is not ideal for creating parsimonious models when large sets of potentially correlated features are available. To induce sparsity in interaction models, ℓ_1 penalization (lasso) can be applied [13]. Incorporating both main and interaction effects in the lasso is often referred to as the *all pairs lasso* or *sparse quadratic model.* This approach, however, does not differentiate between main and interaction effects, increasing the chance of selecting an interaction effect due to the substantially larger number of interaction features ($p(p-1)/2$) compared to main effects. A popular way to enhance model interpretability involves the incorporation of the statistical principle of hierarchy, also referred to as marginality or heredity [50, 66, 52, 67]. This principle allows the presence of an interaction effect $\hat{\Theta}_{jk}$ in the model in Eq. 1 only if either both of the corresponding main effects $\hat{\beta}_j$ and $\hat{\beta}_k$ are included in the model, which is known as *strong hierarchy*

$$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0 \text{ and } \hat{\beta}_k \neq 0,$$

or if at least one main effect is included in the model, known as *weak hierarchy*

$$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0 \text{ or } \hat{\beta}_k \neq 0.$$

While the principle of hierarchy may initially seem like a strong structural assumption, arguments have been made that models that do not adhere to a strong hierarchy are impractical [68, 69] and that it is more likely that large main effect features are also involved in interaction effects, simply due to statistical power [70]. Another argument for hierarchy is its practical aspect of re-using features [69] in order to reduce the number of experiments required to confirm a result. For a specific example, assume a hierarchical model with strong hierarchy that identifies relevant effects of drug A, drug B, and their combination effect AB. In contrast, assume a non-hierarchical model would identify effects of drug A, drug B, as well as a combination effect for drugs C and D (CD). Consequently, the experimentalist would need to perform three experiments in the first scenario (A, B, AB) and five experiments in the second scenario (A, B, C, D, CD) to verify the results. Indeed, the concept of incorporating hierarchy in penalized quadratic interaction models has gained considerable popularity and the field has evolved to address the complexity of modern datasets. Numerous efforts have been made to integrate the hierarchy assumption within multi-stage procedures, which primarily select main effects in the initial stage [71, 72, 73, 74], as well as through a single optimization problem [69, 75, 76]. For a comprehensive list of references on this topic, see [77].

2.2 Interaction models for different data modalities

This section is based on, and partly identical to, manuscript [2].

Given a set of p features $X = (X_1, \dots, X_p)$, which could represent various biological attributes such as chromatin modifications, microbial species abundances, or components of combinatorial drug designs, I consider their impact on an outcome variable $y \in \mathbb{R}^n$ (e.g., protein binding, microbial community function, cellular features). The simplest model to uncover the additive effects of these features on y is a linear or main effect model

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon,$$

where β_0 is the intercept term, each β_j represents the effect of the j-th feature on y, and ϵ models the technical and biological noise term.

For many predictive tasks in biological research, a simple linear model, which only considers main effects, may be insufficient to capture the complex dynamics of biological systems, where interactions between components can significantly influence the outcome. To account for these interactions and introduce greater model complexity while maintaining interpretability, I incorporate quadratic terms into the main effect model, leading to the generic quadratic interaction model

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j,k} \Theta_{jk} X_j X_k + \epsilon, \tag{1}$$

where $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$ is a symmetric matrix of interactions. Depending on the context, it can be meaningful to set $\Theta_{jj} = 0$, which changes the interaction term in Equation 1 to $\frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k$.

To ensure the model's applicability across different types of biological data, I further adapt and instantiate this interaction model to effectively handle:

(i) **Quantitative data:** Where $X \in \mathbb{R}^{n \times p}$. This could for example represent gene expression data or quantitative microbiome data.

(ii) **Binary presence-absence data:** Where $X \in \{0,1\}^{n \times p}$ or $X \in \{-1,1\}^{n \times p}$ indicates the presence or absence of a feature (e.g., presence of a specific drug or gene)

(iii) **Compositional data:** Where the count information of each sample X_i for $i = 1, ..., n$ represents a relative proportion. Mathematically, the features in $x := X_i$ are defined within a simplex Δ^{p-1}, which is defined as

$$\Delta^{p-1} = \left\{ x \in \mathbb{R}^p : x_j \geq 0 \text{ for all } j = 1, ..., p \text{ and } \sum_{j=1}^{p} x_j = 1 \right\}.$$

This formulation ensures that each entry is non-negative and the sum of all components equals one, reflecting the inherent constraints of compositional data. Compositional data are commonly found in microbial studies, arising in high-throughput sequencing experiments where the total data count depends on the capacity of the instrument [78]. Another popular example where compositionality plays a role is in geology. Here, the chemical composition of rock or soil samples is often expressed as proportions of different elements, reflecting the compositional nature of these data [79].

2.2.1 Interaction model for quantitative data

Whenever biological data is given as quantitative data, the forward model corresponds to the model in Eq. 1 and does not require further transformation of the input data X or any constraints on the model coefficients. Throughout this work, we denote the quantitative input data by $A \in \mathbb{R}^{n \times p}$ (often: $A \in \mathbb{R}_{+}^{n \times p}$). Assuming that y depends on the actual amounts of input features, the quadratic interaction model is given by

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j A_j + \frac{1}{2} \sum_{j,k} \Theta_{jk} A_j A_k + \epsilon, \tag{2}$$

where the model parameters follow the description provided in Eq. 1.

2.2.2 Interaction model for presence-absence data

If the information is represented as presence-absence data, given by a binary matrix, I denote the input data as $B \in \{0,1\}^{n \times p}$, where 1 indicates the presence of a feature, and 0 indicates its absence. One common alternative encoding is $B \in \{-1,1\}^{n \times p}$, where the absence is encoded as -1. While the choice of encoding does not affect my model's ability to fit the data, it changes the interpretation of the coefficients.

Assuming that y depends on the presence-absence information of features, the quadratic interaction model is given by

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j B_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} B_j B_k + \epsilon, \tag{3}$$

where the model parameters follow the description provided in the model in Eq. 1. In this model, the condition $\Theta_{jj} = 0$ is particularly meaningful, as the interaction directly corresponds to the main feature, $B_j B_j = B_j$ for $B \in \{0, 1\}^{n \times p}$, or is constant, $B_j B_j = 1$, for $B \in \{-1, 1\}^{n \times p}$. For $B \in \{0, 1\}^{n \times p}$, β_0 is the baseline effect when all features are absent, and β_j for $j = 1, \ldots, p$ represents the effect of the presence of B_j when all other features are absent. The interaction term, Θ_{jk}, accounts for the additional effect when both features B_j and B_k are present. For $B \in \{-1, 1\}^{n \times p}$, β_0 signifies the mean over all group means. Under a completely balanced design, this is equal to the overall mean. For more details on the interpretation of the model coefficients between the encodings, see [2]. When y is characterized as a fitness or phenotypic landscape [32, 30], the encoding $B \in \{-1, 1\}^{n \times p}$ is often preferred. This encoding corresponds to the Fourier expansion, which enables the parameters to describe key landscape properties, such as ruggedness [32, 31, 33].

Transformation between binary encodings There exists a linear transformation between the coefficients of both encodings. I denote all coefficients in the 0 and 1 encoding as $\tilde{\beta}$ and $\tilde{\Theta}$, respectively, and the coefficients in the -1 and 1 encoding as β and Θ, respectively. The transformation between both encodings in the quadratic interaction model is given by the following equation system

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 - \sum_{j=1}^{p} \beta_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \Theta_{jk}, \\
\tilde{\beta}_j &= 2\beta_j - 2 \sum_{k=1, k \neq j}^{p} \Theta_{jk}, \quad \text{for } j = 1, \ldots, p, \\
\tilde{\Theta}_{jk} &= 4\Theta_{jk}, \quad \text{for } j = 1, \ldots, p-1, \text{ and } k = j+1, \ldots, p.
\end{aligned}$$

This transformation from one encoding to the other can be derived by replacing the input matrix in the model with $B^{\{-1,1\}} = 2B^{\{0,1\}} - 1$.

2.2.3 Interaction modeling for compositional data

When the information in X is provided as sparse compositions rather than quantitative data, one approach to modeling interaction effects between features involves converting X to a binary matrix $B = 1_{\{X>0\}}$ that carries the presence-absence information of the features. However, the compositional information might offer valuable insights that presence-absence data do not capture, as discussed and illustrated on real data in [2].

Interaction modeling for compositional data has been described by Aitchison and Bacon-Shone [80], but it is not yet well-established for practical applications or under ℓ_1 penalization. For modeling main effects with compositional input data, a variety of approaches exists, ranging from theoretical concepts in low- and high-dimensional settings [80, 81, 82, 83] to practical implementations [84, 85].

Here, I introduce three approaches for modeling quadratic interactions with compositional input data, building upon existing main effect models:

(a) the additive log-ratio (alr) transformed quadratic model,

(b) the quadratic log-contrast model, and

(c) the quadratic log-ratio model.

These three models differ in terms of interpretability, dimensionality, and optimization, and the choice of model may depend on the dimensionality of the underlying data and the specific biological question being addressed.

There exist certain properties of main effect models for compositional input data that are convenient for practical applications, such as scale invariance and subcompositional coherence [86]—the latter ensuring that the relationships identified remain consistent, irrespective of whether the entire dataset or a subset of it is analyzed. These properties can be directly translated to the interaction models I describe here. Additionally, the considerations for compositional input data in practical applications, such as exact zeros in the data that are typically replaced by pseudo counts [87], can be adopted in the quadratic extensions presented here.

Interaction model (a): alr transformed quadratic model

While comparing the relative count data between different samples might not be biologically meaningful, describing the response as a linear combination of log-ratios derived from the original compositions provides a valid basis for comparison. One popular method for constructing log-ratios is the additive log-ratio (alr) transformation, which involves choosing a common reference feature, denoted here as the p-th feature. The transformed count for each feature j is then given by $C_j = \log\left(\frac{A_j}{A_p}\right)$, for $j = 1, \ldots, p-1$ [80].

The alr transformation allows the modeling of an outcome y based on the $(p-1)$-dimensional compositional input data. This approach assumes that y depends on the composition of X, not on the actual amounts. The main effect model on the transformed features is given by

$$y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \epsilon, \tag{4a}$$

where β_j are coefficients estimating the effect of each log-ratio C_j on y, and ϵ represents the error term.

An extension to include interaction effects was defined by Aitchison and Bacon-Shone [80]. It is described here as the *alr transformed quadratic model* and is given by

$$y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j,k=1}^{p-1} \Theta_{jk} C_j C_k + \epsilon, \tag{4b}$$

where Θ_{jk} represents the interaction coefficients. This formulation aligns with the model in Eq. 1 (when setting $p := p-1$). In the initial definition of the model in [80], the matrix Θ is not necessarily symmetric.

This model allows an interpretation of the effects with respect to a specific reference feature p and is straightforward for parameter estimation and optimization. However, adaptations of this model that enable interpretation without the need to define a reference feature have been proposed [80, 81]. Both approaches can be translated back to the model formulations in Eq. 4a and Eq. 4b, respectively, and will be discussed in the subsequent paragraphs.

Interaction model (b): Constrained quadratic log-contrast model

As shown in [80], a more convenient symmetric expression of the linear alr transformed model in Eq. 4a, which does not require a reference feature and therefore has a better interpretation, can be derived by reformulating the equation as a p-dimensional problem including a zero-sum constraint. This is given by

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \epsilon, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0, \tag{5a}$$

where the main (log) effect coefficients $\beta_j, j = 1, \ldots, p$, sum up to zero. The corresponding extension to the *quadratic log-contrast model* has also been proposed in [80] and is given as

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log\left(\frac{A_j}{A_k}\right)^2 + \epsilon, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0, \tag{5b}$$

where the main (log) effect coefficients $\beta_j, j = 1, \ldots, p$, sum up to zero, with $\beta \in \mathbb{R}^p$, and the interaction effect coefficients Θ_{jk} correspond to the quadratic (log-ratio) interaction effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$. In other words, the main effects are interpreted relative to all main effects, while the interaction effects represent the quadratic effects arising from pairwise comparisons. Note that assuming $\Theta_{jj} = 0$ is particularly meaningful, as $\log\left(\frac{A_j}{A_j}\right) = 0$.

Linear transformation between alr and log-contrast parameters As outlined in [80], there exists a linear transformation between the interaction parameters in the alr transformed quadratic model in Eq. 4b, Θ^{alr}_{jk}, and in the constrained quadratic log-contrast model in Eq. 5b, Θ^{lc}_{jk}:

$$\begin{aligned}
\Theta^{\text{lc}}_{jk} &= -\frac{1}{2}\Theta^{\text{alr}}_{jk}, \text{ for } j = 1, ..., p-1, \; k = j+1, ..., p-1, \\
\Theta^{\text{lc}}_{jp} &= \Theta^{\text{alr}}_{jj} + \frac{1}{2}\sum_{k<j}\Theta^{\text{alr}}_{kj} + \frac{1}{2}\sum_{k>j}\Theta^{\text{alr}}_{jk}, \text{ for } j = 1, ..., p-1.
\end{aligned}$$

For a complete derivation of how to arrive at the constrained quadratic log-contrast model from the alr transformed quadratic model, see manuscript [2].

Interaction model (c): Quadratic log-ratio model

Another way I account for compositionality in regression models is to build log-ratios between all possible pairs of features in $A \in \mathbb{R}^{n\times p}_{+}$. This approach is referred to as the (all-pairs) log-ratio model [81], which is given by

$$y = \beta_0 + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k} \log\left(\frac{A_j}{A_k}\right) + \epsilon, \tag{6a}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k. Rather than modeling each main effect with respect to all features or a reference feature, this model employs pairwise log-ratios to describe the main effects. To incorporate interaction effects, I adopt the approach detailed in Eq. 5b, such that both main and interaction effects are represented as pairwise log-ratios. I have named this the *quadratic log-ratio interaction model*, which is defined as

$$y = \beta_0 + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k} \log\left(\frac{A_j}{A_k}\right) + \frac{1}{2}\sum_{j\neq k}\Theta_{jk} \log\left(\frac{A_j}{A_k}\right)^2 + \epsilon, \tag{6b}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k, with $\beta \in \mathbb{R}^{p(p-1)/2}$ and the interaction effect coefficient Θ_{jk} corresponds to the quadratic (log-ratio) effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p\times p}$.

Linear transformation between log-ratio and log-contrast parameters There exists a linear transformation between the main effect coefficients β_j in the log-contrast model in Eq. 5a and the main effect coefficients $\beta_{j,k}$ in the log-ratio model in Eq. 6a, namely

$$\beta_j = -\sum_{k=1}^{j-1}\beta_{k,j} + \sum_{k=j+1}^{p}\beta_{j,k},$$

implying that the zero-sum constraint on $\beta \in \mathbb{R}^p$ is inherently met in the linear and quadratic log-ratio model.

2.3 Penalized model estimation

To derive parsimonious models, regularized maximum-likelihood estimation incorporating ℓ_1 penalization (lasso) for both linear and interaction coefficients is employed, as proposed in [13]. I introduce a generic optimization problem, consisting of an objective function $\rho(l, \beta_0, \beta, \Theta)$ and a (potential) constraint set $c(\beta_0, \beta, \Theta)$ on the model parameters that allows parameter estimation for all (linear and interaction) models introduced in Section 2.2. The objective function takes the general form

$$\rho(l, \beta_0, \beta, \Theta) = l(\beta_0, \beta, \Theta) + \lambda \left\|\beta\right\|_1 + \frac{\lambda}{2} \left\|\Theta\right\|_1 . \tag{7}$$

Here, $\lambda > 0$ serves as a tuning parameter, regulating the sparsity levels of the coefficients β and Θ, respectively. The loss function $l(\beta_0, \beta, \Theta)$ is specific to each model. Consequently, the generic optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{8}$$

This optimization problem is subsequently instantiated by specific loss functions and constraints.

Sparse quadratic interaction model for quantitative and presence-absence data

The loss function $l(\beta_0, \beta, \Theta)$ for the sparse quadratic interaction model with absolute count input data or presence-absence input data, as introduced in Eqs. 2 and 3, is defined by

$$l^{\text{qi}}(\beta_0, \beta, \Theta) = \left\| y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j,k} \Theta_{jk} X_j X_k \right\|_2^2 ,$$

with $X := A \in \mathbb{R}_+^{n \times p}$ for absolute count data and $X := B \in \{0,1\}^{n \times p}$ (or $B \in \{-1,1\}^{n \times p}$) for presence-absence data (and potentially $\Theta_{jj} = 0$). This model does not require further constraints on the model parameters, so that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is formulated as

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\text{qi}}, \beta_0, \beta, \Theta). \tag{9}$$

In the linear model case, the loss function in the optimization problem simplifies to $l(\beta_0, \beta) = \left\| y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right\|_2^2$.

Sparse alr transformed quadratic model

Given $A \in \mathbb{R}^{n \times p}$ as a matrix containing the relative abundance information of p microbial taxa, the loss function in Eq. 7 for the sparse alr transformed quadratic model, as introduced in Eq. 4b, is defined as

$$l^{\text{qalr}}(\beta_0, \beta, \Theta) = \left\| y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j,k}^{p-1} \Theta_{jk} C_j C_k \right\|_2^2 ,$$

with $C_j = \log\left(\frac{A_j}{A_p}\right)$, for $j = 1, \ldots, p-1$. The model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is formulated as

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\text{qalr}}, \beta_0, \beta, \Theta). \tag{10}$$

In the main effect model case, the loss function in the optimization problem reduces to $l^{\text{alr}}(\beta_0, \beta) = \left\| y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j \right\|_2^2$.

Sparse quadratic log-contrast model

The linear log-contrast model has been extended to the high-dimensional setting, where it is also known as the *sparse log-contrast model* [82, 88, 84, 85] and is supported by software implementations [85]. Yet, the quadratic log-contrast model has not been adapted for high-dimensional settings, nor has it been widely implemented in practical applications. Here, I translate the interaction model proposed in [80] to the high-dimensional setting. The loss function for the sparse quadratic log-contrast model (qlc), corresponding to the interaction model for compositional data introduced in Eq. 5b, is defined as

$$l^{\text{qlc}}(\beta_0, \beta, \Theta) = \left\| y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) - \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log\left(\frac{A_j}{A_k}\right)^2 \right\|_2^2 .$$

As this model incorporates a zero-sum constraint on the main effect coefficients, the constraint set in Eq. 8 is given by

$$c(\beta_0, \beta, \Theta) = \left\{ \sum_{j=1}^{p} \beta_j = 0 \right\}.$$

Thus, the optimization problem for the sparse quadratic log-contrast model is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\text{qlc}}, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{11}$$

In the linear sparse log-contrast model [82], the loss function reduces to $l^{\text{lc}}(\beta_0, \beta) = \left\| y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) \right\|_2^2$, while the constraint on the main effects is maintained.

Scaling of interaction features The main effect covariates, denoted by A_j for $j = 1, \ldots, p$ typically remain unscaled under the zero-sum constraint. However, the interaction features $\log\left(\frac{A_j}{A_k}\right)^2$ for $j = 1, ..., p-1$ and $k = j+1, ..., p$ are not subject to the zero-sum constraint, and can be scaled. The ℓ_2-norm of these interaction features tends to increase with the ℓ_2-norm of their associated main effects. More specifically, the ℓ_2-norm of the main effects after transforming them with the centered log-ratio (clr) transformation is considered. The clr divides each compositional part by the geometric mean of all parts, namely

$$\operatorname{clr}(A) = \left(\log \frac{A_i}{g(A_i)}\right)_{i=1,\ldots,n} \quad \text{with } g(A_i) = \exp\left(\frac{1}{p}\sum_{j=1}^{p} \log(A_{ij})\right).$$

Here, I introduce a scaling that ensures equal penalization of the interaction features. Moreover, I adjust the scale of the interaction features to align with the norm of the average clr transformed ℓ_2-norms of all main effects. Mathematically, this can be expressed as follows: I denote each column of the interaction feature matrix as $A^I_{\cdot jk} = \log\left(\frac{A_j}{A_k}\right)^2$, with $A^I \in \mathbb{R}^{n \times p(p-1)/2}$, and the scaled version is given by

$$\tilde{A}^I_{\cdot jk} = A^I_{\cdot jk}\left(\left\|A^I_{\cdot jk}\right\|_2\right)^{-1} \frac{1}{p}\sum_{k=1}^{p} \left\|A_k^{\text{clr}}\right\|_2, \quad \text{for } j = 1, ..., p-1, \ k = j+1, ..., p,$$

where $A^{\text{clr}} = \operatorname{clr}(A) \in \mathbb{R}^{n \times p}$ is the clr transformed main effects matrix A and $\left\|A_k^{\text{clr}}\right\|_2$ is the ℓ_2-norm of the k-th column of A^{clr}.

Sparse quadratic log-ratio model

The loss function of the sparse quadratic log-ratio (qlr) model corresponding to the interaction model for compositional data, introduced in Eq. 6b, is defined as

$$l^{\text{qlr}}(\beta_0, \beta, \Theta) = \frac{1}{2}\left\| y - \beta_0 - \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k} \log\left(\frac{A_j}{A_k}\right) - \frac{1}{2}\sum_{j \neq k} \Theta_{jk} \log\left(\frac{A_j}{A_k}\right)^2 \right\|_2^2.$$

This model does not require further constraints on the model parameters, so $c(\beta_0, \beta, \Theta) = \emptyset$. The optimization problem for the sparse quadratic log-ratio model is therefore given as

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\text{qlr}}, \beta_0, \beta, \Theta). \tag{12}$$

In the sparse log-ratio model, which is linear in the features and corresponds to the model in Eq. 6a, the loss function reduces to $l^{\text{lr}}(\beta_0, \beta) = \frac{1}{2}\left\| y - \beta_0 - \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k} \log\left(\frac{A_j}{A_k}\right)\right\|_2^2$. The $p(p-1)/2$-dimensional sparse log-ratio problem, labeled as Eq. 6a, is equivalent to the sparse log-contrast model problem for $\lambda^{\text{qlr}} = 2\lambda^{\text{qlc}}$ [81]. This equality can be directly translated to the quadratic extensions of these models. As the dimensionality of the predictor space in the $p(p-1)/2$-dimensional log-ratio model becomes computationally inefficient for large p, the authors in [81] propose a two-stage procedure that involves a pre-selection step for covariates to reduce the predictor space before applying the log-ratio lasso. This two-step procedure can be directly applied to the $2 \cdot p(p-1)/2$-dimensional quadratic log-ratio lasso, introduced here, in scenarios where p is large.

2.4 Hierarchical interaction estimation

The quadratic interaction models previously introduced can potentially enhance predictive performance compared to models that are linear in the features. However, they might not consistently identify stable interaction effects that are critical for further functional analysis. To improve model interpretability, I incorporate the ideas of hierarchical interaction modeling. Specifically, I focus on the single optimization approach introduced in [69]. This section is based on [69].
The concept of hierarchical interactions allows the inclusion of an interaction term Θ_{jk} in the model *only if* both associated main effects are present (strong hierarchy) or at least one of the main effects is included (weak hierarchy). While this assumption may be valid in various biological contexts—for example, two drugs may only interact if they each have individual effects—it proves particularly useful in practical applications as it introduces the concept of "practical sparsity" by re-utilizing features. This is especially relevant when experimental validation of results is necessary, as described in more detail in Section 2.1. Moreover, this approach allows a strong interaction to "pull" itself into the model, ensuring that it cannot be missed, even if it violates the hierarchy assumption.
The hierarchical constraint on the columns (or also rows, if symmetrical) of the interaction effect matrix $\Theta_{\cdot j}$, for each $j = 1, \ldots, p$, can be introduced to the optimization problem by including a constraint set

$$c(\beta_0, \beta, \Theta) = \left\{\Theta = \Theta^T, \|\Theta_{\cdot j}\|_1 \leq |\beta_j|\right\}. \tag{13}$$

Eliminating the symmetry constraint on $\Theta \in \mathbb{R}^{p \times p}$ allows the model to adopt a weak hierarchy among the interaction features. Due to the non-convex nature of Eq. 13, I have adopted a convex relaxation approach as proposed by [69], which is efficiently implemented in the R package `hierNet` [89]. This strategy facilitates handling high-dimensional data by imposing the hierarchical constraint within the generic optimization framework described in Eq. 8, suitable for (i) quantitative data, (ii) presence-absence information, and (iii) relative data processed through alr transformation (Eq. 10). Note, that a direct application of the hierarchical constraint in Eq. 13 might not be meaningful for the models for compositional data in Eqs. 11 and 12.

Application in microbiome data (part I): Deriving a parsimonious quadratic model through hierarchy

In Fig. 4, I illustrate an example from manuscript [2] where I compared the estimated coefficients derived from the sparse quadratic interaction model in Eq. 9 with weak hierarchy (see Eq. 13 without a symmetry constraint on Θ) and without hierarchical constraint. Specifically, the models were applied to a presence-absence microbiome dataset, $B \in \{0,1\}^{n\times p}$, which includes combinations of $p = 25$ bacterial populations across $n = 1561$ experiments, in which the production of butyrate, $Y \in \mathbb{R}_+^n$, a short-chain fatty acid that is beneficial to human health, was measured. The comparison shows that the model incorporating hierarchy yields a substantially reduced set of selected effects, while maintaining good predictive performance, achieving an R^2 of 0.72 with weak hierarchy versus 0.78 without hierarchy on a test set.

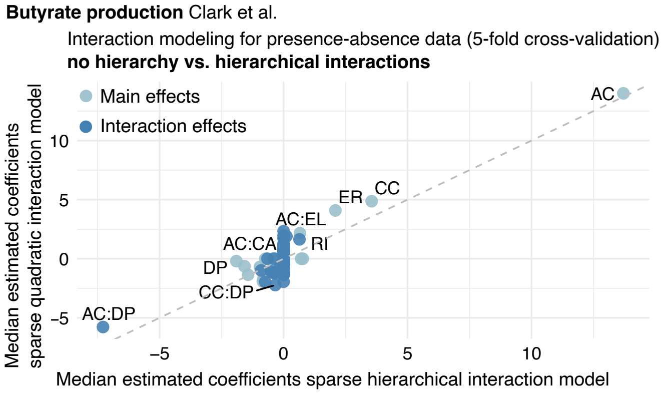


Figure 4: Comparison of median estimated coefficients (over 10 train test splits) between the sparse quadratic interaction model (y-axis) and the sparse hierarchical interaction model with weak hierarchy (x-axis). Labels containing ":" denote interactions (dark blue dots), while those without it represent main effects (light blue dots). Many of the very small coefficients in the model without hierarchy are exactly zero in the model with weak hierarchy. For more details and the full feature names, see [2]. Adapted from [2].

3 Model selection

One of the key challenges in the realm of penalized regression involves determining the optimal regularization parameter, λ, to balance the sparsity (i.e., interpretability) of the model's coefficients with the model's out-of-sample predictive performance [90, 91]. Standard approaches for selecting λ in interaction models include cross-validation [69] and information criteria such as the Akaike information criterion (AIC) and the extended Bayesian information criterion (EBIC) [74].

However, both simulations and practical experiences have shown that cross-validation and information criteria often choose more predictors and interactions than necessary, which can complicate the model [74]. If the main aim lies in detecting reproducible effects, one way of accounting for the potential limitations caused by cross-validation in penalized regression models is the concept of stability selection [92].

This section is based on, and partly identical to, the manuscript [1].

3.1 Cross-validation

This chapter is based on [93], if not stated otherwise. Most implementations of penalized regression models use K-fold cross-validation as the default model selection technique [94, 89, 95] to choose the optimal regularization parameter λ. The principle idea of K-fold cross-validation is to divide the dataset into K subsets, or folds, and iteratively train the model on $K-1$ folds while using the remaining fold for validation. This process is repeated K times, with each fold serving as the validation set once, allowing for the comprehensive assessment of the model's performance. Cross-validation as a model selection approach seeks to find the regularization parameter λ that minimizes the cross-validated prediction error, essentially balancing the trade-off between bias and variance in the model. Specifically, a lasso estimator $\hat{\beta}_{-k}(\lambda)$ for $k = 1, ..., K$ is derived on each of the K train sets for each λ. Typically, the optimal λ is then derived by minimizing the mean-squared cross-validation error. For the main effect model this is given by

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{k=1}^{K} \sum_{i \in I_k} (y_i - X_i^T \hat{\beta}_{-k}(\lambda))^2,$$

where I_k denotes the subset of observations without fold k. Notably, two specific values of λ are often considered: λ_{min}, which minimizes the cross-validation error, and λ_{1se}, which is the largest λ such that the error is within one standard error of the minimum error. The latter can result in a more parsimonious model, potentially with slightly higher bias but better generalization properties due to its simplicity [13].

3.2 Stability selection

Stability selection was first introduced by [92] and has shown effectiveness across various scientific domains, ranging from network learning [96, 97] to data-driven partial differential equation identification [98, 99]. In the context of regression, stability selection involves iteratively learning sparse regression models from subsamples of the data, recording the frequency of selected predictors across models, and selecting the most frequent predictors for the final model. A variant of stability selection, complementary pairs stability selection (CPSS) [100], is particularly advantageous for handling unbalanced experimental designs, as it ensures that individual subsamples contribute equally often. CPSS draws b subsamples as complementary pairs $\{(a_{2l-1}, a_{2l}) : l = 1, ..., b\}$, with $a_{2l-1} \cap a_{2l} = \emptyset$ from samples $\{1, ..., n\}$ of size $\lfloor n/2 \rfloor$. Applying a variable selection procedure S (such as the k first predictors entering the penalized model on the regularization path or cross-validation) to each subsample allows defining a feature-specific selection probability $\hat{\pi}_i$ for $i = 1, ..., p + (p+1)/2$ that is given by

$$\hat{\pi}_i = \frac{1}{2b} \sum_{l=1}^{2b} 1_{\{i \in \hat{S}(a_l)\}}. \tag{14}$$

The final selection set, denoted as $\hat{S}^{\text{CPSS}}$, consists of features for which the estimated selection probability $\hat{\pi}_i$ exceeds a predefined threshold π_{thr}, that represents the minimum selection frequency required for a predictor to be included in the final set. The CPSS approach involves defining several hyperparameters, including the set of regularization parameters Λ, the threshold $\pi_{\text{thr}} \in [0, 1]$, the number of initial predictors k entering the sparse model, and the number of complementary splits b. The CPSS procedure in [101] can be applied to linear and quadratic models and makes no distinction between main and interaction effects. An integration of the CPSS procedure within the hierarchical interaction modeling framework has been introduced in [1].

In [1], I also demonstrate, using a realistic synthetic scenario, how CPSS reduces the number of spuriously detected effects compared to model selection with cross-validation (see [1] and Fig. 9).

Application in microbiome data (part II): Inferring stable estimates through stability selection

Here, I revisit the data example from earlier (see Fig. 4) from manuscript [2]. Employing hierarchical interaction modeling under 5-fold cross-validation (CV) has already facilitated the derivation of a more parsimonious model that explains butyrate production via bacterial species and their interactions. To assess whether all identified effects, including smaller ones, are stable enough to be reported, I used CPSS instead of using 5-fold CV in the interaction model with weak hierarchy. The selection probabilities $\hat{\pi}_i$ for $i = 1, ..., p + p(p-1)/2$ are shown in Fig. 5. Stability selection reveals that there is only one stable interaction effect for predicting butyrate production between *A. caccae* (AC) and *D. piger* (DP) (AC:DP), while other interactions identified by CV fail to demonstrate stability across multiple subsamples, exhibiting very low selection probabilities $\hat{\pi}_i$.

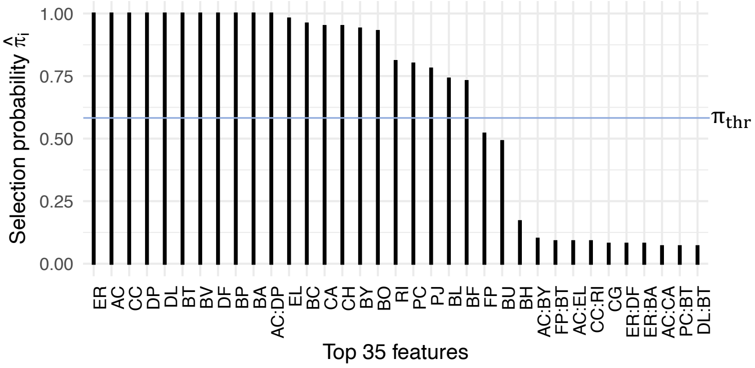


Figure 5: Top 35 selection probabilities from complementary pairs stability selection (CPSS) in the sparse hierarchical interaction model with $\pi_{\mathrm{thr}} = 0.6$. Labels containing ":" denote interactions, while those without it represent main effects. For more details and the full names of features, see [2]. Adapted from [2].

4 Mitigating noise through outlier removal

High-throughput technologies, such as high-throughput sequencing (HTS), has revolutionized biological research by providing an unprecedented resolution of DNA fragments, but comes with the cost of amplified noise. Effectively minimizing this random noise to reveal functionally meaningful biological signals remains a challenge [102, 103]. Mitigating the impact of noise and thereby reducing the chance of identifying spurious interactions is a key aspect of this thesis. In [1], I developed ways to remove outliers by incorporating replicate consistency mechanisms within the interaction modeling framework. In [3], I introduced a way to efficiently learn robust interaction models by being sensitive to outliers in a more general way.

4.1 Replicate consistency with few replicates

This section is based on, and partly identical to, manuscript [1].
Biological datasets often include a small number of replicated measurements to probe different sources of variability in the underlying experimental procedure or study object [104]. Replicate consistency, i.e., assessing how consistent two or multiple replicated measurements are in terms of sign or distance, is an important property to evaluate experimental protocols and downstream analysis quality (see, e.g., [105] for a discussion in the context of RNA sequencing data). Here, I introduce two filtering steps that can be performed when few and inconsistent replicates are available. These filtering steps are major components of the `asteRIa` workflow in the manuscript [1]. Specifically, I propose two replicate consistency mechanisms: (i) data sign consistency and (ii) nested model consistency. While there are alternative ways of performing filtering, data sign consistency can be considered as a data filtering step that ensures that replicated measurements agree on the direction, i.e., the sign of the measured unit, and removes experiments where sign consistency does not hold. A sensitivity analysis of the type of filter in step (i), comparing data sign consistency with various distance-based filters and with no pre-filtering, can be found in [1]. The replicate consistency mechanism (i) is particularly useful when two replicates are available. Assuming that each outcome $y = Y_i$ for $i = 1, ..., p_2$ was measured twice, the first data sign consistency filtering step removes observations of inconsistent signs in their observations (see Fig. 6, Step 1). Although this reduction in sample size (for each $y = Y_i$ $n_i \leq n$ samples are available) decreases the power for subsequent hierarchical interaction modeling, the filtering increases the chance of estimating pairs of consistent interaction models. In a second post-hoc step, nested model consistency further ensures that only pairs of consistent interaction models are considered for downstream analysis. Nested model consistency deems estimated interaction models valid only if they comprise the same set of features (main and interaction coefficients) across replicates *or* one model comprises a nested subset of main and interaction effects of the other model (see Fig. 6, Step 2, for illustration). Depending on the desired level of rigor, these consistency checks can be extended to scenarios with more than two, yet still few, replicates. In cases with

slightly more replicates, one could proceed similarly by removing observations that show inconsistent signs, or by eliminating measurements where the distances between them are large.

For a complete description of how these replicate consistency mechanisms function within the `asteRIa` workflow, see Fig. 6 and [1].

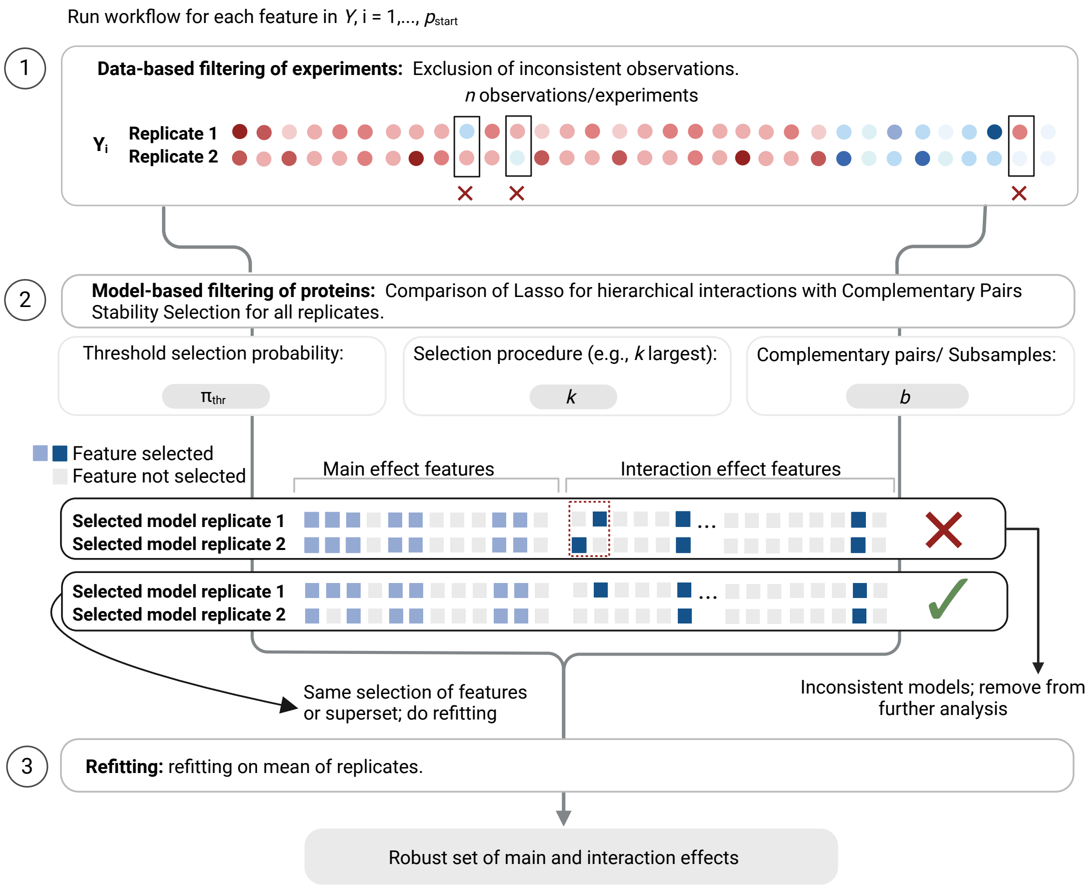

Run workflow for each feature in Y_i i = 1,..., p_{start}
1
Data-based filtering of experiments: Exclusion of inconsistent observations.
n observations/experiments
Y_i
Replicate 1
Replicate 2
2
Model-based filtering of proteins: Comparison of Lasso for hierarchical interactions with Complementary Pairs Stability Selection for all replicates.
Threshold selection probability:
π_{thr}
Selection procedure (e.g., k largest):
k
Complementary pairs/ Subsamples:
b
Feature selected
Feature not selected
Main effect features
Interaction effect features
Selected model replicate 1
Selected model replicate 2
Selected model replicate 1
Selected model replicate 2
Same selection of features or superset; do refitting
Inconsistent models; remove from further analysis
3
Refitting: refitting on mean of replicates.
Robust set of main and interaction effects

Figure 6: Graphical representation of two replicate consistency mechanisms for the stable detection of hierarchical interactions (created with BioRender.com). This is a generalized version of the `asteRIa` workflow introduced in [1]. Step 1: Observed outcome values $y = Y_i$ (two replicates). Removal of observations with different signs in the replicates (sign consistency). Step 2: Interaction modeling approach. Hierarchical interaction modeling with default complementary pairs stability selection (CPSS) parameters (`asteRIa`). Comparison of selected features for each replicate (nested model consistency): The first example shows a prediction model for $y = Y_i$ that gets filtered out since the selected features from the replicates are neither identical nor nested. The second example shows a "consistent" model for $y = Y_i$ where the selected features learned from replicate 1 are a nested subset of the features learned from replicate 2. Step 3: Least-squares refitting on averaged replicate data for final prediction model building. The intersection of two selected feature sets is used for refitting. Adapted from [1].

4.2 Robust learning with Huber loss

Here, I present an approach to enhance robustness toward each outcome $y = Y_i$ for $i = 1, ..., p_2$ of a regression model in a more general manner compared to the previously described replicate consistency procedure. This section is based on, and partly identical to, draft manuscript [3].

The Huber loss function, introduced in [106], a robust alternative to the squared error loss (L2), has been widely adopted in (penalized) regression analysis due to its robustness to outliers [107]. It combines the L2 loss for small residuals with an absolute loss (L1) for larger ones, mitigating the influence of outliers on parameter estimates. The convexity of the Huber loss ensures the existence of a unique minimum, making it a suitable choice for optimization problems in statistical learning.

To enhance the robustness of the results in the interaction modeling framework, I integrate the Huber loss as a robust alternative. Specifically, in [3], I integrate hierarchical interaction modeling with this robust loss function.

The Huber loss function for $r = l(\beta_0, \beta, \Theta)^{1/2}$ is given as

$$l_{\text{Huber}}(\beta_0, \beta, \Theta, \delta) = \begin{cases} \frac{1}{2}r^2 & \text{for } |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise}, \end{cases} \tag{15}$$

where δ is a tuning parameter that determines the point where the loss transitions from quadratic (L2) to linear behavior (L1) [106]. The Huber loss can be incorporated into the interaction modeling strategies by replacing the L2 loss with it within the optimization framework.

Particularly when dealing with multiple replicates, and some are identified as outliers, the concept of replicate consistency checks in section 4.1 becomes less straightforward. In such instances, the Huber loss offers a robust alternative. Note that these models are employed as exploratory tools without further statistical inference. This allows including replicates within the same model, as one can argue that the estimates remain consistent even with correlations among the observations [108].

5 Simulation approaches

The interaction modeling strategies proposed in this thesis are generic and applicable to a wide range of data, which may vary in structural properties such as amounts of experimental noise or data sparsity, influencing the performance of the models. Unlike traditional settings where standard statistical methods are constrained by specific data assumptions, the models introduced in this thesis rely heavily on simulations. These simulations are crucial for providing insights into how effectively these models can identify the correct effects [9].

The aim is to generate a semi-synthetic outcome $Y_{\text{syn}} \in \mathbb{R}^{n \times p_2}$ that can be used for testing how well the underlying statistical approach works. More specifically, Y_{syn} is constructed as a linear combination of the observed input data X, incorporating fixed main and interaction effects, intercepts, and noise terms. Generating synthetic outcomes from real data has the advantage of preserving the distributional properties of the real data when evaluating the models.

Here, I present two specific examples of how I generated realistic synthetic data scenarios in [1] and in [2] to evaluate the performance of the models. Finally, I use the synthetic data to explore the performance for different model selection approaches, varying noise levels, and different levels of sparsity in the features.

Example 1: Synthetic data generation in a large-scale proteomics study

To illustrate the generation of a realistic semi-synthetic data scenario with a multiple outcome comprising p_2 features, I use the modification atlas of regulation by chromatin states (MARCS) data, introduced in [4] and further analyzed in [1]. This paragraph is based on the manuscript [1].

To be consistent with the notation in Eq. 3, the binary input data is denoted as $B \in \{0,1\}^{n \times p_1}$, which consists of a designed library of engineered di-nucleosomes comprising $p_1 = 12$ chromatin modifications analyzed in $n = 33$ different combinations. The outcome, the observed binding profiles of $p_2 = 1915$ proteins, is denoted by $Y = (Y_1, \ldots, Y_{p_2}) \in \mathbb{R}^{n \times p_2}$.

In [1], sparse main effects $\hat{\beta}_i \in \mathbb{R}^{p_1}$ and interaction effects $\hat{\Theta}_i \in \mathbb{R}^{p_1(p_1-1)/2}$ for each of the p_2 features in Y were derived from the interaction modeling strategy for presence-absence data in Eq. 3 (specifically, these are the interactions derived from the `asteRIa` workflow). The base assumption in this simulation setup is that the estimated coefficients of $p_1 + p_1(p_1-1)/1$ features in B and p_2 features in Y follow a joint distribution. Thus, all coefficients were organized into a common coefficient matrix with main and interaction effects $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}] \in \mathbb{R}^{p_1 + p_1(p_1-1)/2 \times p_2}$. Creating a histogram of the non-zero coefficients in $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}]$ suggested that the coefficients jointly follow an asymmetric Laplace distribution [109]. This observation justified the generation of new simulated model coefficients from an asymmetric Laplace distribution, which was fitted to the estimated coefficients. For a comparison of the histogram of the estimated coefficients and the fitted asymmetric

Laplace distribution, see Fig. 7a and b.

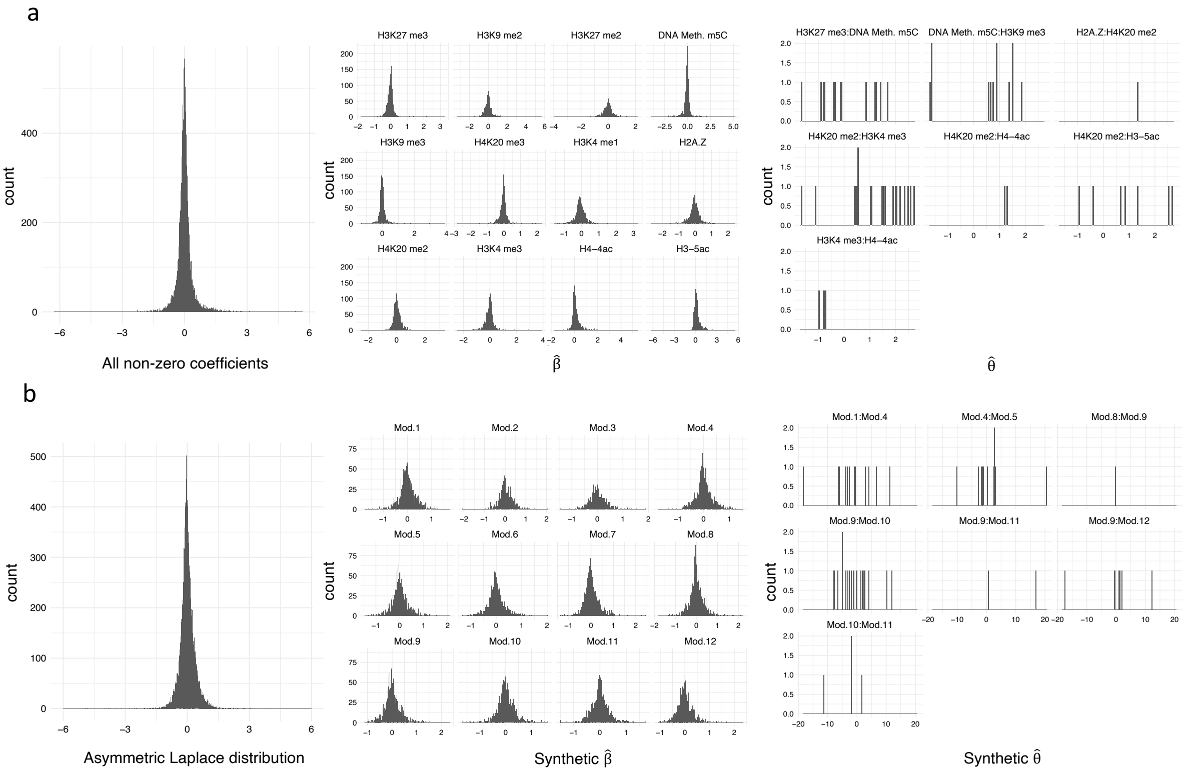


Figure 7: **a**, Joint distribution (left) and feature-wise distributions (center and right) of all non-zero estimated coefficients $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}] \in \mathbb{R}^{p_1+p_1(p_1-1)/2 \times p_2}$ from [1]. Only combinations of two chromatin modifications with at least one non-zero estimate are shown. **b**, Asymmetric Laplace distribution fitted to the estimated coefficients in **a**. Adapted from [1].

In order to maintain the sparsity structure from $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}]$, the same entries in the synthetic coefficient matrix $[\boldsymbol{\beta}, \boldsymbol{\Theta}]_{\text{syn}}$ were set to zero as those in the observed (or estimated) coefficient matrix. Similarly, distributions were fitted to the intercept and error terms to generate synthetic versions $\beta_{0_{\text{syn}}}$ and ϵ_{syn} (see [1] for more details). To vary the signal-to-noise ratios (SNR) in the data, the noise term was multiplied by different constants. Combining the simulated parts and the true experimental design B as a linear combination according to the interaction model in Eq. 3 gives a new outcome $Y_{\text{syn}} \in \mathbb{R}^{n \times p_2}$. This can be written in mathematical terms as

$$Y_{\text{syn},i} = \beta_{0\text{syn}} + \sum_{j=1}^{p_1} \boldsymbol{\beta}_{\text{syn},j} B_j + \frac{1}{2} \sum_{j \neq k} \boldsymbol{\Theta}_{\text{syn},jk} B_j B_k + \epsilon_{\text{syn}}, \text{ for } i = 1, ..., p_2.$$

The joint semi-synthetic outcome $Y_{\text{syn}} = (Y_{\text{syn},1}, ..., Y_{\text{syn},p_2}) \in \mathbb{R}^{n \times p_2}$ recovers the main structures of the observed data $Y = (Y_1, ..., Y_{p_2}) \in \mathbb{R}^{n \times p_2}$ making it an appropriate outcome to test the performance of the interaction model (see Fig. 8 for a visual comparison).

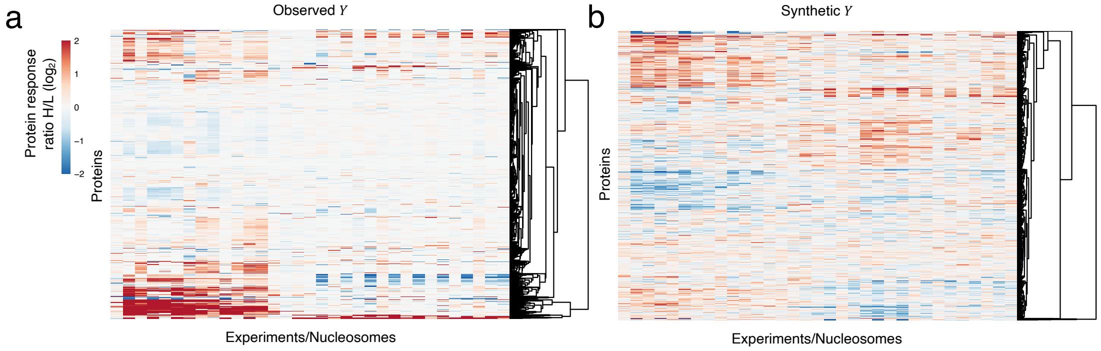


Figure 8: **a**, Clustered heatmap of observed protein binding measures Y. **b**, Clustered heatmap of synthetic protein binding measures Y_{syn}. Adapted from [1].

This newly generated outcome Y_{syn} was used to compare two model selection approaches for the hierarchical interaction model employed in `asteRIa`: 5-fold cross-validation with a regularization parameter $\lambda_{1\text{se}}$ that is the largest within one standard error of the minimum error and complementary pairs stability selection (CPSS). The comparison on semi-synthetic data shows that hierarchical interaction modeling with stability selection greatly reduces the number of spuriously detected effects compared to cross-validation in terms of F1 score and Hamming distance (see Fig. 9).

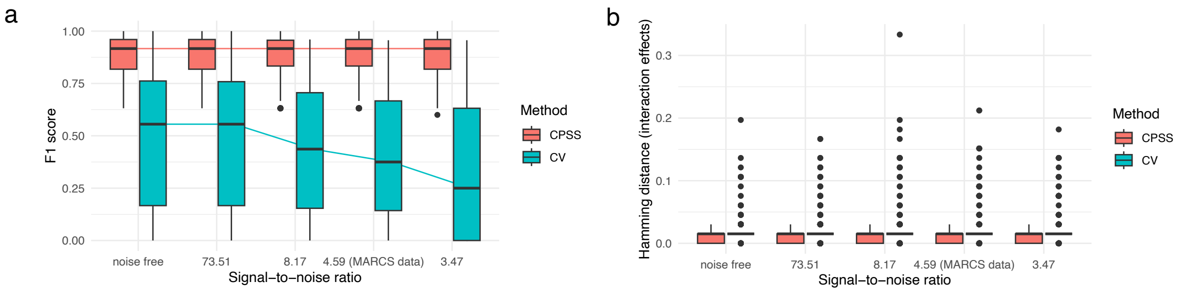


Figure 9: **a**, F1 score for five signal-to-noise ratios (SNR) for the hierachical interaction model with CPSS and 5-fold cross-validation ($\lambda_{1\text{se}}$) over 20 repetitions. **b**, Hamming distance of interaction effects for CPSS and 5-fold cross-validation ($\lambda_{1\text{se}}$) for five SNRs. Adapted from [1].

Example 2: Synthetic data-generation in a compositional microbiome dataset

For a second illustration, I use a compositional microbiome dataset $A \in \mathbb{R}_{+}^{n\times p}$ as input to generate a one-dimensional semi-synthetic outcome $y \in \mathbb{R}^{n}$. This paragraph is based on, and partly identical to, the manuscript [2].

The simulation setup described here uses the model formulation of the quadratic log-contrast model in Eq. 5b. Microbiome data from observational studies derive large numbers of microbial taxa that do not necessarily appear in every sample, leading to large amounts of zeros in the data and complicating statistical analyses [110].

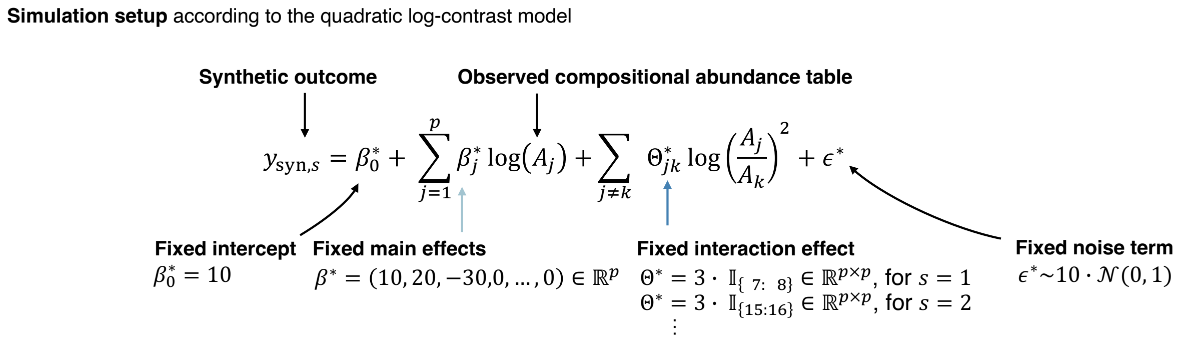

Simulation setup according to the quadratic log-contrast model
Synthetic outcome
Observed compositional abundance table
$y_{\text{syn},s} = \beta_0^* + \sum_{j=1}^{p} \beta_j^* \log(A_j) + \sum_{j \neq k} \Theta_{jk}^* \log\left(\frac{A_j}{A_k}\right)^2 + \epsilon^*$
Fixed intercept
$\beta_0^* = 10$
Fixed main effects
$\beta^* = (10, 20, -30, 0, \ldots, 0) \in \mathbb{R}^p$
Fixed interaction effect
$\Theta^* = 3 \cdot \mathbb{I}_{\{7:8\}} \in \mathbb{R}^{p \times p}$, for $s = 1$
$\Theta^* = 3 \cdot \mathbb{I}_{\{15:16\}} \in \mathbb{R}^{p \times p}$, for $s = 2$
$\vdots$
Fixed noise term
$\epsilon^* \sim 10 \cdot \mathcal{N}(0,1)$

Figure 10: Simulation setup for generating a synthetic outcome $y_{\text{syn},s}$ for $s = 1, ..., S$ based on the quadratic log-contrast model formulation in Eq. 5b. To account for the problem of having zeros in the data when building log-ratios a pseudo count of one is added to each entry in A. Adapted from [2].

Thus, the aim of this simulation was to elucidate the conditions under which accurate parameter estimation is feasible, by varying the degree of sparsity of the interaction features. I used a real-world compositional microbial count data matrix $A \in \mathbb{R}_+^{n \times p}$ (here: $p := p_1$) from the American Gut cohort [111], derived from 16s rRNA sequencing [112], to generate S synthetic single ($p_2 = 1$) outcomes $y_{\text{syn},s} \in \mathbb{R}^n$ for $s = 1, ..., S$. The full simulation setup is shown in Fig. 10 (for more details on the underlying data see Fig. 11 and [2]).

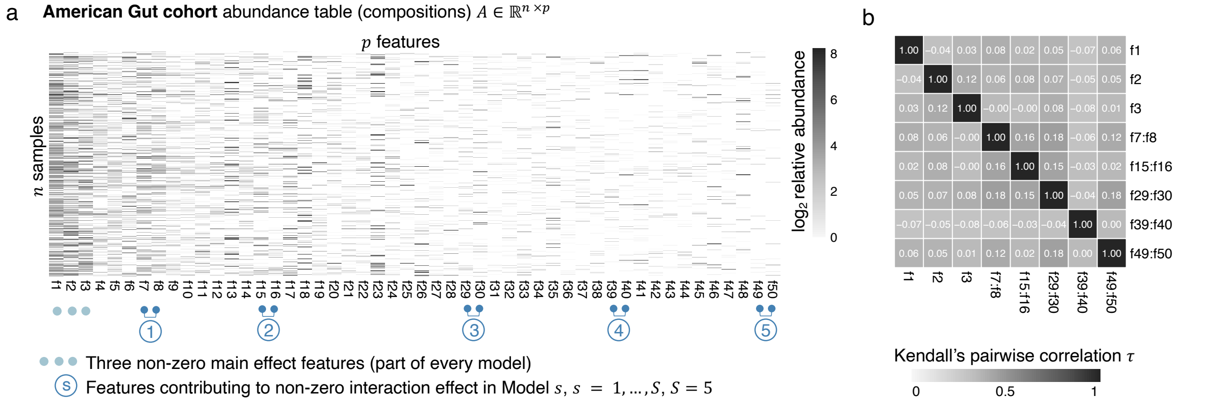


Figure 11: **a**, Heatmap of the microbial abundance table carrying compositional information for a subset of $p = 50$ microbial taxa sorted by sparsity in descending order. Non-zero main effects contributing to each of the $S = 5$ semi-synthetic scenarios (light blue) and features contributing to the non-zero interaction effect in model scenario s for $s = 1, ..., S$ (dark blue) are highlighted. **b**, Kendall's pairwise correlations τ between features that have non-zero effects in the models. Adapted from [2].

In total, I defined $S = 5$ semi-synthetic scenarios $y_{\text{syn},s} \in \mathbb{R}^n$ for $s = 1, ..., S$. To avoid bi-

ases due to correlation effects among predictive features in the model, interaction features with non-zero coefficients were carefully defined to ensure they are uncorrelated with the main effects (with absolute Kendall's pairwise correlation $|\tau| < 0.2$; see Fig. 11b).
By fitting the sparse quadratic log-contrast model to the semi-synthetic data, I showed that as interaction features become sparser, the accuracy of feature estimates decreases (see Tab. 1). This simulation demonstrates the limits of the sparse quadratic log-contrast model in presence of extreme data sparsity. For a more detailed description of the simulation setup see manuscript [2].

Table 1: Median and Variance of the estimation error of the interaction coefficient $\sqrt{(\Theta^*_{jk} - \hat{\Theta}_{jk})^2}$ for $S = 5$ semi-synthetic scenarios.

	Sparsity interaction feature				
	36% (f7:f8)	52% (f15:f16)	67% (f29:f30)	74% (f39:f40)	88% (f49:f50)
Median estimation error	0.15	0.30	0.82	0.87	1.78
Variance estimation error	0.04	0.09	1.15	0.21	1.22

Moreover, I used this simulation setup to illustrate how a misspecified main effect model impacts the estimated coefficients. In other words, the aim was to analyze how accurately the sparse linear log-contrast model from Eq. 5a estimates main effects when there exist true interaction effects that define the outcome y_{syn}. To do so, a fixed intercept term $\beta^*_0 = 10$ was defined, and six non-zero main effects, summing up to zero, were assigned to the first six features as $\beta^* = (10, 20, 30, -10, -20, -30, 0, \dots, 0) \in \mathbb{R}^p$. Additionally, three non-zero interaction effects were introduced (between A_1 and A_3, A_8 and A_{10}, and A_9 and A_{10}) as $\Theta^* = 10 \cdot 1_{\{1:3,\ 8:10,\ 9:10\}} \in \mathbb{R}^{p\times p}$. The noise term ϵ^* was assumed to follow a normal distribution with mean 0 and variance 1. Consequently, the synthetic outcome y_{syn} (with a pseudo count of 1 on A) is given by

$$y_{\text{syn}} = \beta^*_0 + \sum_{j=1}^{p} \beta^*_j \log(A_j) + \sum_{j\neq k} \Theta^*_{jk} \log\left(\frac{A_j}{A_k}\right)^2 + \epsilon^*.$$

Fitting both models, the sparse linear log-contrast model (sparse lc) and the sparse quadratic log-contrast model (sparse qlc), to the newly generated data gives the following results: Notably, features f2, f4, f5, and f6 show no contribution to interaction effects in our synthetic example and are accurately estimated by the (misspecified) sparse lc model (see Fig. 12a and b, first row). However, for the features f1 and f3, which both have non-zero main effects as well as a common interaction effect, the sparse lc model accommodates the positive interaction effect between f1 and f3 within their main effect estimates, leading to an overestimation of the true positive main effect of f1 ($\beta^*_1 = 10$) and an underestimation of the true positive main effect of f3 ($\beta^*_3 = 30$). Moreover, the sparse lc model selects f8, f9, and f10 as relevant main effect features despite their lack of true non-zero main effects, in order to integrate their true underlying interaction effects. In contrast, the sparse qlc model captures the coefficients accurately.

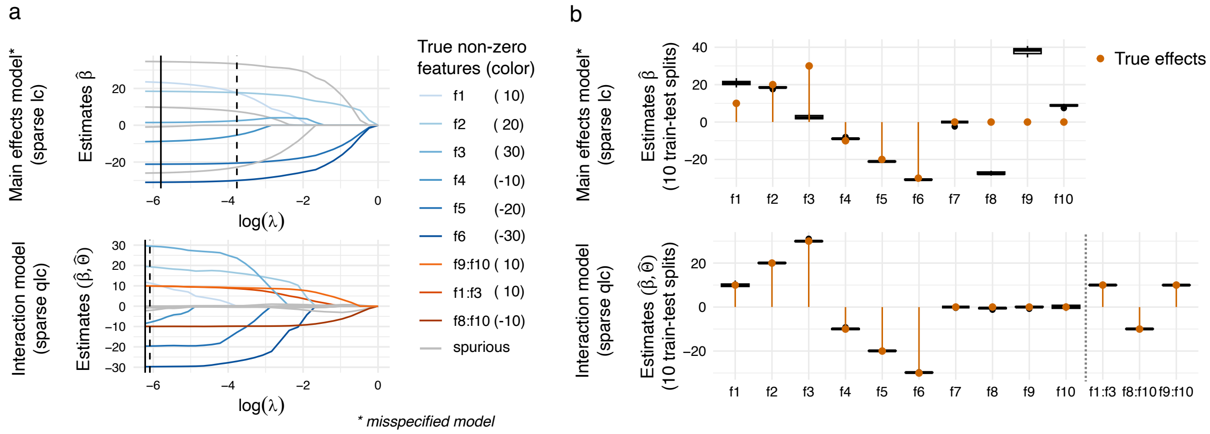


Figure 12: Influence of a misspecified log-contrast (lc) model in a semi-synthetic scenario. **a**, Solution path for the misspecified main effect model (sparse lc) and the interaction model (sparse qlc) for one train test split. **b**, Estimated coefficient distributions over 10 train test splits corresponding to the solution paths in **a**. For the interaction model only three non-zero interaction features are shown for visualization purposes. Adapted from [2].

In summary, the semi-synthetic simulation scenarios for compositional input data demonstrated that very sparse interaction features might not be accurately detected by the model. Furthermore, these scenarios illustrated how a misspecified main effect model tends to inaccurately estimate effects when true interactions are present. The latter observation suggests that extensions to the quadratic components can not only enable accurate estimation of interactions but also aid in deriving more accurate coefficients for main effects. For more details on the simulation setup as well as the influence of noise in the data on the performance of the sparse quadratic log-contrast model see [2].

6 Post estimation clustering and visualization

Providing clear and visually attractive illustrations of results from statistical analyses is crucial not only for interpretation but also for the successful communication of scientific findings [113, 9]. This is particularly important when multiple input features p_1 and interactions $(p_1(p_1-1)/2)$, as well as numerous outcome features p_2, are analyzed. In this final section, I propose two methods for summarizing and visually representing the feature effects derived from the interaction modeling strategies introduced in this thesis.

For the visualizations presented herein, I assume that the underlying data were analyzed by quadratic models of the from in Eq. 1 and are structured as depicted in Fig. 1, where $Y_{n\times p_2}$ comprises multiple features p_2 (e.g., a large set of proteins or multiple features describing the morphology of a cell).

Despite the independent analysis of each outcome Y_i for $i = 1,\ldots,p_2$, based on the same underlying input data $X_{n\times p_1}$, the model coefficients may exhibit similar patterns that are of interest. Organizing the vector representations of these estimated coefficients, $\hat{\beta}_i \in \mathbb{R}^{p_1}$ or $\hat{\Theta}_i \in \mathbb{R}^{p_1(p_1-1)/2}$ for $i = 1,\ldots,p_2$, into a common coefficient matrix, represented as $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p_1\times p_2}$ for main effects and $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}] \in \mathbb{R}^{p_1+p_1(p_1-1)/2\times p_2}$ for main and interaction effects, enables downstream tasks such as clustering or common visualizations.

First, I use a representation that summarizes all estimated coefficients within a clustered heatmap. To characterize each cluster in an interpretable way, I employ a technique that assigns a prototypical feature to each cluster [114], effectively representing the cluster while providing a condensed view of the results. This clustering technique is applicable to coefficients derived from both main effect and interaction models.

The second visual summary of the results corresponds to the modes of combinatorial behavior depicted in Fig. 2 and contrasts the individual effects $\hat{\beta}_j$ and $\hat{\beta}_k$ with the corresponding interaction effects $\hat{\Theta}_{jk}$. This representation provides an overview of the number of synergistic, antagonistic, and other effects identified.

In the following two subsections, I detail the post-estimation clustering methods and visualizations, and illustrate them with examples from the manuscripts [4, 1, 3].

6.1 Hierarchical clustering representation with prototypes

To represent the derived model coefficients in a clustered representation with prototypes, I follow the hierarchical clustering with prototypes via minimax linkage approach, introduced in [114]. This method extends traditional agglomerative hierarchical clustering by introducing the concept of prototypes, which are representative data points within each cluster. The key idea is to minimize the maximum dissimilarity between any point and its prototype. This concept is known as the minimax linkage and for a matrix of dissimilarities $d(x,x')$ it is mathematically represented as

$$\min_{x\in C}\max_{x'\in C} d(x,x') \tag{16}$$

where C is a cluster, x and x' are points in C. The inner part defines for any point $x \in C$ the farthest point in C from x. This maximum distance is then minimized to find the prototype of cluster C which is the point $x \in C$ whose farthest point is the closest. Finally, the minimax linkage between two clusters C_1 and C_2 takes the form in Eq. 16 by replacing C by $C_1 \cup C_1$. This method allows structuring the data into clusters and returns a subset of features (from Y) that can serve as a condensed view of the data (see [114] for the full algorithm).

In the following, I demonstrate this clustering technique in two scenarios.

Application in epigenetics: individual binding responses of chromatin readers to chromatin modifications

In the first example, I demonstrate how post-estimation clustering can be applied to a common coefficient matrix, characterized solely by main effects. This can be a meaningful step to gain a first impression of the main effect patterns within large-scale datasets before deriving more complex interaction effects, which might exhibit different clustering patterns. This analysis is detailed in the publication [4].

The underlying data for this example consists of a binary experimental design matrix $B \in \{0,1\}^{n \times p_1}$, which represents combinations of $p_1 = 15$ distinct chromatin modifications, and an outcome matrix $Y = (Y_1, ..., Y_{p_2}) \in \mathbb{R}^{n \times p_2}$ containing the observed binding profiles of $p_2 = 352$ proteins. Fig. 13 illustrates a clustered heatmap of the estimated coefficients $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p_1 \times p_2}$, showing the main effects of each chromatin modification on each protein. Each cluster is characterized by a prototypical protein, providing a straightforward description of each cluster. For more details, see [4].

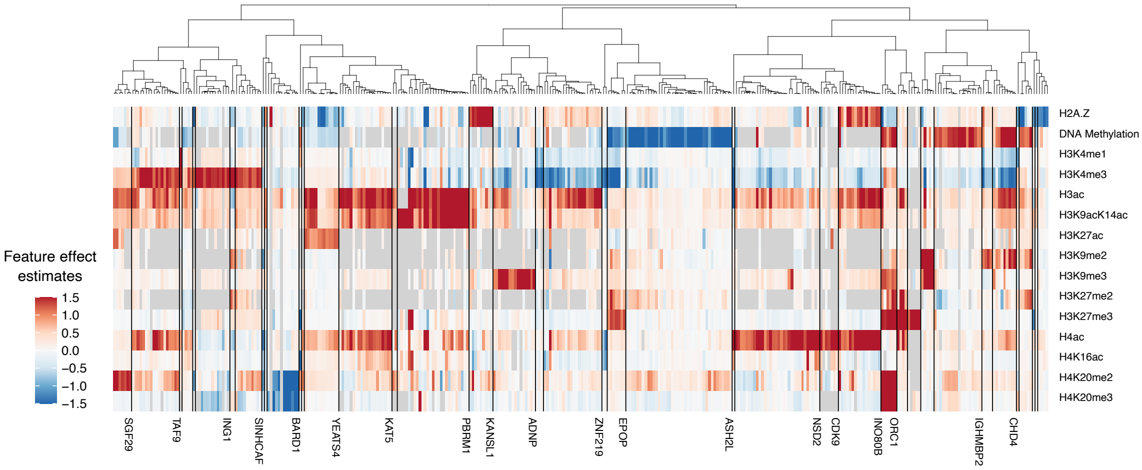


Figure 13: Hierarchical clustering with prototypes via minimax linkage on a main effects coefficient matrix $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p_1 \times p_2}$, as derived in [4], representing the effects of chromatin modifications (rows) on the binding response of chromatin reader proteins (columns). Prototypical proteins for clusters with more than five members are highlighted with labels. Adapted from [4].

Application in epigenetics: individual and combinatorial binding responses of chromatin readers to chromatin modifications

In this example, I utilize the concept of post-estimation clustering on a common coefficient matrix comprising main and interaction effects. Here, the underlying data is given by a binary experimental design matrix $B \in \{0,1\}^{n \times p_1}$ comprising combinations of $p_1 = 12$ distinct chromatin modifications across $n = 33$ experiments and observed binding profiles of $p_2 = 1915$ proteins $Y = (Y_1, ..., Y_{p_2}) \in \mathbb{R}^{n \times p_2}$. Note that while the underlying data is the same as in the previous example, the features are summarized differently prior to performing interaction modeling, and the full list of proteins is considered here. The main and interaction effect coefficients are derived from the forward model in Eq. 3 within the multi-stage statistical workflow `asteRIa` (for more details on the exact modeling strategy, see [1]). A clustered representation of the common coefficient matrix $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}] \in \mathbb{R}^{p_1 + p_1(p_1-1)/2 \times p_2}$ with prototypical proteins is shown in Fig. 14. This representation provides a joint representation of the coefficients from all p_2 models and shows that groups of proteins exhibit similar binding behavior to chromatin modifications. Gray represents exact zeros. Prototypical proteins for each cluster are labeled. Some of the few proteins that respond to combinatorial chromatin modification effects also serve as representatives of entire clusters (SAP30, URB2, ZMYM4).

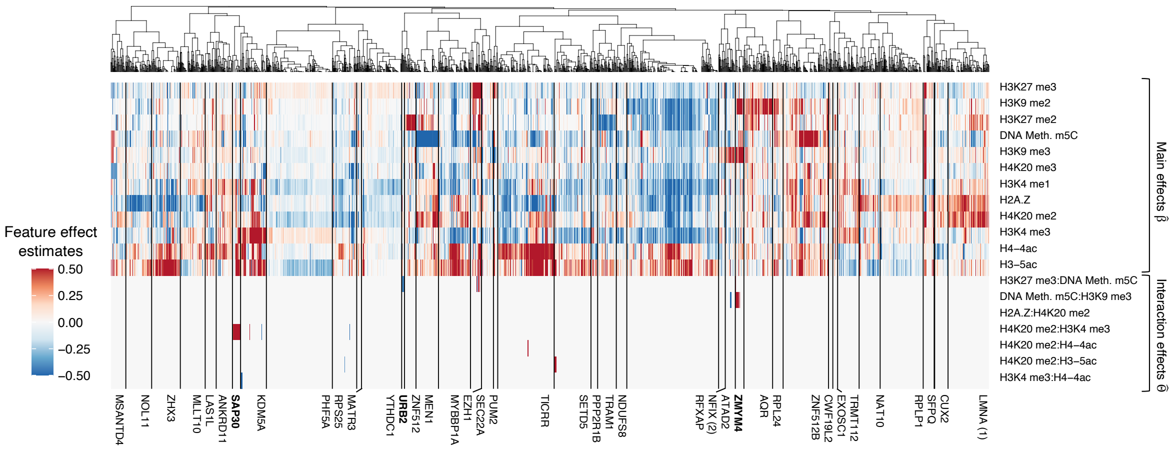


Figure 14: Hierarchical clustering with prototypes via minimax linkage on the common coefficient matrix of main and interaction effects $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}] \in \mathbb{R}^{p_1 + p_1(p_1-1)/2 \times p_2}$ derived in [1]. Only features (rows) that influence at least one protein are shown. Prototypical proteins are highlighted with labels. Proteins responding to interactions between chromatin modifications are marked in bold.

6.2 Visual summary of modes of interactions

The primary aim of the statistical approaches outlined in this thesis is the estimation and interpretation of stable interaction effects derived from the statistical models of the form in Eq. 1, as depicted by the modes in Fig. 2. In this section, I propose a method for visualizing these results through scatterplots that contrast the individual effects, $\hat{\beta}_j$ and $\hat{\beta}_k$, with their corresponding interaction effects, $\hat{\Theta}_{jk}$, directly linking the results to these modes. These three dimensions of information are incorporated as follows: the estimated main effects are displayed on the x- and y-axes, while the interaction effect is indicated by the color of each point. The sign of the interaction coefficient, along with its location in the scatterplot (quadrant), determines the mode of interaction. While there are established methods for contrasting main and interaction effects, particularly in studies of fitness landscapes as shown in Fig. 3, the approach I introduce provides a unified way to depict modes of combinatorial effects for outcomes, $Y \in \mathbb{R}^{n \times p_2}$, that involve multiple features p_2. I illustrate this representation with two applications, one detailed in the manuscript [1] and another in the draft manuscript [3].

Application in epigenetics: Modes of chromatin modification protein interactions

Here, subsets of the coefficients depicted in Fig. 14 are presented from a different perspective.

In [1], the primary goal was to identify stable interaction effects between pairs of $p_1 = 12$ chromatin modifications within a binary experimental design $B \in \{0,1\}^{n \times p_1}$ and their influence on the binding behavior of $p_2 = 1915$ proteins, given by $Y = (Y_1, ..., Y_{p_2}) \in \mathbb{R}^{n \times p_2}$. A small subset of 55 proteins, which respond to stable interaction effects between chromatin modifications, was identified. These effects are summarized in Fig. 15, providing an overview of the modes of combinatorial effects from Fig. 2.

For instance, proteins exhibiting positive interaction effects (red) in the first quadrant, such as the protein UHRF1, demonstrate a positive synergistic interaction between two chromatin modifications. This visualization highlights that chromatin modification interactions can exhibit various modes of combinatorial behavior. For more detailed findings from this study, see [1].

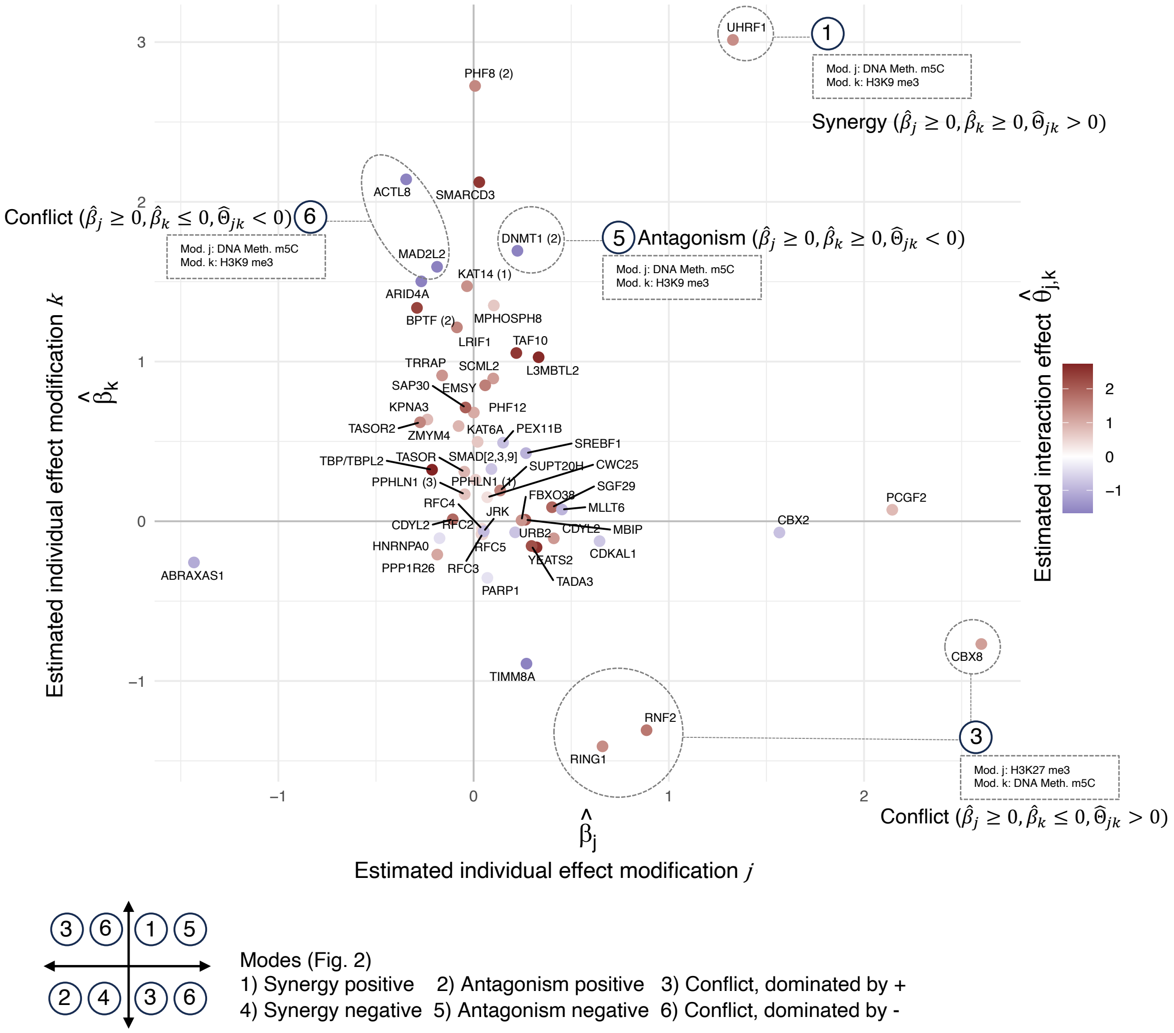
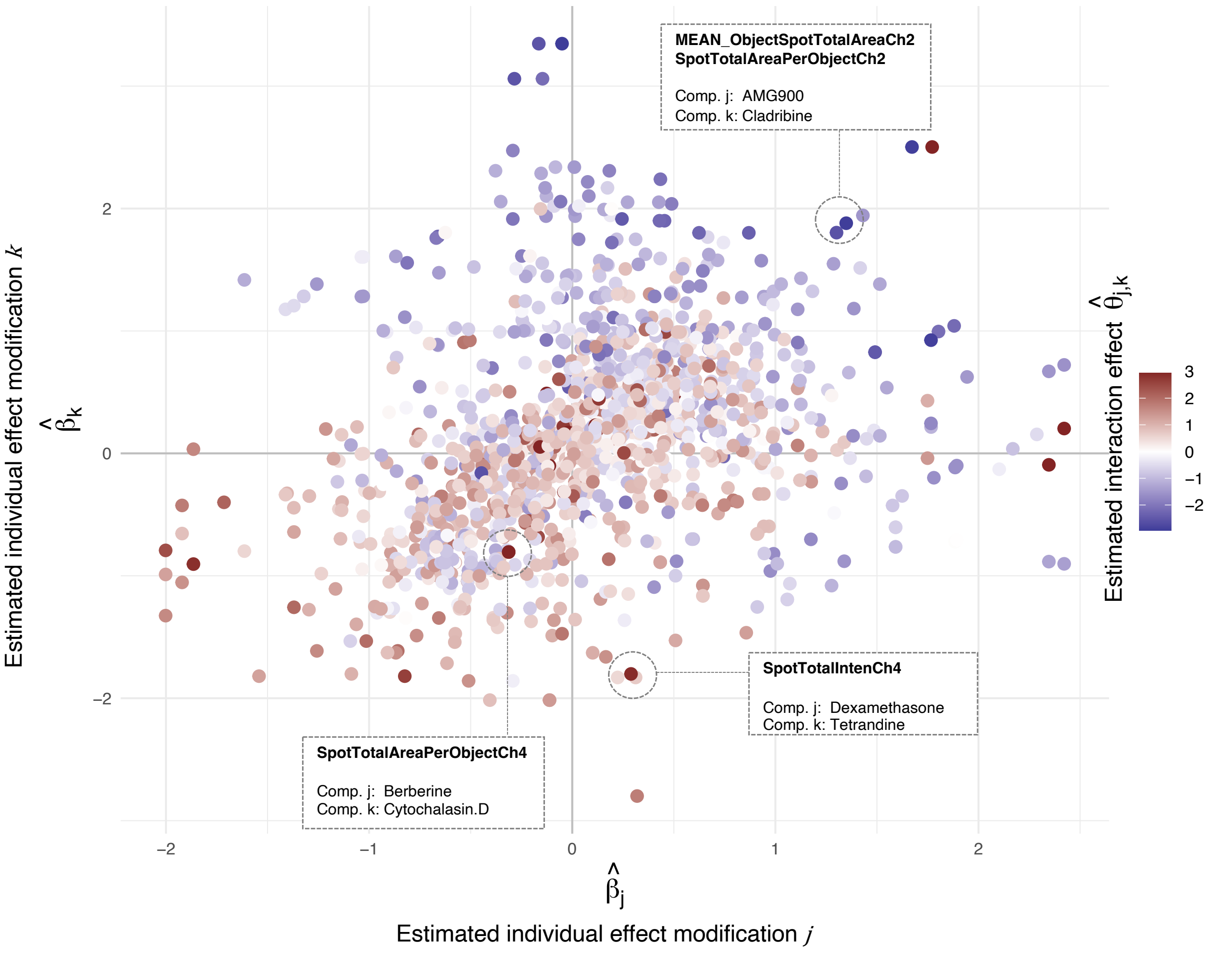


Figure 15: Scatterplot of protein binding effects with unspecific linear effects $\hat{\beta}_j$ and $\hat{\beta}_k$ on the x- and y-axis and corresponding additional combination effect $\hat{\Theta}_{jk}$ represented by color. For some example proteins, detailed information on the chromatin modifications is provided in the figure. For instance, the proteins RING1, RNF2, and CBX8 all show a conflicting behavior between the chromatin modifications H3K27me3 and DNA Meth. m5C. The bottom part of the figure includes a description indicating where each category defined in Fig. 2 is located within the scatterplot presented here. Adapted from [1].

Application in pharmacology: Modes of drug combination cell morphology interaction

In a second example, I illustrate the results from [3] using the same scatterplot representation. Here, the primary aim was to estimate drug interactions among $p_1 = 20$ compounds within a binary experimental design, represented as $B \in \{0,1\}^{n \times p_1}$, which includes $n = 408$ experiments. This analysis examines their influence on a set of $p_2 = 68$ cellular features, represented by $Y = (Y_1, ..., Y_{p_2}) \in \mathbb{R}^{n \times p_2}$. The identified interaction effects $\hat{\Theta}_{jk}$ and the corresponding main effects $\hat{\beta}_j$ and $\hat{\beta}_k$, derived within the interaction modeling strategy in [3], are presented in Fig. 16. This analysis identified a large set of interaction effects and labeling all dots is only practical when stratifying this representation, for example, by showing only a specific drug combination or specific features in Y corresponding to a certain region (channel) of the cell. Nevertheless, this joint representation provides an overall impression, indicating that large main effects of the same sign tend to exhibit antagonistic interaction effects. Moreover, some prototypical features, derived from the hierarchical clustering approach described in Section 6.1, and their corresponding interaction effects are labeled.



MEAN_ObjectSpotTotalAreaCh2
SpotTotalAreaPerObjectCh2

Comp. j: AMG900
Comp. k: Cladribine

SpotTotalIntenCh4

Comp. j: Dexamethasone
Comp. k: Tetrandine

SpotTotalAreaPerObjectCh4

Comp. j: Berberine
Comp. k: Cytochalasin.D

Overview Channels (Ch)
Ch1: Nucleus
Ch2: Endoplasmic reticulum
Ch3: Nucleoli, cytoplasmic RNA
Ch4: Golgi, plasma membrane, F-actin, cytoskeleton
Ch5: Mitochondria

Figure 16: Scatterplot of drug effects on features describing cellular morphology with unspecific linear effects $\hat{\beta}_j$ and $\hat{\beta}_k$ on the x- and y-axes, and the corresponding interaction effect $\hat{\Theta}_{jk}$ represented by color. Some prototypical morphological features are labeled for an illustration. Adapted from [3].

7 Summary and outlook

The main goal of this thesis was to advance toward a more interpretable and consistent estimation of combinatorial effects in data-driven biological research. The focus was on data derived from high-throughput technologies in settings with limited sample size or unbalanced experimental designs, and potentially large numbers of features and pairwise interactions. A major challenge was reducing the number of spurious effects under biological and experimental noise. Through the generation of simulated data and domain-expert feedback, I evaluated the stability and trustworthiness of detected effects in penalized quadratic regression models and defined how existing models need to be extended to derive parsimonious models with functionally meaningful biological signals.

In particular, the contributions [1], [2], and [3] introduce statistical workflows that combine two concepts, both of which have been shown to enhance stable model estimation: hierarchical interaction modeling [114] and stability-based model selection [100]. As outlined in [102] and [103], effectively minimizing experimental and biological noise to reveal stable and biologically relevant signals remains challenging. To address this issue, I have extended the statistical workflows to mitigate noise by introducing outlier removal mechanisms through replicate-consistency checks in contribution [1], and by integrating robust alternatives within the optimization problem in contribution [3]. Moreover, interaction modeling has proven to be very effective in many prediction tasks [14, 32]; however, so far, penalized interaction modeling strategies have not been defined for compositional data, which are an important component in biological sequencing data. To address this gap, I have introduced penalized interaction modeling strategies for compositional data in contribution [2], and have integrated these strategies with methods for deriving stable interactions as outlined in [1]. All these strategies are unified and generalized within one framework as part of this thesis. Finally, to interpret and effectively communicate the results, I have employed post-estimation clustering strategies in the contributions [4] and [3] to provide a condensed view of the findings and developed visualization strategies that summarize the derived modes of combinatorial behavior.

Overall, the statistical approaches introduced in this thesis allowed me to gain novel biological insights and develop generally applicable tools that enable data analyses of specific types. For instance, based on a large set of proteins and combinations of chromatin modifications, I was able to uncover a set of stable and previously unknown epigenetic reader protein candidates and prove the biological relevance by validating one of these findings with external data sources in various cell types in contribution [1]. In contribution [2], I identified sparse microbial interaction models that accurately predict the abundance of antimicrobial resistance genes, enabling the formulation of novel biological hypotheses about microbial community composition and antimicrobial resistance.

In conclusion, this thesis contributes to the development of reproducible statistical tools that enhance interpretability by adhering to concepts of stability [115, 116] and by integrating simulations to evaluate the robustness and trustworthiness in data-driven biological research [9]. This integration effectively bridges the gap between statistical theory and real-world applications.

As more datasets with a greater number of experiments become available, the quadratic regression models I defined in this thesis can be conveniently extended to study more and higher-order interactions. Particularly, as the numbers of features and higher-order terms increase, the currently used solvers for hierarchical interactions [69] can be replaced with computationally more efficient algorithms, such as those proposed in [74], or by adopting less strict assumptions on the interaction features through approaches that emphasize reluctance rather than strict hierarchy [77]. Furthermore, the ideas developed to estimate stable interaction effects can be easily adapted to accommodate arbitrary nonlinear effects of the form

$$y = \beta_0 + \sum_{j=1}^{p} X_j + g(X) + \epsilon,$$

where $g(X)$ can represent a nonlinear component such as those found in generalized additive models (GAMs) or a feed-forward neural network. Similar to the concept of hierarchy in interaction modeling, methods have been introduced that employ reluctance to nonlinearities in generalized additive models (GAMs) [117] and weak hierarchy assumptions in feed-forward neural networks [118]. This allows integrating feature sparsity, thereby maintaining interpretability while allowing the modeling of more complex relationships. In this thesis, I exclusively focused on regression approaches as they allow inferring actual effect sizes with straightforward interpretation, which are crucial for the modes of combinatorial behavior I introduced. However, particularly when considering expansion to more complex nonlinear effects, random forests (RF) present powerful alternatives [119, 120]. In a manner similar to the principles I adhered to in this thesis, an extension of RF approaches, the iterative random forests (iRF), for discovering predictive, stable, and interpretable higher-order interactions have also been introduced [121].
Moreover, the workflows developed in this thesis require each feature $y = Y_i$, for $i = 1, \dots, p_2$ to be analyzed within an individual model. Future work will have to integrate correlation structures among the features in Y to capture the complex dynamics of biological systems. These models could resemble ideas from sparse matrix models used in analyzing linear associations [122, 123, 124]. In such joint approaches, pre-imposed group penalties on the features in Y could be integrated (e.g., proteins within the same complex should respond jointly), for instance, inspired by group lasso approaches [125].

Bibliography

[1] M. Stadler, S. Lukauskas, T. Bartke, and C. L. Müller. “asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers”. In: *Nucleic Acids Research* 52.11 (2024), pp. 6129–6144.

[2] M. Stadler, C. L. Müller, and J. Bien. “Predictive modeling of microbial data with interaction effects”. In: *bioRxiv* 2024–04 (2024).

[3] M. Stadler, E. Kupczyk, L. Buckett, X. Zhang, and C. Müller. “A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach”. In: *Draft manuscript* (2024).

[4] S. Lukauskas, A. Tvardovskiy, N. V. Nguyen, M. Stadler, P. Faull, T. Ravnsborg, B. Özdemir Aygenli, S. Dornauer, H. Flynn, R. G. Lindeboom, et al. “Decoding chromatin states by proteomic profiling of nucleosome readers”. In: *Nature* (2024), pp. 1–9.

[5] J. A. Reuter, D. V. Spacek, and M. P. Snyder. “High-throughput sequencing technologies”. In: *Molecular cell* 58.4 (2015), pp. 586–597.

[6] D. Deshpande, K. Chhugani, Y. Chang, A. Karlsberg, C. Loeffler, J. Zhang, A. Muszyńska, V. Munteanu, H. Yang, J. Rotman, et al. “RNA-seq data science: From raw data to effective interpretation”. In: *Frontiers in Genetics* 14 (2023), p. 997383.

[7] B. J. Callahan, K. Sankaran, J. A. Fukuyama, P. J. McMurdie, and S. P. Holmes. “Bioconductor workflow for microbiome data analysis: from raw reads to community analyses”. In: *F1000Research* 5 (2016).

[8] F. Zanella, J. B. Lorens, and W. Link. “High content screening: seeing is believing”. In: *Trends in biotechnology* 28.5 (2010), pp. 237–245.

[9] S. H. Holmes and W. Huber. *Modern statistics for modern biology.* Cambridge university press, 2018.

[10] R. A. Fisher, R. A. Fisher, S. Genetiker, R. A. Fisher, S. Genetician, G. Britain, R. A. Fisher, and S. Généticien. *The design of experiments.* Vol. 21. Oliver and Boyd Edinburgh, 1966.

[11] J. Neyman and E. Pearson. “Sufficient statistics and uniformly most power-ful tests of statistical hypotheses”. In: *Joint Statistical Papers* (1936), p. 240.

[12] J. W. Tukey et al. *Exploratory data analysis.* Vol. 2. Reading, MA, 1977.

[13] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[14] A. L. Gould, V. Zhang, L. Lamberti, E. W. Jones, B. Obadia, N. Korasidis, A. Gavryushkin, J. M. Carlson, N. Beerenwinkel, and W. B. Ludington. “Microbiome interactions shape host fitness”. In: *Proceedings of the National Academy of Sciences* 115.51 (2018), E11951–E11960.

[15] T. Kouzarides. “Chromatin modifications and their function”. en. In: *Cell* 128.4 (2007), pp. 693–705.

[16] B. A. Garcia, J. J. Pesavento, C. A. Mizzen, and N. L. Kelleher. “Pervasive combinatorial modification of histone H3 in human cells”. en. In: *Nat. Methods* 4.6 (2007), pp. 487–489.

[17] J. J. Pesavento, C. R. Bullock, R. D. LeDuc, C. A. Mizzen, and N. L. Kelleher. “Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry”. en. In: *J. Biol. Chem.* 283.22 (2008), pp. 14927–14937.

[18] E. Shema, D. Jones, N. Shoresh, L. Donohue, O. Ram, and B. E. Bernstein. “Single-molecule decoding of combinatorially modified nucleosomes”. en. In: *Science* 352.6286 (2016), pp. 717–721.

[19] A. Tvardovskiy, V. Schwämmle, S. J. Kempf, A. Rogowska-Wrzesinska, and O. N. Jensen. “Accumulation of histone variant H3.3 with age is associated with profound changes in the histone methylation landscape”. en. In: *Nucleic Acids Res.* 45.16 (2017), pp. 9272–9289.

[20] P. Voigt, G. LeRoy, W. J. Drury 3rd, B. M. Zee, J. Son, D. B. Beck, N. L. Young, B. A. Garcia, and D. Reinberg. “Asymmetrically modified nucleosomes”. en. In: *Cell* 151.1 (2012), pp. 181–193.

[21] N. L. Young, P. A. DiMaggio, M. D. Plazas-Mayorca, R. C. Baliban, C. A. Floudas, and B. A. Garcia. “High throughput characterization of combinatorial histone codes”. en. In: *Mol. Cell. Proteomics* 8.10 (2009), pp. 2266–2284.

[22] S. Li, Y. Peng, and A. R. Panchenko. “DNA methylation: Precise modulation of chromatin structure and dynamics”. In: *Current Opinion in Structural Biology* 75 (2022), p. 102430.

[23] B. M. Turner. “Decoding the nucleosome”. en. In: *Cell* 75.1 (1993), pp. 5–8.

[24] B. D. Strahl and C. D. Allis. “The language of covalent histone modifications”. en. In: *Nature* 403.6765 (2000), pp. 41–45.

[25] T. Jenuwein and C. D. Allis. “Translating the histone code”. en. In: *Science* 293.5532 (2001), pp. 1074–1080.

[26] P. C. Phillips. “Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems”. In: *Nature Reviews Genetics* 9.11 (2008), pp. 855–867.

[27] D. M. Weinreich, Y. Lan, J. Jaffe, and R. B. Heckendorn. “The influence of higher-order epistasis on biological fitness landscape topography”. In: *Journal of statistical physics* 172 (2018), pp. 208–225.

[28] N. Beerenwinkel, L. Pachter, and B. Sturmfels. “Epistasis and shapes of fitness landscapes”. In: *Statistica Sinica* (2007), pp. 1317–1342.

[29] H. J. Cordell. “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans”. In: *Human molecular genetics* 11.20 (2002), pp. 2463–2468.

[30] J. Diaz-Colunga, A. Skwara, K. Gowda, R. Diaz-Uriarte, M. Tikhonov, D. Bajic, and A. Sanchez. “Global epistasis on fitness landscapes”. In: *Philosophical Transactions of the Royal Society B* 378.1877 (2023), p. 20220053.

[31] F. J. Poelwijk, V. Krishna, and R. Ranganathan. “The context-dependence of mutations: a linkage of formalisms”. In: *PLoS computational biology* 12.6 (2016), e1004771.

[32] A. Skwara, K. Gowda, M. Yousef, J. Diaz-Colunga, A. S. Raman, A. Sanchez, M. Tikhonov, and S. Kuehn. “Statistically learning the functional landscape of microbial communities”. In: *Nature Ecology & Evolution* 7.11 (2023), pp. 1823–1833.

[33] E. D. Weinberger. “Fourier and Taylor series on fitness landscapes”. In: *Biological cybernetics* 65.5 (1991), pp. 321–330.

[34] E. J. Culp and A. L. Goodman. “Cross-feeding in the gut microbiome: Ecology and mechanisms”. In: *Cell Host & Microbe* 31.4 (2023), pp. 485–499.

[35] W. Z. Lidicker Jr. “A clarification of interactions in ecological systems”. In: *Bioscience* 29.8 (1979), pp. 475–477.

[36] K. Faust and J. Raes. “Microbial interactions: from networks to models”. In: *Nature Reviews Microbiology* 10.8 (2012), pp. 538–550.

[37] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. “Sparse and compositionally robust inference of microbial ecological networks”. In: *PLoS computational biology* 11.5 (2015), e1004226.

[38] S. Peschel, C. L. Müller, E. Von Mutius, A.-L. Boulesteix, and M. Depner. “NetCoMi: network construction and comparison for microbiome data in R”. In: *Briefings in bioinformatics* 22.4 (2021), bbaa290.

[39] N. Meinshausen and P. Bühlmann. “High-dimensional graphs and variable selection with the lasso”. In: (2006).

[40] G. Yoon, I. Gaynanova, and C. L. Müller. “Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data”. In: *Frontiers in genetics* 10 (2019), p. 516.

[41] T. A. Joseph, L. Shenhav, J. B. Xavier, E. Halperin, and I. Pe’er. “Compositional Lotka-Volterra describes microbial dynamics in the simplex”. In: *PLoS computational biology* 16.5 (2020), e1007917.

[42] M. E. Muscarella and J. P. O’Dwyer. “Species dynamics and interactions via metabolically informed consumer-resource models”. In: *Theoretical Ecology* 13.4 (2020), pp. 503–518.

[43] J. Foucquier and M. Guedj. “Analysis of drug combinations: current methodological landscape”. In: *Pharmacology research & perspectives* 3.3 (2015), e00149.

[44] C. I. Bliss. “The toxicity of poisons applied jointly 1”. In: *Annals of applied biology* 26.3 (1939), pp. 585–615.

[45] W. Zhao, K. Sachsenmeier, L. Zhang, E. Sult, R. E. Hollingsworth, and H. Yang. “A new bliss independence model to analyze drug combination data”. In: *Journal of biomolecular screening* 19.5 (2014), pp. 817–821.

[46] W. A. Shewhart and S. S. Wilks. *Wiley series in probability and mathematical statistics*. Wiley, 1984.

[47] S. t. Loewe. “Effect of combinations: mathematical basis of problem”. In: *Arch. Exp. Pathol. Pharmakol.* 114 (1926), pp. 313–326.

[48] D. M. Jonker, S. A. Visser, P. H. Van Der Graaf, R. A. Voskuyl, and M. Danhof. “Towards a mechanism-based analysis of pharmacodynamic drug–drug interactions in vivo”. In: *Pharmacology & therapeutics* 106.1 (2005), pp. 1–18.

[49] Y. S. Low, A. C. Daugherty, E. A. Schroeder, W. Chen, T. Seto, S. Weber, M. Lim, T. Hastie, M. Mathur, M. Desai, et al. “Synergistic drug combinations from electronic health records and gene expression”. In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 565–576.

[50] J. Nelder. “A reformulation of linear models”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 140.1 (1977), pp. 48–63.

[51] L. S. Aiken, S. G. West, and R. R. Reno. *Multiple regression: Testing and interpreting interactions*. sage, 1991.

[52] M. Hamada and C. J. Wu. “Analysis of designed experiments with complex aliasing”. In: *Journal of quality technology* 24.3 (1992), pp. 130–137.

[53] R. P. Duncan and B. J. Kefford. “Interactions in statistical models: three things to know”. In: *Methods in Ecology and Evolution* 12.12 (2021), pp. 2287–2297.

[54] U. Simonsohn. “Interacting with curves: How to validly test and probe interactions in the real (nonlinear) world”. In: *Advances in Methods and Practices in Psychological Science* 7.1 (2024), p. 25152459231207787.

[55] L. Tolosa, M. J. Gómez-Lechón, and M. T. Donato. “High-content screening technology for studying drug-induced hepatotoxicity in cell models”. In: *Archives of toxicology* 89 (2015), pp. 1007–1022.

[56] D. Siegismund, M. Fassler, S. Heyse, and S. Steigele. “Benchmarking feature selection methods for compressing image information in high-content screening”. In: *SLAS technology* 27.1 (2022), pp. 85–93.

[57] C. A. Musselman, M.-E. Lalonde, J. Côté, and T. G. Kutateladze. “Perceiving the epigenetic landscape through histone readers”. In: *Nature structural & molecular biology* 19.12 (2012), pp. 1218–1227.

[58] A. J. Bannister and T. Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell research* 21.3 (2011), pp. 381–395.

[59] M. V. Greenberg and D. Bourc’his. “The diverse roles of DNA methylation in mammalian development and disease”. In: *Nature reviews Molecular cell biology* 20.10 (2019), pp. 590–607.

[60] T. Bartke, M. Vermeulen, B. Xhemalce, S. C. Robson, M. Mann, and T. Kouzarides. "Nucleosome-interacting proteins regulated by DNA and histone methylation". In: *Cell* 143.3 (2010), pp. 470–484.

[61] L. R. Comolli. "Intra-and inter-species interactions in microbial communities". In: *Frontiers in microbiology* 5 (2014), p. 122250.

[62] M. Bickle. "The beautiful cell: high-content screening in drug discovery". In: *Analytical and bioanalytical chemistry* 398 (2010), pp. 219–226.

[63] Y. E. Pearson, S. Kremb, G. L. Butterfoss, X. Xie, H. Fahs, and K. C. Gunsalus. "A statistical framework for high-content phenotypic profiling using cellular feature distributions". In: *Communications Biology* 5.1 (2022), p. 1409.

[64] A. Kümmel, P. Selzer, M. Beibel, H. Gubler, C. N. Parker, and D. Gabriel. "Comparison of multivariate data analysis strategies for high-content screening". In: *Journal of biomolecular screening* 16.3 (2011), pp. 338–347.

[65] F. Reisen, X. Zhang, D. Gabriel, and P. Selzer. "Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery". In: *Journal of biomolecular screening* 18.10 (2013), pp. 1284–1297.

[66] J. L. Peixoto. "Hierarchical variable selection in polynomial regression models". In: *The American Statistician* 41.4 (1987), pp. 311–313.

[67] H. Chipman. "Bayesian variable selection with related predictors". In: *Canadian Journal of Statistics* 24.1 (1996), pp. 17–36.

[68] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. London: Chapman Hall, 1983.

[69] J. Bien, J. Taylor, and R. Tibshirani. "A lasso for hierarchical interactions". In: *The Annals of Statistics* 41.3 (2013). ISSN: 0090-5364.

[70] D. R. Cox. "Interaction". In: *International Statistical Review* 52 (1984), pp. 1–31.

[71] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. "Genome-wide association analysis by lasso penalized logistic regression". In: *Bioinformatics* 25.6 (2009), pp. 714–721.

[72] N. Hao and H. H. Zhang. "Interaction screening for ultrahigh-dimensional data". In: *Journal of the American Statistical Association* 109.507 (2014), pp. 1285–1301.

[73] R. D. Shah. "Modelling interactions in high-dimensional data with backtracking". In: *Journal of Machine Learning Research* 17.207 (2016), pp. 1–31.

[74] N. Hao, Y. Feng, and H. H. Zhang. "Model selection for high-dimensional quadratic regression via regularization". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 615–625.

[75] M. Lim and T. Hastie. "Learning Interactions via Hierarchical Group-Lasso Regularization". In: *Journal of Computational and Graphical Statistics* 24.3 (2015), pp. 627–654.

[76] A. Haris, D. Witten, and N. Simon. "Convex modeling of interactions with strong heredity". In: *Journal of Computational and Graphical Statistics* 25.4 (2016), pp. 981–1004.

[77] G. Yu, J. Bien, and R. Tibshirani. "Reluctant interaction modeling". In: *arXiv preprint arXiv:1907.08414* (2019).

[78] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. "Microbiome datasets are compositional: and this is not optional". In: *Frontiers in microbiology* 8 (2017), p. 294209.

[79] A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn. "Compositional data analysis in the geosciences: from theory to practice". In: Geological Society of London. 2006.

[80] J. Aitchison and J. Bacon-Shone. "Log contrast models for experiments with mixtures". In: *Biometrika* 71.2 (1984), pp. 323–330.

[81] S. Bates and R. Tibshirani. "Log-ratio lasso: scalable, sparse estimation for log-ratio models". In: *Biometrics* 75.2 (2019), pp. 613–624.

[82] P. L. Combettes and C. L. Müller. "Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications". In: *Statistics in Biosciences* 13.2 (2021), pp. 217–242.

[83] A. Mishra and C. L. Müller. "Robust regression with compositional covariates". In: *Computational Statistics & Data Analysis* 165 (2022), p. 107315.

[84] P. Shi, A. Zhang, and H. Li. "Regression analysis for microbiome compositional data". In: (2016).

[85] J. Bien, X. Yan, L. Simpson, and C. L. Müller. "Tree-aggregated predictive modeling of microbiome data". In: *Scientific Reports* 11.1 (2021), p. 14505.

[86] M. Greenacre, E. Grunsky, J. Bacon-Shone, I. Erb, and T. Quinn. "Aitchison's compositional data analysis 40 years on: A reappraisal". In: *Statistical Science* 38.3 (2023), pp. 386–410.

[87] A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. "Analysis of microbiome data in the presence of excess zeros". In: *Frontiers in microbiology* 8 (2017), p. 283205.

[88] W. Lin, P. Shi, R. Feng, and H. Li. "Variable selection in regression with compositional covariates". In: *Biometrika* 101.4 (2014), pp. 785–797.

[89] J. Bien and R. Tibshirani. *hierNet: A Lasso for Hierarchical Interactions.* R package version 1.9. 2020.

[90] J. Lederer and C. Müller. "Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX". In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 29. 1. 2015.

[91] Y. Wu and L. Wang. "A survey of tuning parameter selection for high-dimensional regression". In: *Annual review of statistics and its application* 7 (2020), pp. 209–226.

[92] N. Meinshausen and P. Bühlmann. "Stability Selection". In: *Journal of the Royal Statistical Society, Series B* 72 (2010), pp. 417–473.

[93] D. Chetverikov, Z. Liao, and V. Chernozhukov. "On cross-validated lasso in high dimensions". In: *The Annals of Statistics* 49.3 (2021), pp. 1300–1317.

[94] J. Friedman, R. Tibshirani, and T. Hastie. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22.

[95] G. Yu. *sprintr: Sparse Reluctant Interaction Modeling*. R package version 0.9.0. 2019.

[96] H. Liu, K. Roeder, and L. Wasserman. “Stability approach to regularization selection (stars) for high dimensional graphical models”. In: *Advances in neural information processing systems* 23 (2010).

[97] B. Bodinier, S. Filippi, T. H. Nøst, J. Chiquet, and M. Chadeau-Hyam. “Automated calibration for stability selection in penalised regression and graphical models”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* (2023), qlad058.

[98] S. Maddu, B. L. Cheeseman, I. F. Sbalzarini, and C. L. Müller. “Stability selection enables robust learning of differential equations from limited noisy data”. In: *Proceedings of the Royal Society A* 478.2262 (2022), p. 20210916.

[99] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton. “Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control”. In: *Proceedings of the Royal Society A* 478.2260 (2022), p. 20210904.

[100] R. D. Shah and R. J. Samworth. “Variable selection with error control: Another look at stability selection”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75.1 (2013), pp. 55–80. ISSN: 13697412.

[101] B. Hofner and T. Hothorn. *stabs: Stability Selection with Error Control*. R package version 0.6-4. 2021.

[102] I. Moutsopoulos, L. Maischak, E. Lauzikaite, S. A. Vasquez Urbina, E. C. Williams, H.-G. Drost, and I. I. Mohorianu. “noisyR: enhancing biological signal in sequencing datasets by characterizing random technical noise”. In: *Nucleic Acids Research* 49.14 (2021), e83–e83.

[103] N. Eling, M. D. Morgan, and J. C. Marioni. “Challenges in measuring and understanding biological noise”. In: *Nature Reviews Genetics* 20.9 (2019), pp. 536–548.

[104] P. Blainey, M. Krzywinski, and N. Altman. “Replication: quality is often more important than quantity”. In: *Nature Methods* 11.9 (2014), pp. 879–881.

[105] M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, et al. “A benchmark for RNA-seq quantification pipelines”. In: *Genome biology* 17.1 (2016), pp. 1–12.

[106] P. J. Huber. “Robust estimation of a location parameter”. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[107] Y. Liu, P. Zeng, and L. Lin. “Degrees of freedom for regularized regression with Huber loss and linear constraints”. In: *Statistical Papers* 62.5 (2021), pp. 2383–2405.

[108] J. C. Gardiner, Z. Luo, and L. A. Roman. “Fixed effects, random effects and GEE: What are the differences?” In: *Statistics in medicine* 28.2 (2009), pp. 221–239.

[109] T. J. Kozubowski and K. Podgórski. “A multivariate and asymmetric generalization of Laplace distribution”. In: *Computational Statistics* 15 (2000), pp. 531–540.

[110] M. C. Tsilimigras and A. A. Fodor. “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. In: *Annals of epidemiology* 26.5 (2016), pp. 330–335.

[111] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, et al. “American gut: an open platform for citizen science microbiome research”. In: *Msystems* 3.3 (2018), pp. 10–1128.

[112] J. M. Janda and S. L. Abbott. “16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls”. In: *Journal of clinical microbiology* 45.9 (2007), pp. 2761–2764.

[113] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. “Visualizing high-dimensional data: Advances in the past decade”. In: *IEEE transactions on visualization and computer graphics* 23.3 (2016), pp. 1249–1268.

[114] J. Bien and R. Tibshirani. “Hierarchical clustering with prototypes via minimax linkage”. In: *Journal of the American Statistical Association* 106.495 (2011), pp. 1075–1084.

[115] B. Yu. “Three principles of data science: predictability, computability, and stability (PCS)”. In: (2018).

[116] B. Yu. “Veridical data science”. In: *Proceedings of the 13th international conference on web search and data mining.* 2020, pp. 4–5.

[117] J. K. Tay and R. Tibshirani. “Reluctant generalised additive modelling”. In: *International Statistical Review* 88 (2020), S205–S224.

[118] I. Lemhadri, F. Ruan, L. Abraham, and R. Tibshirani. “Lassonet: A neural network with feature sparsity”. In: *Journal of Machine Learning Research* 22.127 (2021), pp. 1–29.

[119] X. Chen and H. Ishwaran. “Random forests for genomic data analysis”. In: *Genomics* 99.6 (2012), pp. 323–329.

[120] R. Hornung and A.-L. Boulesteix. “Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects”. In: *Computational Statistics & Data Analysis* 171 (2022), p. 107460.

[121] S. Basu, K. Kumbier, J. B. Brown, and B. Yu. “Iterative random forests to discover predictive and stable high-order interactions”. In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948.

[122] G. Yoon, R. J. Carroll, and I. Gaynanova. “Sparse semiparametric canonical correlation analysis for data of mixed types”. In: *Biometrika* 107.3 (2020), pp. 609–625.

[123] D. Kobak, Y. Bernaerts, M. A. Weis, F. Scala, A. S. Tolias, and P. Berens. “Sparse reduced-rank regression for exploratory visualisation of paired multivariate data”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 70.4 (2021), pp. 980–1000.

[124] A. Mishra, D. K. Dey, Y. Chen, and K. Chen. “Generalized co-sparse factor regression”. In: *Computational statistics & data analysis* 157 (2021), p. 107127.

[125] J. Friedman, T. Hastie, and R. Tibshirani. “A note on the group lasso and a sparse group lasso”. In: *arXiv preprint arXiv:1001.0736* (2010).

A Contributions as first author

A.1 asteRIa enables robust interaction modeling between chromatin modifications and epigenetic reader

Contributing article

Stadler, M., Lukauskas, S., Bartke, T., and Müller, C.L. (2024). asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers. *Nucleic Acids Research, 6129–6144*. doi: https://doi.org/10.1093/nar/gkae361

Replication code

The source data and code for reproducing all results of this study is available at https://doi.org/10.6084/m9.figshare.25003103.v2. The `asteRIa` workflow is available at https://github.com/marastadler/asteRIa.

Copyright information

Author contributions

M.S. developed asteRIa, conducted the analysis on the MARCS data, ChIP-Seq and bisulfite data and conducted the analysis on synthetic data. C.L.M. and T.B. supervised the work. M.S. and C.L.M. conceived the statistical workflow. S.L. and T.B. analyzed the results and provided feedback. M.S., C.L.M. and T.B. wrote the manuscript. All authors read and approved the final manuscript.

Nucleic Acids Research, 2024, **52**, 6129–6144
https://doi.org/10.1093/nar/gkae361
Advance access publication date: 16 May 2024
Computational Biology

asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers

Mara Stadler[1,2,*], **Saulius Lukauskas**[3], **Till Bartke**[3,†] **and Christian L. Müller**[1,2,4,†]

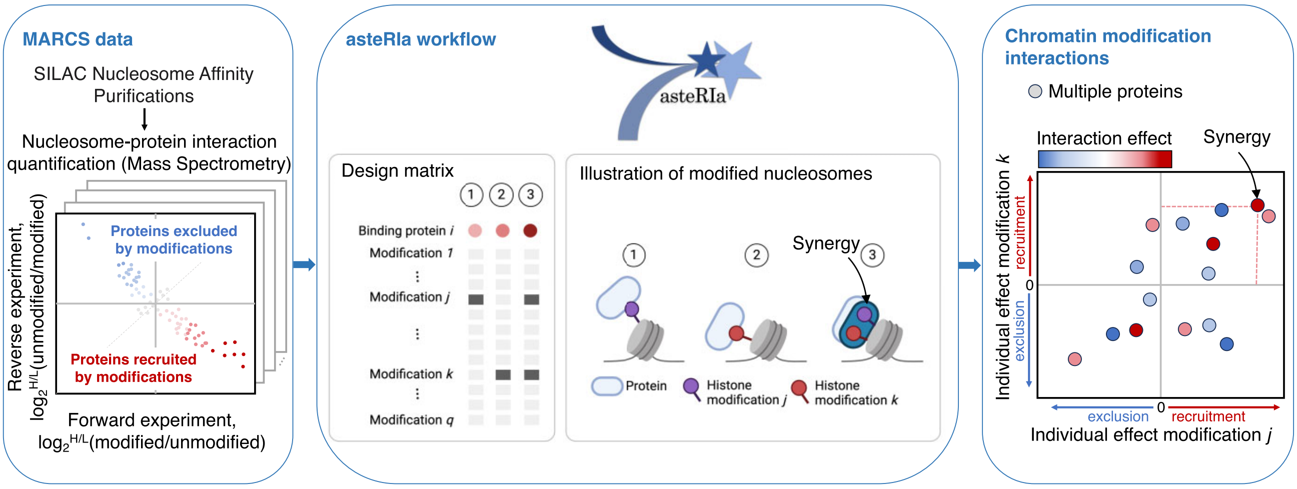
[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Department of Statistics, Ludwig-Maximilians-University Munich, 80539 Munich, Germany
[3]Institute of Functional Epigenetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[4]Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA

[*]To whom correspondence should be addressed. Tel: +49 8921803466; Email: mara.stadler@stat.uni-muenchen.de
[†]Christian L. Müller and Till Bartke co-supervised the project.

Abstract

Chromatin, the nucleoprotein complex consisting of DNA and histone proteins, plays a crucial role in regulating gene expression by controlling access to DNA. Chromatin modifications are key players in this regulation, as they help to orchestrate DNA transcription, replication, and repair. These modifications recruit epigenetic 'reader' proteins, which mediate downstream events. Most modifications occur in distinctive combinations within a nucleosome, suggesting that epigenetic information can be encoded in combinatorial chromatin modifications. A detailed understanding of how multiple modifications cooperate in recruiting such proteins has, however, remained largely elusive. Here, we integrate nucleosome affinity purification data with high-throughput quantitative proteomics and hierarchical interaction modeling to estimate combinatorial effects of chromatin modifications on protein recruitment. This is facilitated by the computational workflow asteRIa which combines hierarchical interaction modeling, stability-based model selection, and replicate-consistency checks for **a st**able **e**stimation of **R**obust **I**nter**a**ctions among chromatin modifications. asteRIa identifies several epigenetic reader candidates responding to specific *interactions* between chromatin modifications. For the polycomb protein CBX8, we independently validate our results using genome-wide ChIP-Seq and bisulphite sequencing datasets. We provide the first quantitative framework for identifying cooperative effects of chromatin modifications on protein binding.

Graphical abstract


MARCS data
SILAC Nucleosome Affinity Purifications
Nucleosome-protein interaction quantification (Mass Spectrometry)
Proteins excluded by modifications
Proteins recruited by modifications
Reverse experiment, $\log_2^{H/L}$(unmodified/modified)
Forward experiment, $\log_2^{H/L}$(modified/unmodified)
asteRIa workflow
asteRIa
Design matrix
1
2
3
Binding protein *i*
Modification *1*
Modification *j*
Modification *k*
Modification *q*
Illustration of modified nucleosomes
1
2
Synergy
3
Protein
Histone modification *j*
Histone modification *k*
Chromatin modification interactions
Multiple proteins
Interaction effect
Synergy
Individual effect modification *k*
recruitment
exclusion
0
exclusion
recruitment
Individual effect modification *j*

Introduction

Eukaryotic cells store the genetic material in the nucleus where it is packaged into chromatin, a nucleo-protein complex made up primarily of DNA and histone proteins. Both DNA and histones carry chemical modifications that can either directly affect chromatin structure or recruit so-called epigenetic reader proteins that mediate downstream events. As these modifications are involved in the regulation of all DNA-templated processes, such as transcription, DNA replication, or DNA repair, they play central roles in controlling chromatin function (1). The basic repeating unit of chromatin is the nucleosome, which coordinates 147 bp of DNA wrapped around an octamer consisting of two copies each of the core histones H2A, H2B, H3 and H4 (2). Nucleosomes are folded into higher-order structures to form chromatin. Since DNA and histone modifications show extensive overlap in the genome (3) and

Received: January 17, 2024. **Revised:** March 15, 2024. **Editorial Decision:** April 20, 2024. **Accepted:** April 24, 2024

decorate histones and nucleosomes in specific combinations (4–10), it is likely that these modifications act in a concerted manner. This is supported by the observation that most chromatin regulators contain multiple modification binding domains or are part of multi-subunit complexes harbouring multiple such domains, and are therefore likely to read out multiple chromatin modifications (11). Indeed, the idea that combinations of histone modifications may form a 'histone code' that together with DNA modifications could store epigenetic information in the chromatin template, thereby expanding the genetic information encoded in the DNA sequence, has been around for over two decades (12–14).

To date, the functions and readers of a host of *individual* chromatin modifications have been described (see, e.g., (15,16) and references therein for an overview). Moreover, epigenetic regulators that read the modification status of more than one epigenetic mark on histones or the DNA have been described using functional and structural studies (17–23). Several DNA repair factors were also found to recognize dual histone modification signatures, ranging from individual interactions (24–29) to combinatorial ones (30,31). One prime example is the ubiquitin ligase UHRF1, an essential player in DNA methylation maintenance, that recognizes a triple modification signature on histone H3 (32–34) and the DNA (35–37).

The gap in knowledge about the combinatorial nature of factors that read multiple DNA and histone modifications can be partially attributed to the fact that one of the most prevailing high-throughput technology to study histone modifications and their readers is chromatin immunoprecipitation followed by deep sequencing (ChIP-seq). Here, antibodies are used to detect the localization of specific modifications *or* chromatin-binding proteins at a genome-wide scale (38). Despite its groundbreaking influence on our understanding of the histone code through community efforts such as the NIH Roadmap Epigenomics Mapping Consortium (39), ENCODE (40), and ChIP-Atlas (41), ChIP-seq alone can only probe a single modification or reader protein in each experiment, thus making it difficult to assess combinatorial synergies or antagonistic effects on epigenetic readers. However, careful integration of multiple genome-wide ChIP-seq experiments of individual modifications enabled the application of *multivariate* statistical analysis techniques to uncover chromatin states and interactions. For example, using hidden Markov modeling techniques, the `ChromHMM` method (42,43) revealed cell-type specific discrete chromatin states that characterize the combinatorial presence or absence of modifications on the genome. Alternatively, sparse partial correlation estimation techniques were proposed to learn multivariate association networks between histone modifications (44). The latter framework was extended in (45,46) to include both histone modifications and a small set of chromatin modifiers. Using linear regression and sparse partial correlation estimation, the studies derived *de novo* high-confidence backbones of 'chromatin signaling networks' from ChIP-Seq data. There, the inferred network edges are to be interpreted as additive (or main) effects between histone modifications on chromatin modifiers and vice versa. The analysis of the derived chromatin signaling networks revealed both histone-protein interactions known from literature and several novel hypothetical interactions. To show the power of the network approach, the authors were also able to experimentally verify the statistically hypothesized interactions between H4K20me1 and members of the polycomb repressive complexes 1 and 2 (PRC1 and PRC2, respectively) (46). Nevertheless, none of these ChIP-Seq-based computational approaches allow the statistical estimation of how *multiple* histone modifications co-operate in recruiting epigenetic regulators.

In this contribution, we present a statistical interaction modeling approach, termed `asteRIa`, that tackles this challenge. Rather than considering genome-wide ChIP-Seq data, `asteRIa` uses novel nucleosome affinity purification data with high-throughput quantitative proteomics, as provided in the Modification Atlas of Regulation by Chromatin States (MARCS), to make robust and reproducible predictions of combinatorial effects of chromatin modifications on chromatin-interacting proteins. The MARCS data, available at https://marcs.helmholtz-munich.de comprises a collection of Stable Isotope Labeling with Amino acids in Cell culture (SILAC) nucleosome affinity purification (SNAP) experiments (47) that probe the binding of proteins from HeLa S3 nuclear extracts to a library of semi-synthetic di-nucleosomes (referred to as nucleosomes throughout the manuscript) incorporating biologically meaningful combinations of chromatin modifications representing promoter, enhancer and heterochromatin modification states. Each affinity purification measures the relative abundances of nuclear proteins on a modified nucleosome in relation to an unmodified control nucleosome using the SILAC labelling and quantitative proteomics as a read out. This allows the high-throughput identification of proteins that are either recruited or excluded by the modification(s), and also indicates the relative extent of the recruitment or exclusion. Collectively, the MARCS data set catalogs the binding responses of 1915 nuclear proteins to nucleosomes carrying 55 different modification signatures. The constructive nature of these data, paired with an appropriate statistical model, thus enables the direct analysis of combinatorial effects of different modification features on the nucleosome binding of the measured proteins. At its core, `asteRIa` uses a linear regression model with pairwise (or 'two-way') interactions among chromatin modifications to predict the binding affinities of each protein. Regression models with pairwise interactions have a long tradition in statistics and experimental design (48–50) but are notoriously difficult to estimate in the presence of noisy, scarce data and/or incomplete experimental designs, and are prone to misinterpretation (51,52). As we will show, the `asteRIa` framework incorporates several model and design principles that (i) guard against common pitfalls and (ii) take the properties of the MARCS data (and biological data in general) into account. Firstly, we posit that our framework should work in the underdetermined regime, i.e. the number of features q (here the chromatin modifications) and pairwise interactions exceeds the number of measurements n. We achieve this by including sparsity-inducing penalization of the model coefficients (53–55). Secondly, we assume that the underlying interaction model obeys the so-called 'strong hierarchy' principle (50,53,56), i.e. interactions among features are only included in the model if both features are present as main effects. Thirdly, we embrace the principle of statistical 'stability' (57–59) for model selection, implying that interactions are only included when they are reproducibly identified across subsets of the data. To respect the ubiquitous measurement variability of biological systems, we also require replicate consistency (60) of our combinatorial models. This means that models with interactions need to be (at least partially) consistent across available technical or biological replicates, further ensuring the general robustness and validity of the resulting models. While these design principles and the underlying computational workflow, available at https://github.com/marastadler/asteRIa.git, are general, we illustrate the framework to detect novel combinatorial interactions between chromatin modifications on epigenetic reader recruitment.

On the MARCS data, we show that considering interaction effects between chromatin modifications can consistently improve the predictive performance of the binding profiles of a subset of proteins. `asteRIa` not only recovers known binding patterns, such as, e.g. the well-known H3K27me3-CBX8 pairing, but also identifies novel interaction effects between chromatin modifications on the binding behavior of proteins not yet implicated as epigenetic readers (e.g. ACTL8). Our analysis also allows to define and quantify the extent of distinct modes of apparent chromatin modification interactions, ranging from synergistic and antagonistic to competitive effects. Our post-hoc model analysis shows that proteins belonging to the same protein complexes do read combinatorial chromatin modification signatures in a similar fashion, thus allowing the delineation of a protein complex - chromatin modification interaction network.

Independent confirmation of the identified combinatorial interactions is challenging due to the uniqueness of the MARCS data and the accompanying statistical analysis. Nevertheless, we provide a validation workflow on ENCODE ChIP-Seq, ChIP-Atlas ChIP-Seq and WGBS (Whole Genome Bisulfite Sequencing) data that demonstrates that our findings are not limited to a specific cell type or experimental setup. Specifically, we show that one of the found combinatorial interactions for CBX8 are consistent with these orthogonal datasets. The latter analysis also illustrates how to validate other interactions found in this study, thus inviting the generation of new ChIP-Seq data collections for previously understudied proteins.

Materials and methods

The Modification Atlas of Regulation by Chromatin States dataset

The Modification Atlas of Regulation by Chromatin States (MARCS), as introduced in (61), builds on two experimental components: (i) a designed library of engineered dinucleosomes (referred to as nucleosomes throughout the manuscript) comprising combinatorial chromatin modifications and (ii) nucleosome affinity purifications coupled to high-throughput quantitative proteomics measurements employing SILAC labeling (SNAP) (47). The modified nucleosomes were assembled from a biotinylated DNA containing two 601 nucleosome positioning sequences (62) and histone octamers containing semi-synthetic site-specifically modified histones H3.1 and H4 prepared by native chemical ligation (63). Some nucleosomes were also assembled using CpG-methylated DNA (5mC) or the histone variant H2A.Z. The complete library design matrix comprises $n_{\text{total}} = 55$ modified nucleosomes with thirteen possible chromatin modifications (see left panel of Figure 1 for a conceptual picture). The available modifications include six lysine residues on the tails of histone H3 (K4, K9, K14, K18, K23 and K27) and five on histone H4 (K5, K8, K12, K16 and K20) as well as the variant histone H2A.Z and CpG methylated (5mC) DNA on both DNA strands (symmetric methylation), respectively. The lysines are modified with acetylation (ac) or mono-, di-, or tri-methylation (me1, me2, me3). H3-5ac denotes that multiple acetylations (K9, K14, K18, K23, K27) on the tails of histone H3 are present. H4-4ac denotes that multiple acetylations (K8, K5, K12, K16) on the tails of histone H4 are present. For our computational analysis, we do not consider engineered nucleosomes that include subsets of acetylations (namely, not all five acetylations on H3 or not all four acetylations on H4) since building their mathematical products would result in perfectly collinear (thus fully redundant, and therefore not distinguishable) pair-wise interaction features (see **Interaction modeling strategy** for further clarification). Our analysis thus excludes *22* nucleosomes from the initial nucleosome library and considers a subset of $n = 33$ nucleosomes with $q = 12$ different chromatin modifications, resulting in the design matrix $L \in \{0, 1\}^{33 \times 12}$. The (transposed) design matrix L with the available combinatorial modifications is shown in the top panel ((Step 1) of Figure 2). Note that the design pattern in L does not follow any particular statistical experimental design guideline (50) but is driven by biological expertise about common modification co-occurrences.

For each modified nucleosome in MARCS, SNAP experiments are provided in two experimental ‘label-swap’ replicates of the nucleosome affinity purification process, a ‘forward’ (F) and ‘reverse’ (R) nucleosome pull-down. Nucleosomes are immobilized on streptavidin beads and incubated with nuclear extracts from HeLa S3 cells cultured either in isotopically light or heavy-labelled SILAC media. In the ‘forward’ experiments the heavy extracts are incubated with the modified and the light extracts with the unmodified nucleosome, in the ‘reverse’ experiments the extracts are exchanged. Bound proteins are eluted from the beads and identified and quantified by mass spectrometry. For each SNAP experiment the relative abundance of a given protein on the modified nucleosome is determined in relation to the unmodified nucleosome by measuring the ratios between the heavy and the light peptides (H/L ratios) identified for that particular protein (47). The H/L ratios indicate binding preferences to the modified or the unmodified nucleosomes and allow the unbiased identification of proteins that are either recruited or excluded by the modification(s) present on the modified nucleosomes. In addition, the SILAC enrichment ratios also indicate a relative ‘strength’ of the recruitment or exclusion of a given protein by the modifications. In total, the MARCS dataset comprises the binding behavior of $p = 1915$ proteins in the forward (F) and reverse (R) experiments. For our analysis, we consider the protein measurement matrices $P^F, P^R \in \mathbb{R}^{33\times1915}$ that correspond to the subset of $n = 33$ nucleosomes, described above.

Interaction modeling strategy

We aim at predicting the binding profile of each protein captured in MARCS $(P_i)_{1 \leq i \leq 1915}$ (either from the forward or reverse experiment) from the combinations of nucleosome modifications $(L_j)_{1 \leq j \leq 12}$. Given the binary design matrix L, the baseline model of uncovering (joint) additive effects of the modifications on a binding profile $Y = P_i \in \mathbb{R}^n$, $i = 1, \ldots, p$, is the linear model

$$Y = \beta_0 + \sum_{j=1}^{q} \beta_j L_j + \epsilon, \qquad (1)$$

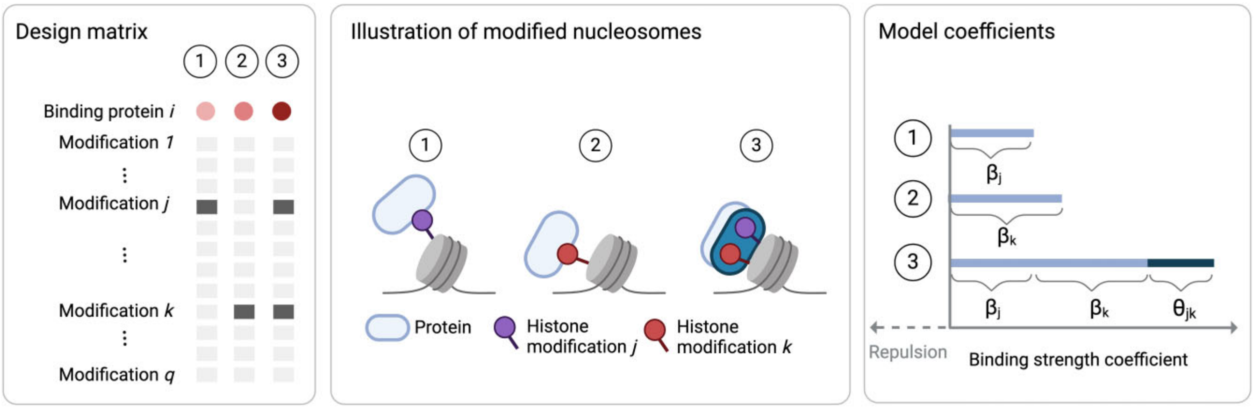


Figure 1. Left: Three exemplary columns of the design matrix *L*. Dark gray boxes indicate that a modification has been installed on the respective nucleosome. Above the design, the binding behavior of an exemplary protein to the modified nucleosomes is shown by color. The shade of red indicates the strength of the binding effect. Center: Illustration of two *individual* binding effects of chromatin modifications *j* and *k* on a protein P_i (1 and 2). Synergistic *combinatorial* effect of the modifications *j* and *k* on protein P_i (dark blue) compared to expected binding effect under independence of modification *j* and *k* (light blue) (3). Right: Model coefficients/estimated binding strength of protein P_i for the three scenarios. Light blue bar in scenario 3 shows the binding strength under independence of modification *j* and *k*, $\beta_j + \beta_k$. Dark blue shows the additional combinatorial effect θ_{jk} that goes beyond additive combinatorial effects (created with BioRender.com).

where $\beta_0 \in \mathbb{R}^n$ is a protein-specific (constant) intercept, β_j is the effect of modification *j* on the binding profile $Y = P_i$ of protein *i*, and ϵ models the technical and biological noise component. In (61), a simplified version of this baseline model was investigated through 'feature effect estimates' via pairwise comparisons of the enrichments of individual proteins on nucleosomes differing by a single modification feature. This, however, only allowed robust prediction of the effects of individual modifications or blocks of modifications and did not provide any information on combinatorial effects. Here, we extend the baseline model by including all pairwise interactions between modifications. For each protein binding profile $Y = P_i$, $i = 1, \ldots, p$, the core model in `asteRIa` thus reads

$$Y = \beta_0 + \sum_{j=1}^{q} \beta_j L_j + \frac{1}{2} \sum_{j=1}^{q} \sum_{k=1}^{q} \Theta_{jk} L_j L_k + \epsilon\,, \qquad (2)$$

where Θ_{jk} models interaction effects between epigenetic readers that cannot be captured by linear additive effects. Robustly and reproducibly estimating non-zero entries in the interaction matrix Θ from replicated data is at the heart of the `asteRIa` workflow. The sign of the interaction coefficients also allows a characterization of epigenetic reader interplay. For example, when $\hat{\Theta}_{jk} > 0$ we interpret the two modifications *j* and *k* to have a synergistic binding effect if both $\beta_j > 0$ and $\beta_k > 0$ (see Figure 1 for illustration).

To guarantee identifiability and interpretability of individual interaction models, we first need to ensure that the interaction design matrix $L_j L_k$ has no co-linear columns. In the concrete example of the MARCS data, we group modifications of the complete design matrix to a set of $n = 33$ non-redundant nucleosomes (see top panel (Step 1) of Figure 2). Secondly, to enable estimation in the present underdetermined regime ($q(q + 1)/2 > n$) with $q(q + 1)/2 = 78$, we perform regularized maximum-likelihood estimation with ℓ_1-norm (lasso) penalization (64) on the linear and interaction coefficients, respectively. Given the log-likelihood function of the model $l(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - L\beta - \frac{1}{2} L\Theta L^T \right\|_2^2$, the (all-pairs) lasso problem reads

$$\min_{\beta_0, \beta, \Theta} \; l(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \tfrac{\lambda}{2} \|\Theta\|_1\,, \qquad (3)$$

where $\lambda > 0$ is a tuning parameter and controls the sparsity levels of the coefficients β and Θ, respectively. To further ease model interpretability, we follow the statistical principle of hierarchy (also known as marginality or heredity) and allow the presence of an interaction in the model *only if* the associated linear (main) effects are in the model as well (see (53), and references therein). In mathematical terms, this so-called strong hierarchy principle can be expressed as

$$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0 \text{ and } \hat{\beta}_k \neq 0,$$

implying that interaction effects are only present if both linear effects enter the model. This hierarchy can be achieved by adding a constraint on the interaction effects $\Theta_j \in \mathbb{R}^q$ and a symmetry constraint on Θ. The corresponding optimization problem with hierarchical interactions thus reads

$$\begin{aligned} &\min_{\beta, \Theta} \; l(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \tfrac{\lambda}{2} \|\Theta\|_1 \\ &\text{s.t. } \Theta = \Theta^T, \quad \left\|\Theta_j\right\|_1 \leq |\beta_j|\,. \end{aligned} \qquad (4)$$

To solve the non-convex optimization problem in (4), we follow Bien et al. (53) who proposed a convex relaxation of the problem and provide an efficient implementation in the corresponding R package `hierNet` (65) (v1.9). In `asteRIa`, we use `hierNet` to model each protein binding profile $Y = P_i$, $i = 1, \ldots, p$ with hierarchical interactions. Apart from reducing the number of spurious interaction effects, a major advantage of the strong hierarchy constraint is the so called 'practical sparsity'. The strong hierarchy constraint favors models that 'reuse' measured variables. In the context of the MARCS data, this becomes important when generating hypotheses for follow-up functional analysis (where experiments are complex and costly). Concretely, our models assumes that a protein or protein complex must have a domain capable of recognizing a particular chromatin modification. Thus, if there exists a response of a protein to an interaction effect between two modifications, a (possibly small) linear effect to both modifications is expected.

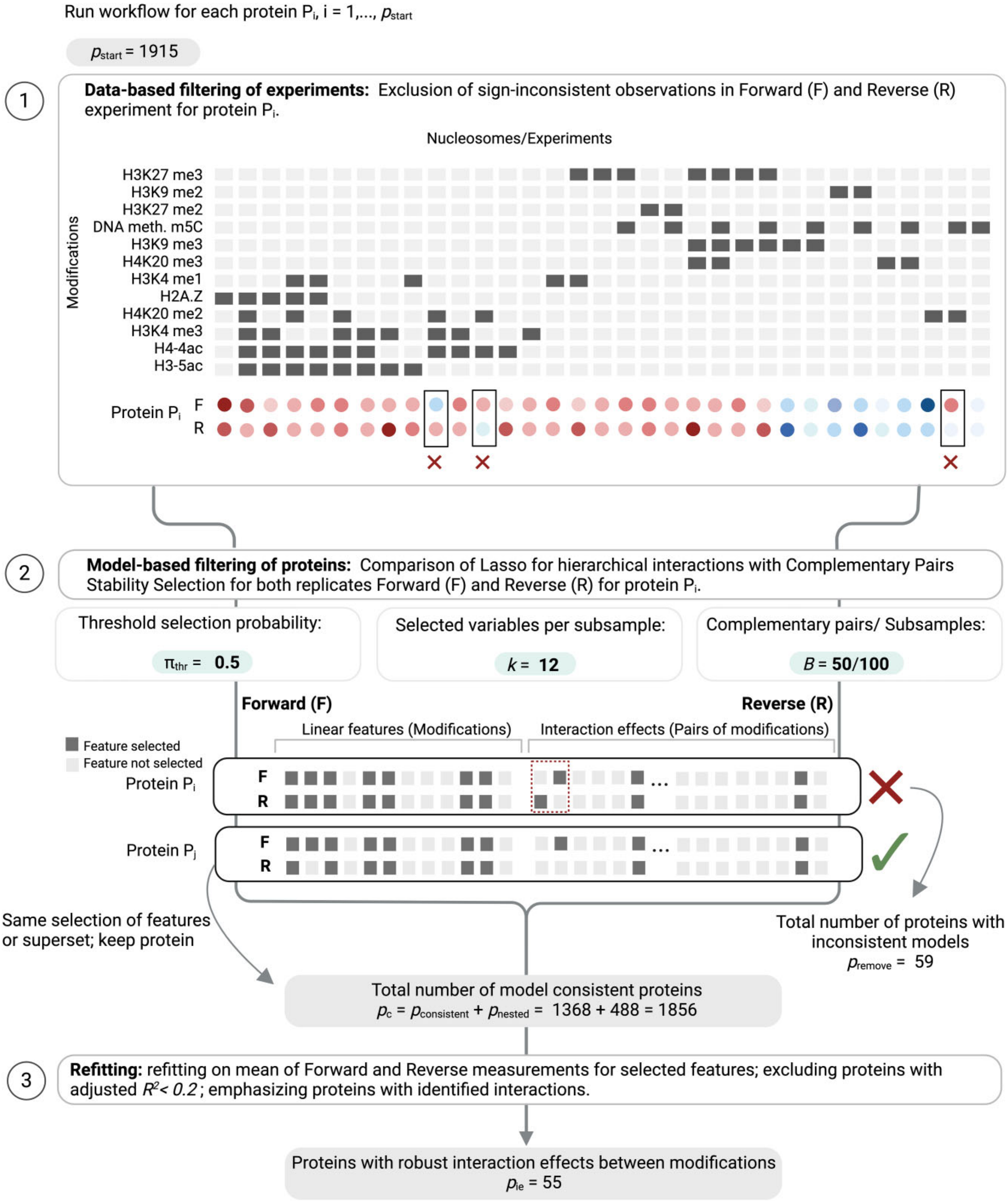

Run workflow for each protein P_i, i = 1,..., p_{start}
p_{start} = 1915
1
Data-based filtering of experiments: Exclusion of sign-inconsistent observations in Forward (F) and Reverse (R) experiment for protein P_i.
Nucleosomes/Experiments
Modifications
H3K27 me3
H3K9 me2
H3K27 me2
DNA meth. m5C
H3K9 me3
H4K20 me3
H3K4 me1
H2A.Z
H4K20 me2
H3K4 me3
H4-4ac
H3-5ac
Protein P_i
F
R
2
Model-based filtering of proteins: Comparison of Lasso for hierarchical interactions with Complementary Pairs Stability Selection for both replicates Forward (F) and Reverse (R) for protein P_i.
Threshold selection probability:
π_{thr} = 0.5
Selected variables per subsample:
k = 12
Complementary pairs/ Subsamples:
B = 50/100
Forward (F)
Reverse (R)
Linear features (Modifications)
Interaction effects (Pairs of modifications)
Feature selected
Feature not selected
Protein P_i
Protein P_j
Same selection of features or superset; keep protein
Total number of proteins with inconsistent models
p_{remove} = 59
Total number of model consistent proteins
p_c = $p_{consistent}$ + p_{nested} = 1368 + 488 = 1856
3
Refitting: refitting on mean of Forward and Reverse measurements for selected features; excluding proteins with adjusted R^2 < 0.2 ; emphasizing proteins with identified interactions.
Proteins with robust interaction effects between modifications
p_{ie} = 55

Figure 2. Graphical representation of the asteRIa workflow for the robust detection of hierarchical interactions (created with BioRender.com). Step 1: Design matrix and measured binding behavior for a protein (two replicates F and R). Removal of observations with different signs in the replicates (sign consistency). Step 2: Hierarchical interaction modeling with default complementary pairs stability selection (CPSS) parameters. Comparison of selected features for each replicate (nested model consistency): The first example shows a protein prediction model that gets filtered out since the selected features from the forward and reverse replicate are neither identical nor nested. The second example shows a 'consistent' protein model where the selected features learned from the reverse replicate is a nested subset of the features learned from the forward replicate. Step 3: Least-squares refitting on averaged replicate data for final prediction model building. The intersection of two selected feature sets is used for refitting. Models with adjusted R^2 < 0.2 are discarded.

Stability-based model selection for hierarchical interactions

One of the core challenges in high-dimensional penalized regression is determining a suitable regularization parameter λ that trades off sparsity (i.e. interpretability) of the model coefficients and out-of-sample predictive performance of the model (66,67). Standard procedures for (hierarchical) interaction models include cross-validation (53) and Information Criteria, including the Akaike (AIC) and the extended Bayesian Information Criterion (BIC) (55). However, it has been observed that, both in simulation and practice, cross-validation and Information criteria tend to select more predictors (and interactions) than necessary (55).

To address this shortcoming, we follow the principle of stability (57) in `asteRIa` and introduce stability selection (58) for the identification of a reproducible set of predictive features *and* interactions. Stability selection has been proven useful across several scientific applications, ranging from network learning (68,69) to data-driven partial differential equation identification (70,71). In the regression context, stability selection repeatedly learns sparse regression models from subsamples of the data of fixed size (e.g. $n_s = \lfloor n/2 \rfloor$), records the frequency of all selected predictors across the models, and selects the most frequent predictors to fit the final regression model. Here, we use a variant of stability selection, the so-called complementary pairs stability selection (CPSS) (59) which draws B subsamples as complementary pairs $\{(A_{2h-1}, A_{2h}): h = 1, \ldots, B\}$, with $A_{2h-1} \cap A_{2h} = \emptyset$ of samples $\{1, \ldots n\}$ of size $\lfloor n/2 \rfloor$. Drawing complementary pairs is particularly beneficial when dealing with unbalanced experimental designs, as the resulting random splits ensure that individual subsamples are independent of each other. After applying a variable selection procedure S (e.g. using the first k predictors that enter the penalized model), each feature j in the model gets an individual estimated selection probability $\hat{\pi}(j)$, given by

$$\hat{\pi}(j) = \frac{1}{2B} \sum_{h=1}^{2B} \mathbb{1}_{\{j \in \hat{S}(A_h)\}}, \qquad (5)$$

and the final selection set is given by $\hat{S}^{\text{CPSS}} = \{j : \hat{\pi}(j) \geq \pi_{\text{thr}}\}$, for a threshold π_{thr} defining the minimum selection frequency. In our workflow we use the corresponding R package `stabs` (72) (v0.6-4) that provides an efficient implementation of CPSS. The CPSS procedure includes the following hyperparameters: The set of regularization parameters Λ, a threshold $\pi_{\text{thr}} \in [0, 1]$, the number of predictors k that first enter the sparse model, and the number of complementary splits B. In `asteRIa`, we set as default parameters Λ to be the internal λ-path in Bien and Tibshirani (65), $\pi_{\text{thr}} = 0.5$, $k = 12$ and $B = 50$, resulting in 100 subsamples. For the MARCS data, this means that chromatin modifications (as main or interaction effects) are part of the pairwise interaction model 2 for protein binding profile i, $Y = P_i$, if it is among the $k = 12$ selected modifications in at least 50 subsamples. While these default values may need to be tuned in other scenarios, we verified in a realistic semi-synthetic simulation scenario (see Supplementary information and Supplementary Figures S1 and S2 for details) that hierarchical interaction modeling with stability selection greatly outperforms cross-validation, particularly in terms of false positive rate.

Replicate consistency

Biological datasets typically include replicated measurements (replicates) to probe different sources of variability in the underlying experimental procedure or study object (73). The MARCS dataset, for example, comprises two technical replicates of the SILAC-based protein binding affinities. Replicate consistency, i.e. assessing how consistent two or multiple replicated measurements are in terms of direction or size, is an important property to evaluate experimental protocols and downstream analysis quality (see, e.g. (74) for a discussion in the context of RNA sequencing data).

In `asteRIa`, we propose and include two replicate-consistency mechanisms: (i) data sign-consistency and (ii) nested model consistency. While there are alternative ways of performing filtering, data sign-consistency can be considered as a data filtering step that ensures that replicated measurements agree on the direction, i.e., the sign of the measured unit, and removes experiments where sign consistency does not hold. In MARCS, we perform data sign consistency for each protein P_i separately using the forward and reverse replicates (see Figure 2, Step 1) and remove nucleosomes (experiments) where measured protein binding affinities disagree in sign. Although this reduction in sample size (for each protein $n_i \leq n$ samples are available) decreases the power for subsequent hierarchical interaction modeling, the filtering increases the chance of estimating pairs of consistent interaction models. In a second post-hoc step, nested model consistency further ensures that only pairs of consistent interaction models are considered for downstream analysis. Nested model consistency deems estimated interaction models valid only if they comprise the same set of features (main and interaction coefficients) across replicates *or* one model comprises a nested subset of main and interaction effects of the other model (see Figure 2, Step 2, for illustration).

The `asteRIa` workflow

The `asteRIa` workflow incorporates the described model and design principles as illustrated in Figure 2 on the MARCS data. `asteRIa` comprises three main steps: Step (1) uses sign consistency to filter pairs of forward and reverse experiments for each protein $(P_i)_{1 \leq i \leq p = 1915}$. Step (2) comprises model estimation using the hierarchical interaction model, CPSS-based model selection, and the post-hoc nested model consistency filter. Step (3) performs least-squares 'refitting' to estimate main and interaction effect sizes on the selected model coefficients from averaged replicate data. The resulting signed model coefficients are then used for functional categorization and downstream analysis.

On the MARCS data, the experiment filtering step (1) removes on average 11 experiments across all proteins. In step (2), using the internal λ-path in Bien and Tibshirani (65), and CPSS parameters $\pi_{\text{thr}} = 0.5$, $k = 12$, and $B = 50$, `asteRIa` learns $p_{\text{consistent}} = 1368$ fully consistent regression models across forward and reverse replicates, as well as $p_{nested} = 488$ models that obey the nested model consistency criterion. Only $p_{\text{remove}} = 59$ models are inconsistent across replicates. Among all $p_{\text{c}} = 1856$ consistent models, `asteRIa` identifies 58 models that include robust interaction coefficients. The refitting estimation process in step (3) uses the averaged binding affinities as outcome and performs least-squares refitting on the *intersection* of the per-replicate selected features. The refit coefficients are the final effect sizes. For downstream analysis, asteRIa removes poorly-performing prediction models with adjusted R^2 below 0.2 (three out of 58).

Results

Enhanced predictive performance of protein binding through chromatin modification interaction

We first quantify the overall predictive performance of asteRIa models for all proteins included in the MARCS dataset and then assess the degree to which hierarchical interaction modeling improves overall predictive performance of protein binding affinities. For a majority of the p=1915 protein binding profiles, asteRIa deems main effects models (i.e., the baseline linear model in 1) to be sufficient for robust prediction. For more than 200 proteins, main effects models achieve adjusted $R^2 > 0.8$, and for more than 500 proteins, main effects models achieve adjusted $R^2 > 0.5$ (see Supplementary Figure S4 for a list of top protein binding models and associated coefficients). The top-six protein binding models achieve near-perfect predictive performance and include the protein ING5, a dimeric, bivalent reader of histone H3K4 me3 (75), with an $R^2 = 0.99$, the methyl–lysine histone-binding protein L3MBTL3 ($R^2 = 0.99$), SMARCC2 ($R^2 = 0.99$) which is part of the chromatin remodeling complex SNF/SWI, the histone acetyltransferase KAT7 ($R^2 = 0.98$), the YAF2 protein ($R^2 = 0.98$), and the histone lysine demethylase KDM2B ($R^2 = 0.98$).

However, asteRIa also identifies a set of $p_{\text{ie}} = 55$ models that comprise stable interaction effects among modifications with enhanced predictive performance. This provides statistical evidence that cooperative effects between chromatin modifications may play a crucial role in the binding of specific reader proteins and thus in controlling chromatin function. Figure 3A shows the modification design matrix (left panel) and binding profiles (both the 'forward' and the 'reverse' experiments) of the 55 proteins explained by interaction models. The proteins are sorted by data density (i.e., in terms of number of experiments removed due to sign consistency filtering step (1) in asteRIa, Figure 3A, gray boxes). Figure 3C shows the corresponding predictive performance of the models in terms of adjusted R^2 both for main effects (light blue) and interaction models (dark blue), respectively. While the light blue segment denotes the proportion of variance explained by all selected main effects combined, the dark blue portion represents the additional explained variance attributed solely to one interaction. We observe that the inclusion of robust interaction among modifications can boost the performance of up to 0.5 (e.g., for proteins CDKAL1 and PEX11B). For others, such as, e.g., RFC3, the binding behavior can only be sufficiently described by taking into account interaction effects. While the improvement is less dramatic for proteins with well-performing main effects models, asteRIa still provides evidence for stable interactions among modifications. Figure 3B illustrates the stabilities (inclusion probabilities) $\hat{\pi}$ of all model coefficients for the protein CBX8. On both forward and reverse experimental data, asteRIa estimates a high selection probability (≈0.7) of an interaction effect between DNA methylation m5C and H3K9me3 while all other interaction effects emit a low inclusion probability. For detailed model inspection, we provide similar stability plots for all other proteins in the Supplementary Material. To illustrate the improvement in binding prediction, Figure 3C (right panel) shows predicted vs. observed binding profiles for the protein RNF2. Comparison of the fits of both the main effect (light gray) and interaction model (dark blue) visually and quantitatively ($R^2 = 0.76$ versus $R^2 = 0.9$) confirm the enhanced predictive performance of the interaction model.

Modes of chromatin modification interactions

To categorize the interaction effects uncovered in asteRIa, we establish potential modes of chromatin modification interactions. This is achieved by contrasting the effects of individual chromatin modifications (modification *j* and *k*) on the binding behavior of specific proteins, represented by the linear model coefficients $\hat{\beta}_{\text{j}}$ and $\hat{\beta}_{\text{k}}$ with the combinatorial effects identified during our analysis, represented by $\hat{\Theta}_{\text{j,k}}$ for the corresponding pair (see Figure 4A and B). We define three major modes: synergistic combinatorial behavior, antagonistic combinatorial behavior, and conflicting combinatorial behavior. We further divide these into two sub-modes each of which describes the direction of the combinatorial effect, either towards binding (b, $\Theta_{j,k} > 0$) or towards repulsion (r, $\Theta_{j,k} < 0$). The direction and strength of the combinatorial effect is color-coded in Figure 4B.

The 'Synergy b+b+b' category (shown in blue in Figure 4) includes proteins that bind to two modifications individually and exhibit particularly strong binding, i.e., stronger than the sum of the two individual effects when both modifications are present. For example, we uncover that UHRF1 (Figure 4B, 1st quadrant) responds in a synergistic way to an interaction effect between DNA methylation m5C and H3K9me3. UHRF1 is a RING-type E3 ubiquitin ligase that plays an essential role in DNA methylation by mediating the recruitment of the maintenance DNA methyltransferase DNMT1 (76). UHRF1 is known to bind to H3K9me3 via a tandem tudor domain and to recognize hemi-methylated DNA via a SRA domain. Our analysis therefore validates previously known binding behaviors and, additionally, unveils that there is a true synergistic effect between H3K9me3 and DNA methylation in the recruitment of UHRF1.

For the maintenance DNA methyltransferase DNMT1 (77), we identify an individual binding effect to H3K9me3 and a modest individual binding to DNA methylation. Furthermore, we also identify an interaction effect between DNA methylation m5C and H3K9me3. In this case, however, the addition of DNA methylation m5C leads to a reduction in binding of DNMT1 to H3K9me3. We define this behavior as 'Antagonism b+b+r' or preferential binding (pink category in Figure 4). UHRF1 and DNMT1 were found to interact with each other (see references in (76)), and binding of DNMT1 to H3K9me3 is likely mediated through UHRF1 (see above). Both UHRF1 and DNMT1 are flexible multi-domain proteins, that consist of several different domains and can change their shape or structure. They are involved in a complex network of interactions, both within themselves (intra-molecular) and with each other (inter-molecular). This network helps control their function through allosteric regulation events involving conformational rearrangements of autoinhibitory domains (changes in the structure of certain domains within the proteins) in both molecules (76,78). The antagonistic effect of DNA methylation on the recruitment of DNMT1 to H3K9me3 indicates that while symmetric DNA methylation stimulates binding of UHRF1 to the doubly modified nucleosomes (see above), it disrupts the interaction with DNMT1.

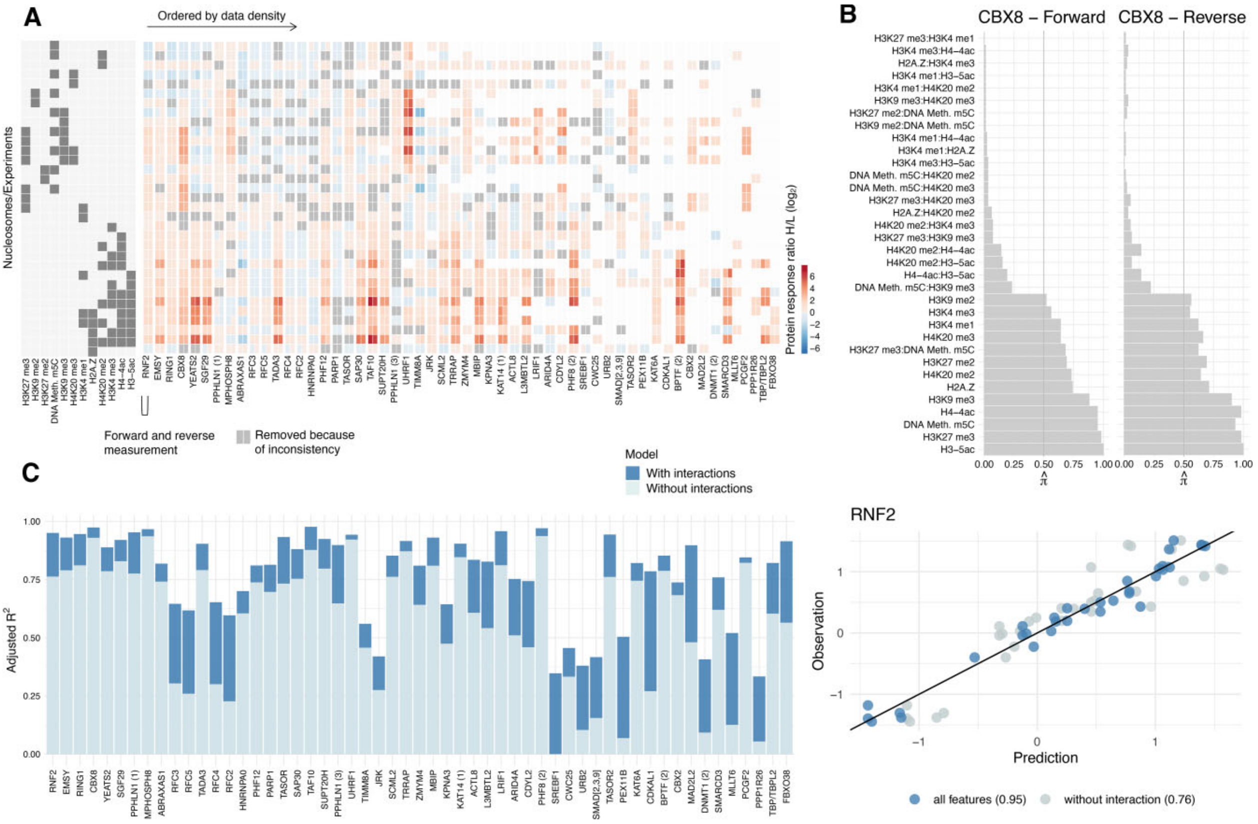


Figure 3. (**A**) Observed protein binding profiles (forward and reverse experiment) for the $p_{ie} = 55$ proteins for which interactions between modifications have been detected. Proteins are arranged from left to right based on data density (number of non-zero measurements), with proteins with the highest data density being on the left. (**B**) Stability plots for CBX8 of the hierarchical interaction model with complementary pairs stability selection (CPSS). Vertical lines show the threshold for the selection probability threshold $\pi_{\text{thr}} = 0.5$. Stability plots for all proteins are provided in Extended (B). (**C**) Adjusted R^2 for all $p_{ie} = 55$ proteins of the main effect (light blue) and interaction model (dark blue) (left panel). Scatter plot of observed vs. predicted values for the protein RNF2 (right panel). Scatter plots for all proteins are provided in Extended (C).

This suggests a mechanism within DNMT1 that senses symmetrically methylated DNA (the end product of the DNA methylation reaction) and triggers the release from chromatin upon completion of its enzymatic reaction. Apart from the catalytic domain of DNMT1, which is responsible for the main activity of the protein, this observed behavior could involve a CXXC domain that has a special ability to bind to certain DNA sequences, specifically sequences that contain unmethylated CpG nucleotides, and could contribute to sensing the DNA methylation status.

Two proteins, MAD2L2 and ACTL8, exhibit a similar behavior with respect to DNA methylation m5C and H3K9me3. However, for these proteins, DNA methylation m5C exhibits a slight repulsive effect on its own. These proteins belong to the category ‘Conflict, dominated by repulsion b+r+r’ (grey category in Figure 4).

Proteins in the ‘Conflict, dominated by binding b+r+b’ category (yellow category in Figure 4) are repelled by one modification and bind to another modification if they are considered individually. In combination, these modifications show a stronger binding effect on the protein than expected under additivity. The chromodomain-containing protein CBX8, which is a component of the polycomb repressive complex 1 (PRC1) (79), also falls into this category. Our analysis reveals that DNA methylation m5C enhances the binding of CBX8 to H3K27me3, while DNA methylation m5C itself exhibits a slight repulsive effect on CBX8. The association of CBX8 with both DNA and H3K27me3 has been investigated in Connelly et al. (80). Here, the authors identified a dual interaction mechanism for the CBX8 chromodomain, where the engagement of both DNA and H3K27me3 mediates the association of CBX8 with chromatin. Similar binding behaviors are observed for the PRC1 subunits RNF2 and RING1. However, in contrast to CBX8, RNF2, and RING1 are shared among multiple complexes, including the canonical polycomb repressive complex 1 (PCR1) and various non-canonical versions of the complex (ncPRC) (79). The nucleosome binding profiles of these shared subunits reflect a superposition of the binding profiles of all the complexes they are associated with. This introduces additional complexity to the interpretation of combinatorial effects.

Chromatin modification interaction in the recruitment of proteins and complexes

Our analysis suggests that proteins within the same protein complex tend to exhibit similar binding patterns not only to individual chromatin modifications, but also with regards to interaction effects of modifications.

Our analysis reveals seven distinct combinations of chromatin modifications demonstrating a robust combinatorial effect on the shortlisted 55 proteins (see Figure 5A). While six of the discovered interactions affect multiple proteins, H2A.Z incorporation appears to interact solely with H4K20me2,

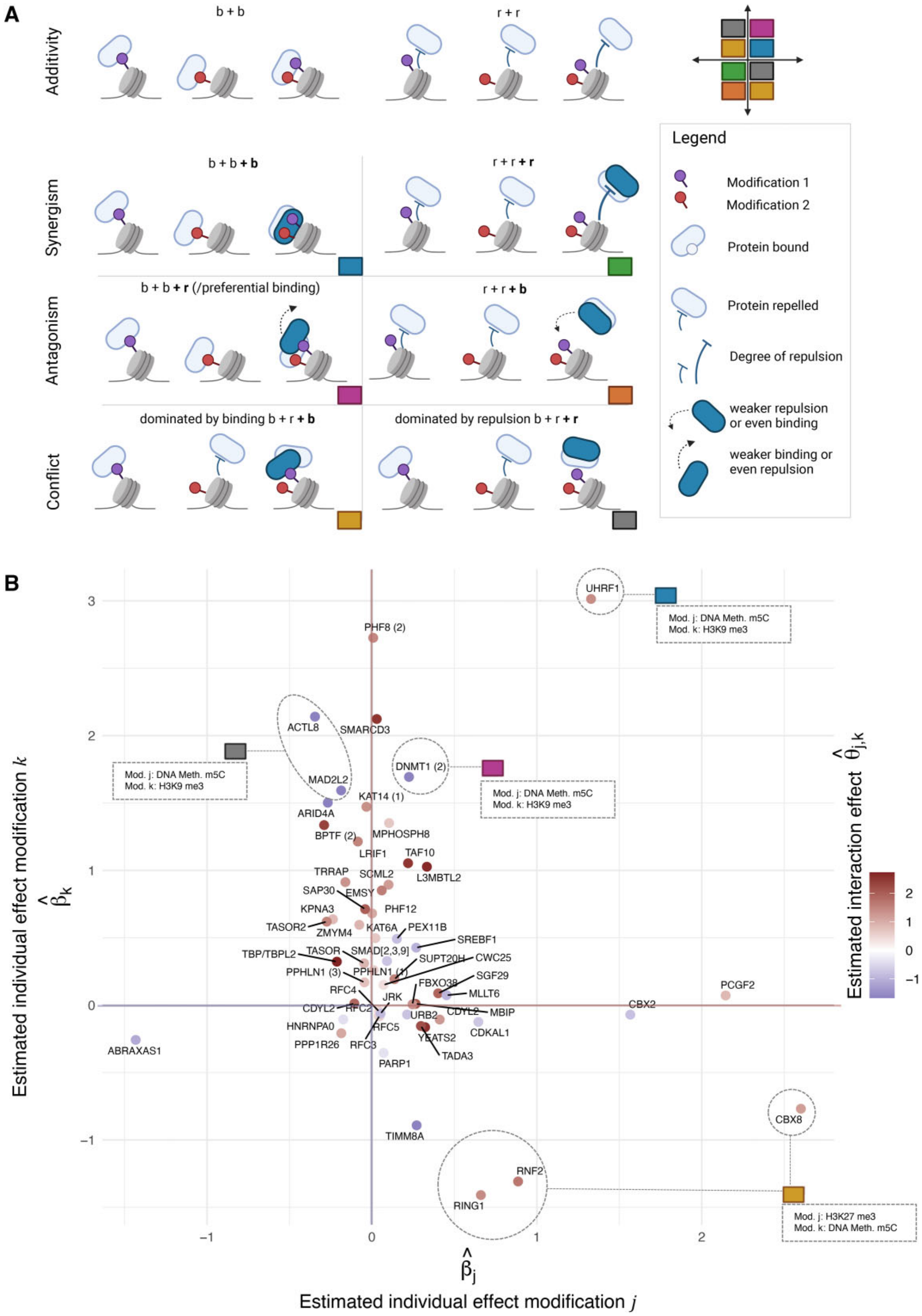


Figure 4. (**A**) Six combinations of combinatorial interaction effects of two modifications on the binding behavior of chromatin-associated proteins (created with BioRender.com). The top row illustrates what would be expected under a purely additive dependence of the effects in a scenario where a protein shows *individual* binding effects to two distinct chromatin modifications (overall *additive* effect is the sum of both individual binding effects, b + b) (left) and where a protein is repelled by two distinct chromatin modifications (right). The rows below shows different modes of deviations due to (directional) interaction effects. (**B**) Scatter plot of protein binding effects with unspecific linear effects β_j and β_k on the *x*- and *y*-axis and corresponding interaction effect $\Theta_{j,\,k}$ represented by color. For some example proteins detailed information is provided.

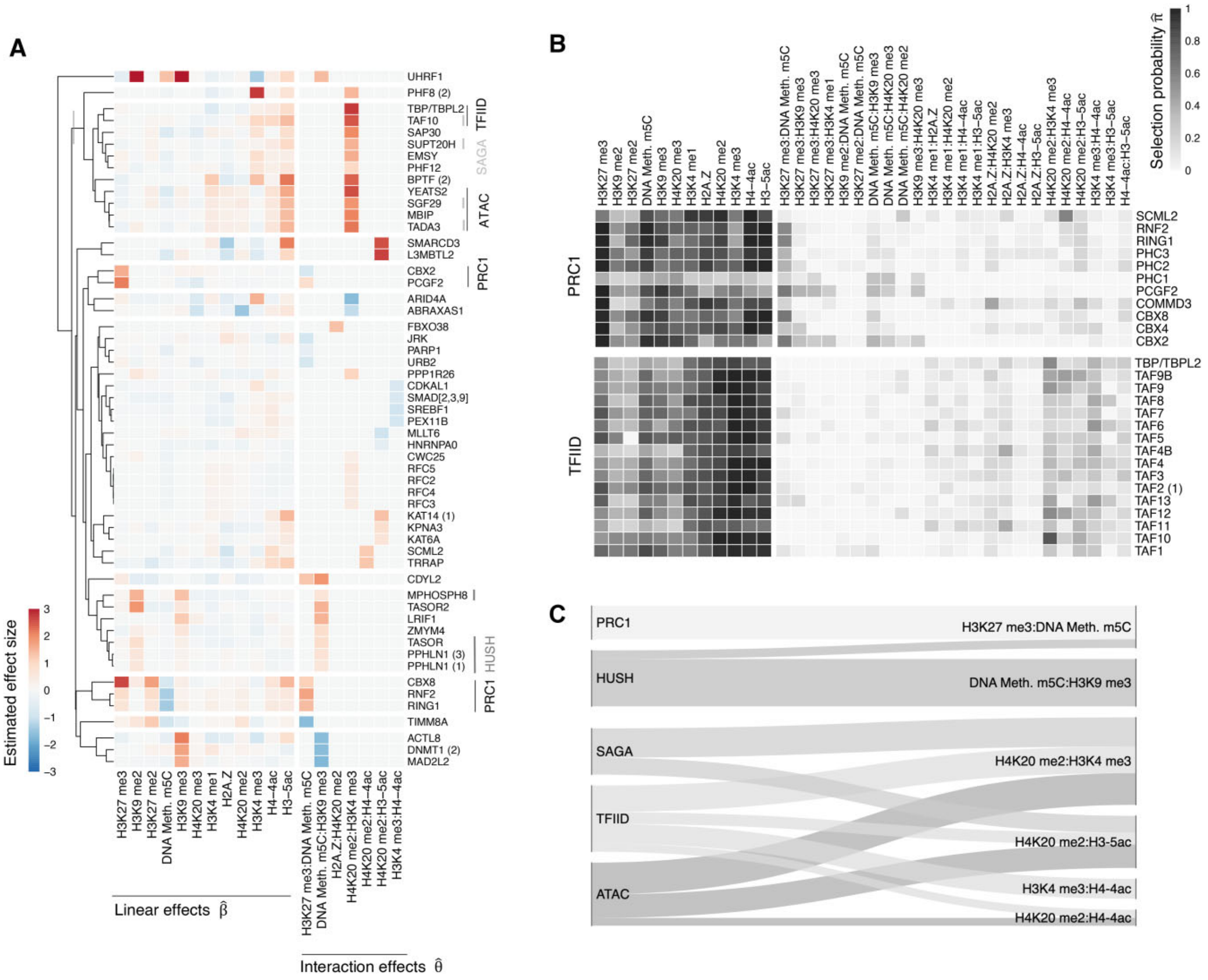


Figure 5. (**A**) Clustered representation of robustly estimated linear and interaction coefficients for $p_l = 55$ proteins for which interaction coefficients have been identified. Proteins belonging to notable protein complexes are highlighted. (**B**) Selection probabilities $\hat{\pi}$ for all proteins in the TFIID and PRC1 complexes. Selection probability plots for all protein complexes are provided in Extended (B). (**C**) Sankey diagram of mean selection probabilities (>0.2) for interaction effects for proteins within a complex.

influencing only the protein FBXO38. Given FBXO38's notably low data density (see Figure 3A, last column), we did not investigate this interaction further. Notably, TAF10 and TBP, which are both part of the Transcription Factor II D (TFIID) complex, respond similarly to the combination of H4K20me2 and H3K4me3. Similarly, members of the PRC1 complex, such as CBX8, RNF2, RING1, CBX2, and PCGF2, are found to respond to the combination of H3K27me3 and DNA methylation m5C (Figure 5A). In addition to examining the effect sizes obtained from asteRIa, our approach allows for the interpretation of protein-specific selection probabilities for each individual chromatin modification and each interaction between chromatin modification combinations. The selection probability indicates how stable a feature is in predicting a proteins binding profile across subsamples.

We observe that proteins belonging to the same complex show similar modification selection probability patterns (see Figure 5B for an illustration using the TFIID and PRC1 complex, respectively). These similarities in selection probability patterns justify the exploration of mean selection probabilities over proteins within the same complexes, leading to a more general analysis of how entire complexes respond to interaction effects between chromatin modifications (see Figure 5C).

One major discovery is that the Ada-Two-A-containing (ATAC), Spt-Ada-Gcn5 acetyltransferase (SAGA), and TFIID complexes exhibit multiple combinations of co-operative chromatin modifications that stimulate their binding (Figure 5C). Notably, our analysis reveals several interactions where H4K20me2 is involved, particularly in conjunction with H3K4me3, H3-5ac and H4-4ac.

SAGA is a highly conserved transcriptional co-activator with four distinct functional modules. Its enzymatic functions, including histone acetylation and deubiquitination modules, play crucial roles in chromatin structure and gene expression (81). The ATAC complex, which shares subunits with SAGA, also exhibits histone acetyltransferase activity (81). TFIID, another essential transcription factor, is also a histone acetyltransferase, but additionally recognizes core promoter sequences, recruits the transcription pre-initiation complex, and interacts with SAGA subunits. TFIID contributes to transcription initiation and gene expression by collaborating with cofactors, gene-specific regulators, and chromatin modifications associated with active genomic regions (82). As such ATAC, SAGA and TFIID are all protein complexes that possess activities that are intricately involved in the process of transcription initiation and that thereby contribute to the regulation of chromatin structure and gene expression.

H4K20me2 is a pervasive modification found on 80% of all histone H4 proteins, marking nearly every nucleosome throughout the genome. Since newly incorporated histone H4 is unmodified at K20 (H4K20me0), the H4K20me2 modification serves as a marker of not yet replicated ‘old’ chromatin, while H4K20me0 marks newly replicated chromatin during the cell cycle. This modification is used by the DNA repair machinery to determine between different DNA repair pathways in different cell cycle phases (83). The synergistic effect between H4K20me2 and active modifications in recruiting protein complexes associated with transcriptional initiation is therefore surprising and hints to a so far unknown possible function of this modification in the context of promoter regulation.

In contrast, members of the repressive PRC1 and HUSH complexes show a response to an interaction effect between H3K27me3 and DNA methylation m5C and an interaction effect between DNA methylation m5C and H3K9me3, respectively.

The human silencing hub (HUSH) complex is well-established for its role in transcriptionally repressing long interspersed element-1 retrotransposons (L1s) and retroviruses through the modification of histone H3 lysine 9 trimethylation (H3K9me3) (84). Our analysis not only confirms H3K9me3 to be an important binding determinant, in line with previous findings, but it also reveals the involvement of DNA methylation m5C in this regulatory process. Furthermore, our analysis uncovers a previously unreported synergistic interaction between these two modifications, indicating a more complex interplay between H3K9me3 and DNA methylation m5C than previously known.

As a last example, we find that for several members of the PRC1 complex, there is an increased likelihood of responding to an interaction between H3K27me3 and DNA methylation m5C, as previously discussed for CBX8 and the subunits RNF2 and RING1. The PRC1 complex is known to be capable of recognizing H3K27me3 and facilitating transcriptional repression (79), while there are no known associations between the PRC1 complex and methylated DNA. Our results suggest a distinct behavior of DNA methylation and H3K27me3 on regulating the recruitment of the PRC1 complex, with DNA methylation m5C having minimal or even a slightly repulsive effect and H3K27me3 having a binding effect on their own. However, in combination, our analysis reveals an interaction between these two modifications that enhances binding.

Validation of the effects of H3K27me3 and DNA methylation on the binding of CBX8 with ChIP-seq and WGBS data

To validate and compare our findings with orthogonal data sources, we leverage publicly accessible ChIP-seq and WGBS (Whole Genome Bisulfite Sequencing) datasets from the ENCODE project (https://www.encodeproject.org) (40,85–87) and ChIP-Atlas (https://chip-atlas.org) (88–90). Specifically, we design a validation workflow that compares partial correlations from modification co-occurrence patterns with `asteRIa`’s linear and interaction coefficients.

Given the unique design of the MARCS data, our ability to independently validate our discoveries hinges on the availability of ChIP-seq/WGBS experiments that encompass chromatin modifications for which we have identified interaction effects *and* are available in the same cell type. After a comprehensive search, we have identified only the trio of H3K27me3 (ChIP-seq), methylated DNA (WGBS), and the CBX8 protein (ChIP-seq) as the only adequate data set.

As previously described, `asteRIa` reveals a modest interaction effect between H3K27me3 and methylated DNA concerning the binding of CBX8 in the nucleosome binding data. This interaction effect is categorized as ’conflict, dominated by binding b+r+b’ (see Figure 4A). We detect a slight repulsive effect of methylated DNA on CBX8 and a recruitment to H3K27me3. Notably, we identify an additional positive interaction effect on CBX8 binding when methylated DNA and H3K27me3 co-occur. Consequently, our results indicate a subtle enhancing effect on CBX8 binding when methylated DNA co-occurs with H3K27me3 (see Figure 4B, lower right corner), resulting in improved predictive accuracy (see Figure 3C).

For this combination, we found matching ChIP-seq and WGBS experiments in A549 (human lung carcinoma epithelial cells), K562 (human myelogenous leukemia cells), and H1 cells (human embryonic stem cells) on ENCODE. Additionally, we use mES cell (mouse embryonic stem cells) data from ChIP-Atlas. For these four cell types, we perform the following analysis workflow: (i) We calculate averages of WGBS data and averages of fold-change values to a reference genome in the ChIP-seq data within consecutive genome bins of 1000 base pairs (bp) with no spacing between bins. We accomplish this by utilizing the ‘bins’ mode within deeptools on the Galaxy web platform (91), and we ensure the exclusion of blacklisted regions (hg38 for A549, K562 and H1 cells and mm9 for mES cells) during these calculations. (ii) We then conduct a genome-wide analysis of the behavior of H3K27me3 and methylated DNA in CBX8 peak regions. We observe increased H3K27me3 fold-changes and simultaneously decreased DNA methylation values (decreased in K562, A549 and mES; unaffected in H1) in CBX8-bound regions across all cell types under investigation (see Figure 6A and Supplementary Figure S3). This substantiates the (linear) dependencies identified in the `asteRIa` workflow. (iii) We compute Kendall’s partial correlations (92) (package version v1.1) of the genome-wide co-occurrence patterns between CBX8, methylated DNA, H3K27me3, and the ‘interaction’ between methylated DNA and H3K27me3 (i.e. the product of WGBS and H3K23me3 ChIP-seq values, denoted by H3K27me3:WGBS). We use this rank-based correlation coefficient to account for the fact that WGBS and ChIP-seq data are measured and interpreted on different scales. The resulting partial correlations patterns are shown in Figure 6B. The interpretation of the partial correlation coefficients aligns with the coefficients in `asteRIa`’s interaction model. Specifically, the first column of each partial correlation matrix (CBX8) can be understood as follows. The partial correlation between CBX8 and H3K27me3, as well as between CBX8 and the WGBS abundances, reflects the individual (linear) effects of these modifications on CBX8 binding (after conditioning on all other effects). We observe that they are (moderately) positive for CBX8 and H3K27me3 across all cell types, and nega-

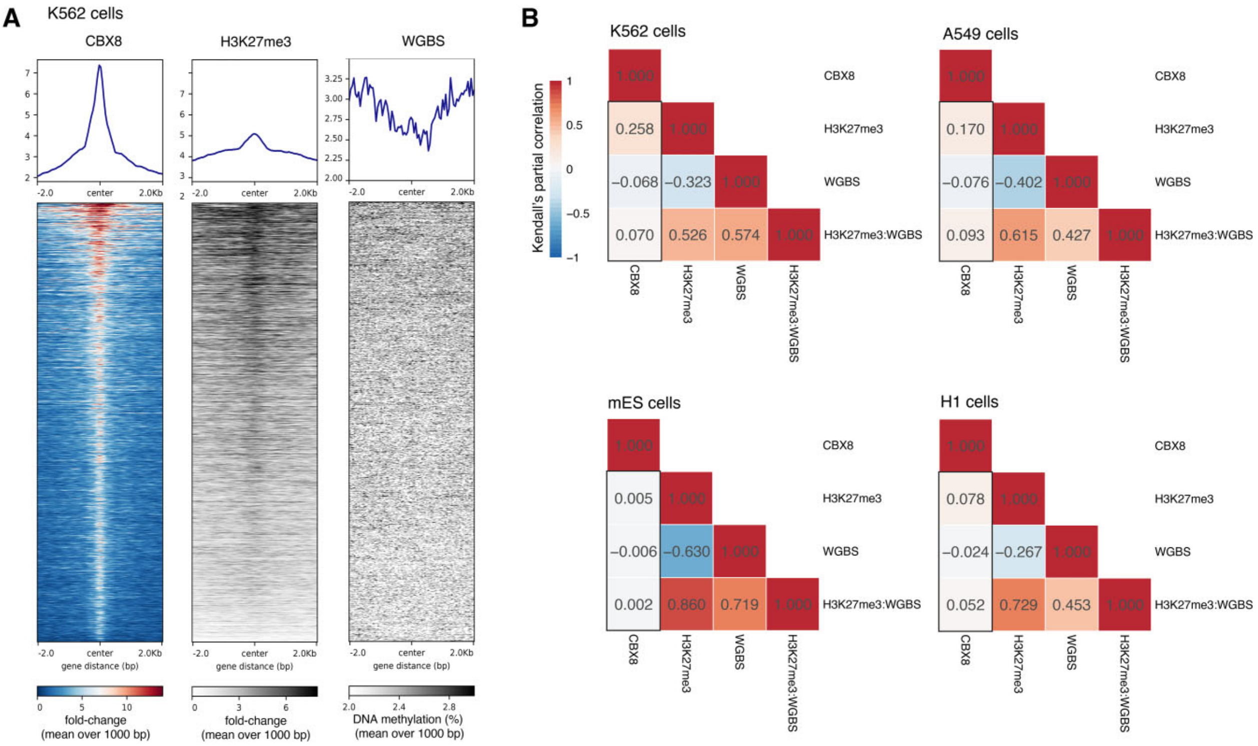


Figure 6. (**A**) Heatmap for score distributions across CBX8 IDR (Irreproducible Discovery Rate) thresholded peaks in K562 cells created with deeptools on the Galaxy web platform. (**B**) Kendall's partial correlation between CBX8, H3K27me3, WGBS data and the product between H3K27me3 and WGBS for A549, K562, H1 and mES cells. The first column in each partial correlation plot recapitulates the main and interaction effects, derived by `asteRIa`. For ENCODE and ChIP-Atlas identifier see caption of Supplementary Figure S3.

tive for CBX8 and WGBS (b+r pattern). Furthermore, the partial correlation between CBX8 and the product of WGBS and H3K27me3 ChIP-seq values represents the additional combinatorial interaction effect, complementing the individual effects. This partial correlation is positive across all cell types, leading to the b+r+b pattern observed in `asteRIa`, Furthermore, it tends to be larger in magnitude than the negative partial correlation between the CBX8 and WGBS data, which also aligns with the `asteRIa` results on the CBX8 nucleosome binding data.

In summary, this analysis provides evidence that `asteRIa`'s estimated main and interaction effects can be recapitulated using other high-throughput experimental data. Moreover, this type of analysis provides a recipe for further validation and invites to perform new ChIP-Seq experiments for other candidate proteins that show evidence of combinatorial interaction effects.

Discussion

While many functions and readers of individual chromatin modifications have been described (15,16), the understanding of how multiple modifications cooperate in recruiting epigenetic regulators has remained largely elusive. To gain insights into these cooperative effects, we have introduced `asteRIa`, a workflow for the robust statistical detection of interaction effects, and applied the workflow to the recently published MARCS nucleosome binding dataset. The MARCS data comprise a library of semi-synthetic di-nucleosomes followed by nucleosome affinity purification with high-throughput quantitative proteomics measurements. Despite MARCS' unique approach to probe the binding behavior of proteins to combinatorial chromatin modifications at a large scale, the imbalanced design matrix and the low sample size pose considerable challenges for consistent statistical interaction estimation. `asteRIa` presents a first step toward identifying robust combinatorial effects between chromatin modifications and is tailored specifically to address these challenges. At its core, `asteRIa` combines the lasso for hierarchical interactions (53) with the complementary pairs stability selection (CPSS) concept (59), and incorporates replicate consistency mechanisms to minimize the identification of spurious interaction effects. We also confirm in a realistic synthetic simulation scenario that combining the interaction model with CPSS reduces the number of spurious effects considerably and leads to more robust results compared to the standard cross-validation procedure (see Supplementary information and Supplementary Figures S1 and S2).

By employing `asteRIa` in conjunction with the MARCS dataset, our study provides the first quantitative framework for the identification of cooperative effects of chromatin modifications on protein binding. We identify a list of 55 epigenetic reader candidates that likely respond to combinatorial modification effects. For the set of 55 proteins we confirmed that interactions enhance predictive performance of protein binding.

To evaluate the validity of `asteRIa`'s data consistency checks, we performed a sensitivity analysis, comparing `asteRIa`'s sign-consistency checks to distance-based consistency filtering and no data filtering. Our analysis demonstrates that requiring data sign-consistency results in the largest number of replicate consistent models and gives the largest set of robustly identified proteins responding to chromatin modification interactions (see Supplementary Figure S5).

For the 55 proteins identified, we observed consistent responses to these combinations across multiple proteins within the same protein complex, further substantiating the robustness of our findings. The derived candidate set also allowed for a quantitative categorization of different modes of potential chromatin modification interactions.

While our analysis is naturally limited to combinations of chromatin modifications that co-occur in at least one MARCS experiment, we were able to both recapitulate established effects of chromatin modifications on protein binding behavior and discover novel interaction effects between chromatin modifications, potentially promising candidates for future functional analyses. An intriguing finding of our analysis is the discovery of several combinations of cooperative chromatin modifications that elicit responses of the ATAC, SAGA and TFIID complexes. In particular, we identified several interactions involving the H4K20me2 modification, especially in combination with H3K4me3, H3-5ac, and H4-4ac. Another intriguing finding from our analysis is the similar binding profile observed for the proteins DNMT1, MAD2L2 and ACTL8 - all exhibiting a repulsive combinatorial effect in response to DNA methylation m5c and H3K9me3. The function of ACTL8 has not been extensively studied. However, its analogous behavior to MAD2L2 and especially DNMT1 provides an initial hint to a potential function of ACTL8.

We demonstrated the generalizability of our findings beyond a specific cell type or experimental setup by comparing the interaction effect of H3K27me3 and methylated DNA on CBX8, as identified by asteRIa, using publicly available ChIP-seq and WGBS data from K562, A549, H1 and mES cells sourced from ENCODE and ChIP-Atlas. Our analysis revealed that, even with the modest improvement in predictive accuracy observed for CBX8 when considering the identified interaction effect between H3K27me3 and methylated DNA, similar patterns are consistently observed in ChIP-seq and WGBS experiments across these diverse cell types.

However, it is important to note that the majority of combinatorial chromatin modification interaction effects identified by asteRIa, particularly those characterized by strong interaction effect sizes, are not present in publicly available ChIP-seq datasets. Consequently, we posit that our study serves as a first unbiased attempt to identify chromatin regulators that respond to more than one modification and thereby act as a hypothesis generator, suggesting specific combinations of proteins and chromatin modifications worthy of further investigation in future biological experiments. In particular, we recommend focusing on proteins that exhibit relatively poor predictive accuracy when considering individual chromatin modification effects alone. For instance, proteins like RFC2, RFC3, RFC4 and RFC5 show a substantial enhancement in predictive accuracy when considering the identified interaction effect between H4K20me2 and H3K4me3.

Moreover, asteRIa functions as a versatile tool that can be readily updated whenever new nucleosome affinity purification experiments become available. As tools are developed to conduct a greater number of experiments with additional combinations of modifications, our workflow can be conveniently extended to explore more and higher-order interaction effects between chromatin modifications, allowing a more comprehensive understanding of the combinatorial complexity of chromatin modifications.

Even though our statistical workflow has been specifically designed and optimized for the MARCS dataset, its methodology and approach can be broadly applied in scenarios where robust assessment of hierarchical interactions is required, particularly in data-scarce regimes with high levels of noise.

In conclusion, our study provides compelling evidence that large-scale SILAC nucleosome affinity purification data, when combined with asteRIa, is a potent resource for generating hypotheses related to epigenetic reader candidates.

Data availability

The asteRIa workflow, the processed data, and the code for reproducing all figures and results are available at https://figshare.com/articles/software/asteRIa/25003103 and (partly, without large files) at https://github.com/marastadler/asteRIa.git. The MARCS data is available at https://marcs.helmholtz-munich.de. Mass spectrometry data for MARCS was submitted to the PRIDE database (https://www.ebi.ac.uk/pride/) (accession number: PXD018966).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We greatly acknowledge Roberto Olayo Alarcon for valuable discussion on the data validation. We acknowledge the ENCODE Consortium and the ENCODE production laboratorys generating the particular datasets used in this study.

ENCODE K562 identifier: ENCFF405HIO, ENCFF687ZGN, ENCFF522HZT, ENCFF459XNY; ENCODE A549 identifier: ENCFF702IOJ, ENCFF081CPV, ENCFF723WVM, ENCFF552VXR; ENCODE H1 identifier: ENCFF345VHG, ENCFF284JDC, ENCFF975NYJ, ENCFF483UZG; ChIP-Atlas mES identifier: SRX426373, SRX006968, DRX001152, SRX5090173.05.

Author contributions: M.S. developed asteRIa, conducted the analysis on the MARCS data, ChIP-Seq and bisulfite data and conducted the analysis on synthetic data. C.L.M. and T.B. supervised the work. M.S. and C.L.M. conceived the statistical workflow. S.L. and T.B. analyzed the results and provided feedback. M.S., C.L.M. and T.B. wrote the manuscript. All authors read and approved the final manuscript.

Funding

M.S. is supported by the Helmholtz Association under the joint research school ‘Munich School for Data Science - MUDS’; T.B. is supported by the Helmholtz Association. Funding for open access charge: Helmholtz Gemeinschaft.

Conflict of interest statement

None declared.

References

1. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

2. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
3. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
4. Garcia,B.A., Pesavento,J.J., Mizzen,C.A. and Kelleher,N.L. (2007) Pervasive combinatorial modification of histone H3 in human cells. *Nat. Methods*, **4**, 487–489.
5. Pesavento,J.J., Bullock,C.R., LeDuc,R.D., Mizzen,C.A. and Kelleher,N.L. (2008) Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. *J. Biol. Chem.*, **283**, 14927–14937.
6. Shema,E., Jones,D., Shoresh,N., Donohue,L., Ram,O. and Bernstein,B.E. (2016) Single-molecule decoding of combinatorially modified nucleosomes. *Science*, **352**, 717–721.
7. Tvardovskiy,A., Schwämmle,V., Kempf,S.J., Rogowska-Wrzesinska,A. and Jensen,O.N. (2017) Accumulation of histone variant H3.3 with age is associated with profound changes in the histone methylation landscape. *Nucleic Acids Res.*, **45**, 9272–9289.
8. Voigt,P., LeRoy,G., Drury,W.J. III, Zee,B.M., Son,J., Beck,D.B., Young,N.L., Garcia,B.A. and Reinberg,D. (2012) Asymmetrically modified nucleosomes. *Cell*, **151**, 181–193.
9. Young,N.L., DiMaggio,P.A., Plazas-Mayorca,M.D., Baliban,R.C., Floudas,C.A. and Garcia,B.A. (2009) High throughput characterization of combinatorial histone codes. *Mol. Cell. Proteomics*, **8**, 2266–2284.
10. Li,S., Peng,Y. and Panchenko,A.R. (2022) DNA methylation: Precise modulation of chromatin structure and dynamics. *Curr. Opin. Struct, Biol.*, **75**, 102430.
11. Ruthenburg,A.J., Li,H., Patel,D.J. and Allis,C.D. (2007) Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell Biol.*, **8**, 983–994.
12. Turner,B.M. (1993) Decoding the nucleosome. *Cell*, **75**, 5–8.
13. Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.
14. Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.
15. Greenberg,M. V.C. and Bourc'his,D. (2019) The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.*, **20**, 590–607.
16. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
17. Li,B., Gogol,M., Carey,M., Lee,D., Seidel,C. and Workman,J.L. (2007) Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science*, **316**, 1050–1054.
18. Tsai,W.-W., Wang,Z., Yiu,T.T., Akdemir,K.C., Xia,W., Winter,S., Tsai,C.-Y., Shi,X., Schwarzer,D., Plunkett,W., *et al.* (2010) TRIM24 links a non-canonical histone signature to breast cancer. *Nature*, **468**, 927–932.
19. Eustermann,S., Yang,J.-C., Law,M.J., Amos,R., Chapman,L.M., Jelinska,C., Garrick,D., Clynes,D., Gibbons,R.J., Rhodes,D., *et al.* (2011) Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. *Nat. Struct. Mol. Biol.*, **18**, 777–782.
20. Ruthenburg,A.J., Li,H., Milne,T.A., Dewell,S., McGinty,R.K., Yuen,M., Ueberheide,B., Dou,Y., Muir,T.W., Patel,D.J., *et al.* (2011) Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions. *Cell*, **145**, 692–706.
21. Su,W.-P., Hsu,S.-H., Chia,L.-C., Lin,J.-Y., Chang,S.-B., Jiang,Z.-d., Lin,Y.-J., Shih,M.-Y., Chen,Y.-C., Chang,M.-S., *et al.* (2016) Combined interactions of plant homeodomain and chromodomain regulate NuA4 activity at DNA double-strand breaks. *Genetics*, **202**, 77–92.
22. Borgel,J., Tyl,M., Schiller,K., Pusztai,Z., Dooley,C.M., Deng,W., Wooding,C., White,R.J., Warnecke,T., Leonhardt,H., *et al.* (2017) KDM2A integrates DNA and histone modification signals through a CXXC/PHD module and direct interaction with HP1. *Nucleic Acids Res.*, **45**, 1114–1129.
23. Jurkowska,R.Z., Qin,S., Kungulovski,G., Tempel,W., Liu,Y., Bashtrykov,P., Stiefelmaier,J., Jurkowski,T.P., Kudithipudi,S., Weirich,S., *et al.* (2017) H3K14ac is linked to methylation of H3K9 by the triple Tudor domain of SETDB1. *Nat. Commun.*, **8**, 2057.
24. Botuyan,M.V., Lee,J., Ward,I.M., Kim,J.-E., Thompson,J.R., Chen,J. and Mer,G. (2006) Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. *Cell*, **127**, 1361–1373.
25. Fradet-Turcotte,A., Canny,M.D., Escribano-Díaz,C., Orthwein,A., Leung,C.C.Y., Huang,H., Landry,M.-C., Kitevski-LeBlanc,J., Noordermeer,S.M., Sicheri,F., *et al.* (2013) 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature*, **499**, 50–54.
26. Nakamura,K., Saredi,G., Becker,J.R., Foster,B.M., Nguyen,N.V., Beyer,T.E., Cesa,L.C., Faull,P.A., Lukauskas,S., Frimurer,T., *et al.* (2019) H4K20me0 recognition by BRCA1-BARD1 directs homologous recombination to sister chromatids. *Nat. Cell Biol.*, **21**, 311–318.
27. Sobhian,B., Shao,G., Lilli,D.R., Culhane,A.C., Moreau,L.A., Xia,B., Livingston,D.M. and Greenberg,R.A. (2007) RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science*, **316**, 1198–1202.
28. Kim,H., Chen,J. and Yu,X. (2007) Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. *Science*, **316**, 1202–1205.
29. Yan,J., Kim,Y.-S., Yang,X.-P., Li,L.-P., Liao,G., Xia,F. and Jetten,A.M. (2007) The ubiquitin-interacting motif containing protein RAP80 interacts with BRCA1 and functions in DNA damage repair response. *Cancer Res.*, **67**, 6647–6656.
30. Wilson,M.D., Benlekbir,S., Fradet-Turcotte,A., Sherker,A., Julien,J.-P., McEwan,A., Noordermeer,S.M., Sicheri,F., Rubinstein,J.L. and Durocher,D. (2016) The structural basis of modified nucleosome recognition by 53BP1. *Nature*, **536**, 100–103.
31. Hu,Q., Botuyan,M.V., Zhao,D., Cui,G., Mer,E. and Mer,G. (2021) Mechanisms of BRCA1-BARD1 nucleosome recognition and ubiquitylation. *Nature*, **596**, 438–443.
32. Rajakumara,E., Wang,Z., Ma,H., Hu,L., Chen,H., Lin,Y., Guo,R., Wu,F., Li,H., Lan,F., *et al.* (2011) PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression. *Mol. Cell*, **43**, 275–284.
33. Arita,K., Isogai,S., Oda,T., Unoki,M., Sugita,K., Sekiyama,N., Kuwata,K., Hamamoto,R., Tochio,H., Sato,M., *et al.* (2012) Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 12950–12955.
34. Rothbart,S.B., Krajewski,K., Nady,N., Tempel,W., Xue,S., Badeaux,A.I., Barsyte-Lovejoy,D., Martinez,J.Y., Bedford,M.T., Fuchs,S.M. and et,al. (2012) Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.*, **19**, 1155–1160.
35. Arita,K., Ariyoshi,M., Tochio,H., Nakamura,Y. and Shirakawa,M. (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*, **455**, 818–821.
36. Avvakumov,G.V., Walker,J.R., Xue,S., Li,Y., Duan,S., Bronner,C., Arrowsmith,C.H. and Dhe-Paganon,S. (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature*, **455**, 822–825.
37. Hashimoto,H., Horton,J.R., Zhang,X., Bostick,M., Jacobsen,S.E. and Cheng,X. (2008) The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature*, **455**, 826–829.
38. Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
39. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A.,

Beaudet,A.L., Ecker,J.R., *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
40. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
41. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public Ch IP-seq data. *EMBO Reports*, **19**, e46255.
42. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
43. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
44. Lasserre,J., Chung,H.-R. and Vingron,M. (2013) Finding associations among histone modifications using sparse partial correlation networks. *PLoS Comput. Biol.*, **9**, e1003168.
45. Perner,J. (2015) Bioinformatic approaches for understanding chromatin regulation. PhD thesis.
46. Perner,J., Lasserre,J., Kinkley,S., Vingron,M. and Chung,H.-R. (2014) Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Res.*, **42**, 13689–13695.
47. Bartke,T., Vermeulen,M., Xhemalce,B., Robson,S.C., Mann,M. and Kouzarides,T. (2010) Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*, **143**, 470–484.
48. Nelder,J. (1977) A reformulation of linear models. *J. R. Stat. Soc. Ser. A: Stat. Soc.*, **140**, 48–63.
49. Aiken,L.S., West,S.G. and Reno,R.R. (1991) Multiple Regression: Testing and Interpreting Interactions. Sage.
50. Hamada,M. and Wu,C.J. (1992) Analysis of designed experiments with complex aliasing. *J. Qual. Technol.*, **24**, 130–137.
51. Duncan,R.P. and Kefford,B.J. (2021) Interactions in statistical models: three things to know. *Methods Ecol. Evol.*, **12**, 2287–2297.
52. Simonsohn,U. (2022) Interacting with curves: How to validly test and probe interactions in the real (nonlinear) world.
53. Bien,J., Taylor,J. and Tibshirani,R. (2013) A lasso for hierarchical interactions. *Ann. Stat.*, **41**, 1111.
54. Lim,M. and Hastie,T. (2015) Learning Interactions via Hierarchical Group-Lasso Regularization. *J. Comput. Graph. Stat.*, **24**, 627–654.
55. Hao,N., Feng,Y. and Zhang,H.H. (2018) Model selection for high-dimensional quadratic regression via regularization. *J. Am. Stat. Assoc.*, **113**, 615–625.
56. Peixoto,J.L. (1987) Hierarchical variable selection in polynomial regression models. *Am. Stat.*, **41**, 311–313.
57. Yu,B. (2013) Stability. *Bernoulli*, **19**, 1484–1500.
58. Meinshausen,N. and Bühlmann,P. (2010) Stability Selection. *J. R. Stat. Soc. Ser. B*, **72**, 417–473.
59. Shah,R.D. and Samworth,R.J. (2013) Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **75**, 55–80.
60. Capraz,T. and Huber,W. (2023) Feature selection by replicate reproducibility and non-redundancy. bioRxiv doi: https://doi.org/10.1101/2023.07.04.547623, 04 July 2023, preprint: not peer reviewed.
61. Lukauskas,S., Tvardovskiy,A., Nguyen,N.V., Stadler,M., Faull,P., Ravnsborg,T., Özdemir Aygenli,B., Dornauer,S., Flynn,H., Lindeboom,R.G., *et al.* (2024) Decoding chromatin states by proteomic profiling of nucleosome readers. *Nature*, **627**, 671–679.
62. Lowary,P. and Widom,J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.
63. Muir,T.W. (2003) Semisynthesis of proteins by expressed protein ligation. *Annu. Rev. Biochem.*, **72**, 249–289.
64. Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **58**, 267–288.
65. Bien,J. and Tibshirani,R. (2020) hierNet: A Lasso for Hierarchical Interactions, R package version 1.9.
66. Lederer,J. and Müller,C. (2015) Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. **29**.
67. Wu,Y. and Wang,L. (2020) A survey of tuning parameter selection for high-dimensional regression. *Annu. Rev. Stat. Appl.*, **7**, 209–226.
68. Liu,H., Roeder,K. and Wasserman,L. (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neu. Inf. Proc. Syst.*, **24**, 1432–1440.
69. Bodinier,B., Filippi,S., Nøst,T.H., Chiquet,J. and Chadeau-Hyam,M. (2023) Automated calibration for stability selection in penalised regression and graphical models. *J. R. Stat. Soc. Ser. C: Appl. Stat.*, **72**, 1375–1393.
70. Maddu,S., Cheeseman,B.L., Sbalzarini,I.F. and Müller,C.L. (2022) Stability selection enables robust learning of differential equations from limited noisy data. *Proc. R. Soc. A*, **478**, 20210916.
71. Fasel,U., Kutz,J.N., Brunton,B.W. and Brunton,S.L. (2022) Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. A*, **478**, 20210904.
72. Hofner,B. and Hothorn,T. (2021) stabs: Stability Selection with Error Control, R package version 0.6-4.
73. Blainey,P., Krzywinski,M. and Altman,N. (2014) Replication: quality is often more important than quantity. *Nat. Methods*, **11**, 879–881.
74. Teng,M., Love,M.I., Davis,C.A., Djebali,S., Dobin,A., Graveley,B.R., Li,S., Mason,C.E., Olson,S., Pervouchine,D., *et al.* (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, **17**, 74.
75. Ormaza,G., Rodríguez,J.A., de Opakua,A.I., Merino,N., Villate,M., Gorroño,I., Rábano,M., Palmero,I., Vilaseca,M., Kypta,R., *et al.* (2019) The tumor suppressor ING5 is a dimeric, bivalent recognition molecule of the histone H3K4me3 mark. *J. Mol. Biol.*, **431**, 2298–2319.
76. Xie,S. and Qian,C. (2018) The growing complexity of UHRF1-mediated maintenance DNA methylation. *Genes (Basel)*, **9**, 600.
77. Petryk,N., Bultmann,S., Bartke,T. and Defossez,P.A. (2021) Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. *Nucleic Acids Res.*, **49**, 3020–3032.
78. Jeltsch,A. and Jurkowska,R.Z. (2016) Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm. *Nucleic Acids Res.*, **44**, 8556–8575.
79. Geng,Z. and Gao,Z. (2020) Mammalian PRC1 Complexes: Compositional Complexity and Diverse Molecular Mechanisms. *Int. J. Mol. Sci.*, **21**, 8594.
80. Connelly,K.E., Weaver,T.M., Alpsoy,A., Gu,B.X., Musselman,C.A. and Dykhuizen,E.C. (2019) Engagement of DNA and H3K27me3 by the CBX8 chromodomain drives chromatin association. *Nucleic Acids Res.*, **47**, 2289–2305.
81. Cheon,Y., Kim,H., Park,K., *et al.* (2020) Dynamic modules of the coactivator SAGA in eukaryotic transcription. *Experimental & Molecular Medicine*, **52**, 991–1003.
82. Timmers,H. T.M. (2021) SAGA and TFIID: Friends of TBP drifting apart. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.*, **1864**, 194604.
83. Chen,B.-R. and Sleckman,B.P. (2022) The Regulation of DNA End Resection by Chromatin Response to DNA Double Strand Breaks. *Front. Cell Dev. Biol.*, **10**, 932633.
84. Seczynska,M., Bloor,S., Cuesta,S.M., *et al.* (2022) Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature*, **601**, 440–445.
85. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K., *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.*, **48**, D882–D889.

86. Kagda,M.S., Lam,B., Litton,C., Small,C., Sloan,C.A., Spragins,E., Tanaka,F., Whaling,I., Gabdank,I., Youngworth,I., *et al.* (2023) Data navigation on the ENCODE portal. arXiv doi: https://arxiv.org/abs/2305.00006, 04 May 2023, preprint: not peer reviewed.
87. Hitz,B.C., Lee,J.-W., Jolanki,O., Kagda,M.S., Graham,K., Sud,P., Gabdank,I., Strattan,J.S., Sloan,C.A., Dreszer,T., *et al.* (2023) The ENCODE Uniform Analysis Pipelines. bioRxiv doi: https://doi.org/10.1101/2023.04.04.535623, 06 April 2023, preprint: not peer reviewed.
88. Oki,S. and Ohta,T. (2015) ChIP-Atlas. https://chip-atlas.org.
89. Zou,Z., Ohta,T., Miura,F. and Oki,S. (2022) ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res*, **50**, W175–W182.
90. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep*, **19**, e46255.
91. Community,T.G. (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.*, **50**, W345–W351.
92. Kim,S. (2015) ppcor: Partial and Semi-Partial (Part) Correlation, R package version 1.1.

Received: January 17, 2024. **Revised:** March 15, 2024. **Editorial Decision:** April 20, 2024. **Accepted:** April 24, 2024

Supplementary information

Synthetic data generation and performance comparison

We construct a realistic synthetic data scenario to demonstrate that employing complementary pairs stability selection yields significantly more robust and accurate outcomes compared to the default implementation of cross-validation in the Lasso for hierarchical interactions (hiernet). To achieve this, we employ an asymmetric Laplace distribution to model the non-zero coefficient estimates derived from actual data across all proteins (see Fig. S1a and b).

In order to maintain consistent sparsity levels for both proteins and features (chromatin modifications or interactions between chromatin modifications), we opt for the simplest approach: retaining the same sparsity pattern as observed in the estimated coefficients from the real data. Additionally, we model the distribution of estimated intercepts for all proteins using a Laplace distribution. Introducing a normally distributed error term, akin to the noise inherent in the actual data, further enhances the fidelity of our synthetic data. Furthermore, we introduce variations in this noise level to illustrate how the quality of outcomes responds to differing degrees of noise. Using the simulated intercept, the product of the simulated coefficients, and the true experimental design matrix containing interaction terms, along with the simulated error term, we generate a synthetic dataset representing protein binding, $P_{syn} = I_{syn} + \Theta_{syn}(L, L_{int}) + E_{syn}$, with $P_{syn},\ I_{syn},\ E_{syn} \sim (p, n)$, $\Theta_{syn} \sim (p, q^2/2)$, $(L, L_{int}) \sim (q^2/2, n)$ (see Fig. S1c).

We compare the following two approaches using the `hierNet` model: one employing 5-fold cross-validation with the 1-standard error (1se) rule, and the other utilizing complementary pairs stability selection (CPSS). These experiments were conducted on a subset of 58 synthetic proteins known to exhibit interactions. Across both experiment sets, we conduct 20 replicates for each configuration to ensure robustness and reliability of our findings (see Fig. S2).

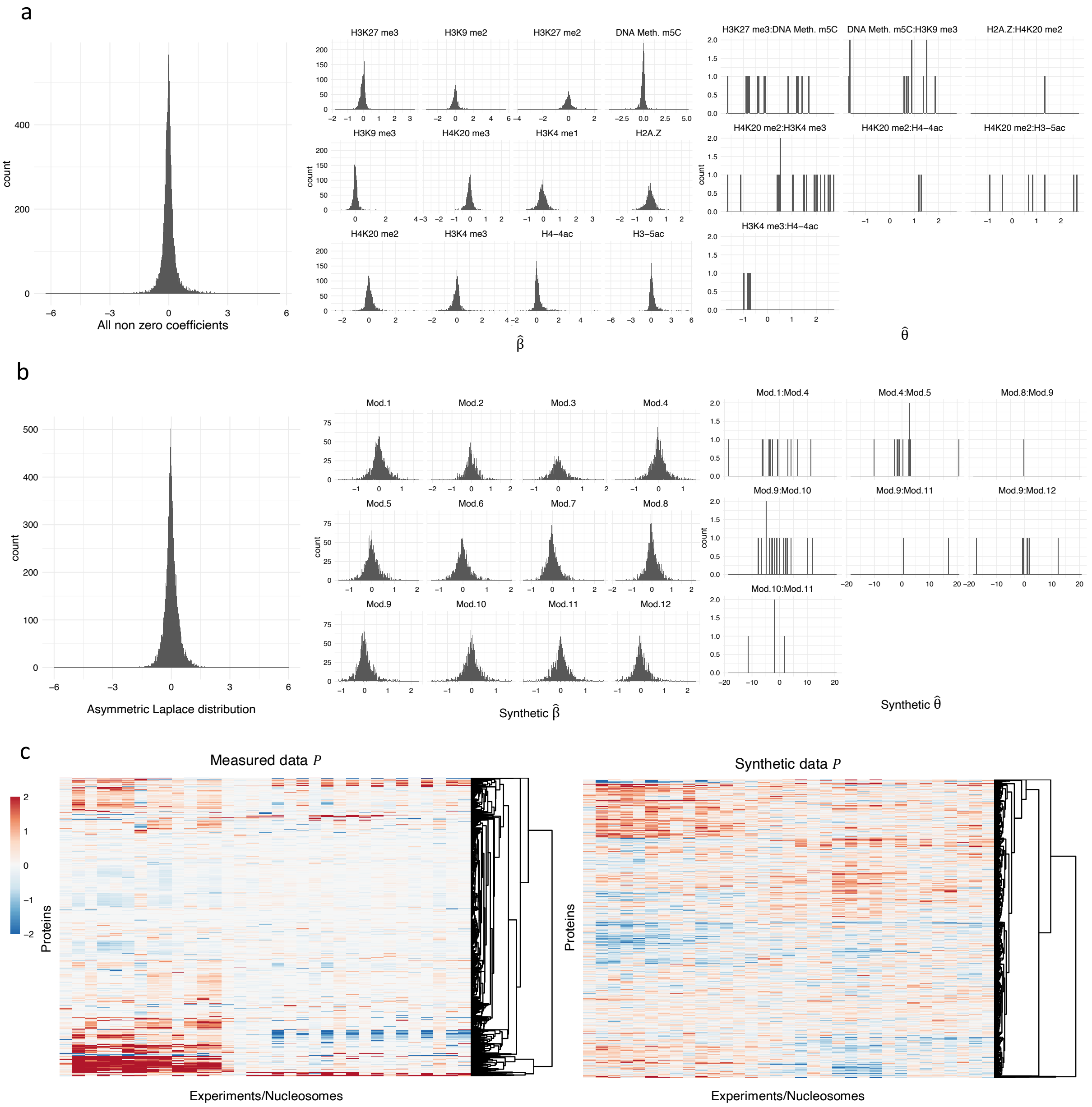


Fig S1. **a**, Joint and individual distributions of all non zero estimated coefficients $\hat{\beta}$ and $\hat{\Theta}$ in the statistical workflow. **b**, Asymmetric Laplace distribution fitted to the joint distribution estimated coefficients in **a**. **c**, Left: clustered heatmap of proteins binding measures P (mean of forward and reverse experiment); right: clustered heatmap of synthetic protein binding measures P.

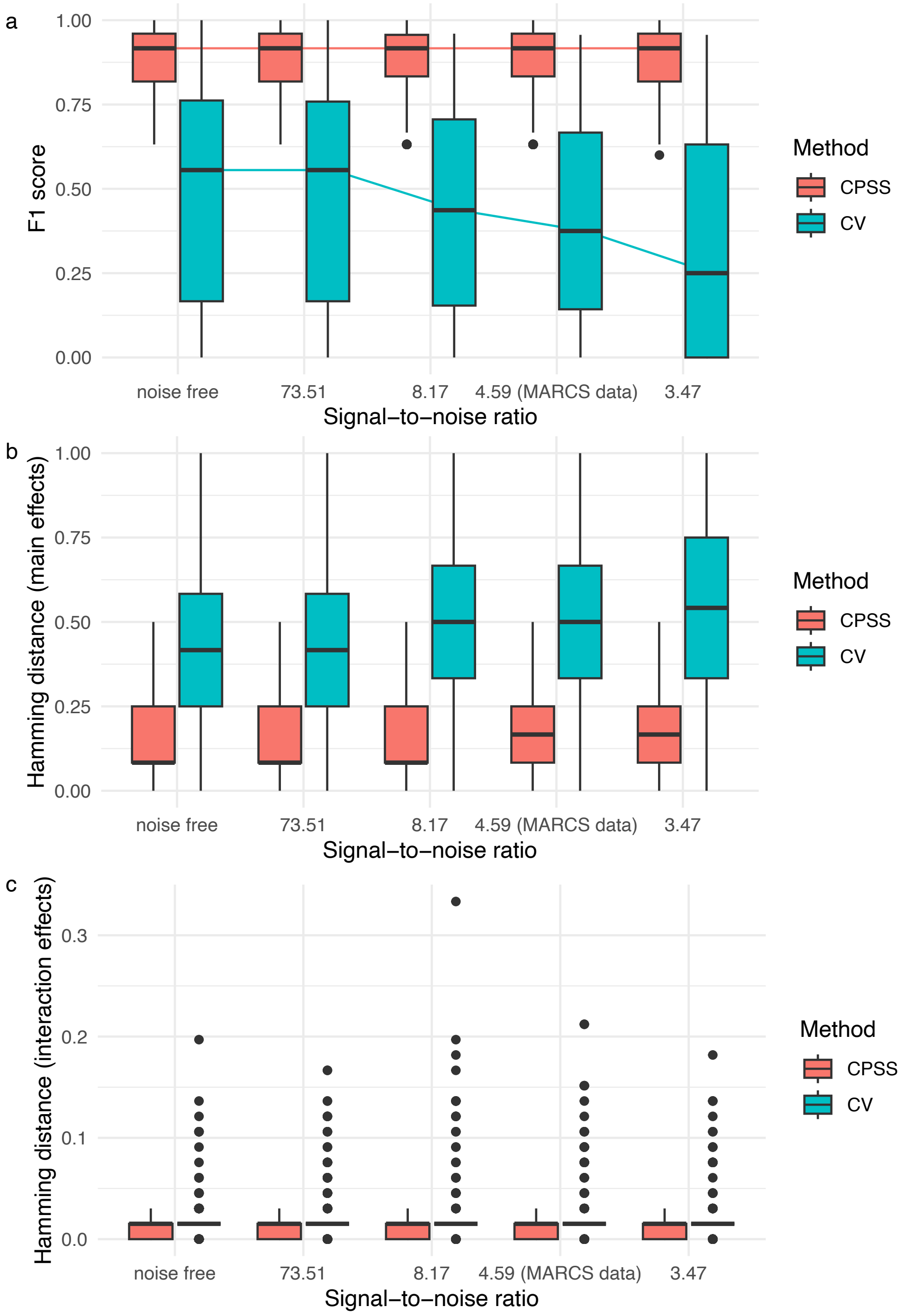


Fig S2. **a**, F1 score for five signal-to-noise ratios (SNR) for hiernet with CPSS and 5-fold cross-validation (1se rule). SNR of 4.59 corresponds to the SNR observed in the MARCS data. Noise free corresponds to 0% noise, SNR = 73.51 corresponds to 25% of the noise observed for the MARCS data; SNR = 8.17 corresponds to 75% of the noise observed for the MARCS data and SNR = 3.47 corresponds to 125% of the noise observed for the MARCS data. **b**, Hamming distance main effects for hiernet with CPSS and 5-fold cross-validation (1se rule) for five SNRs. **c**, Hamming distance interaction effects for hiernet with CPSS and 5-fold cross-validation (1se rule) for five SNRs.

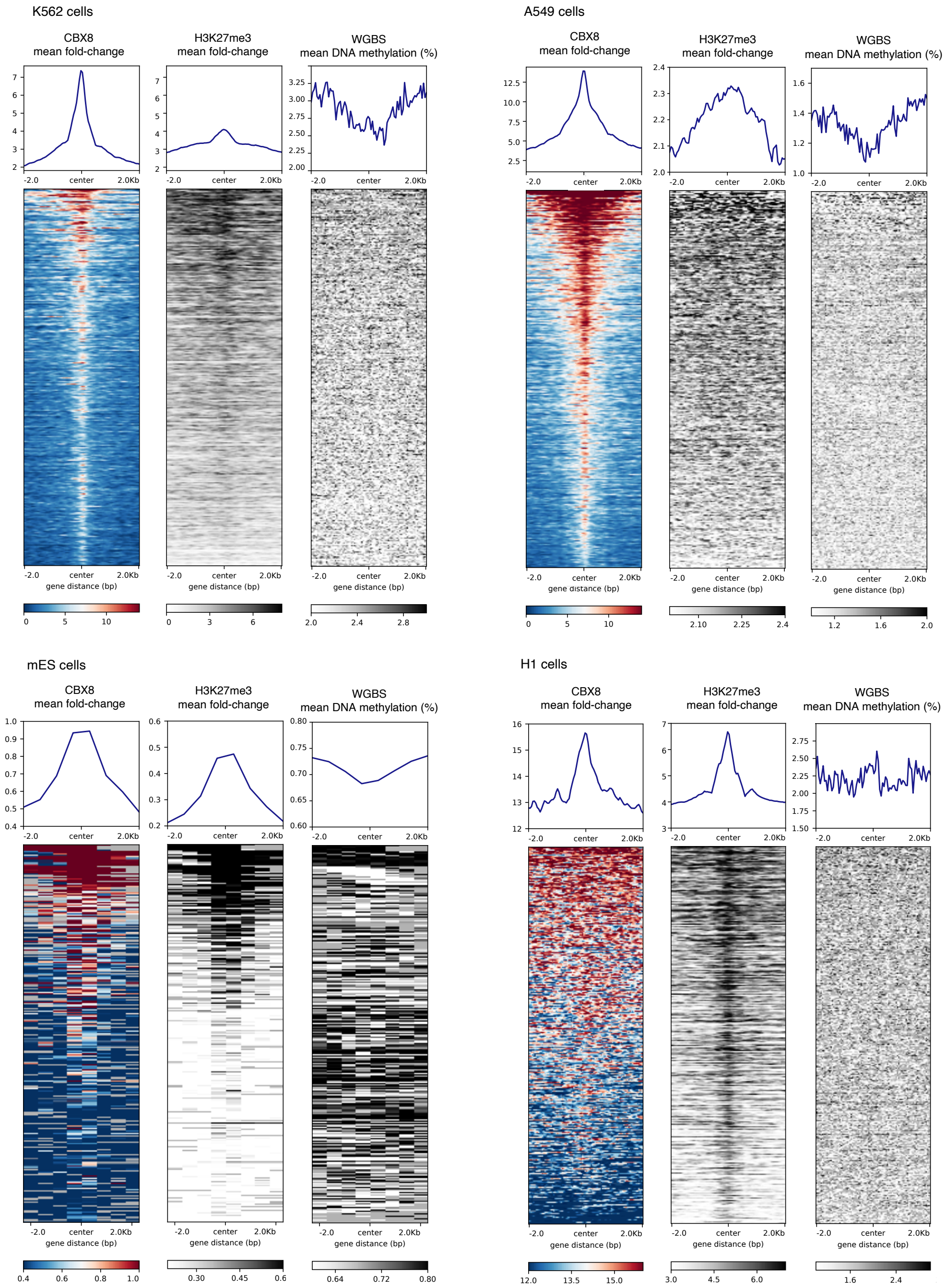


Fig S3. Heatmaps of score distributions across CBX8 IDR thresholded peaks in K562, A549, H1 and mES cells. mES heatmaps are based on 500 bp bins for visualization purposes because of missing values. Fold-changes in mES ChIP-Atlas experiments are scaled between 0 and 1 while mean fold-changes in ENCODE experiments represent raw values. ENCODE K562 identifier: ENCFF405HIO, ENCFF687ZGN, ENCFF522HZT, ENCFF459XNY; ENCODE A549 identifier: ENCFF702IOJ, ENCFF081CPV, ENCFF723WVM, ENCFF552VXR; ENCODE H1 identifier: ENCFF345VHG, ENCFF284JDC, ENCFF975NYJ, ENCFF483UZG; ChIP-Atlas mES identifier: SRX426373, SRX006968, DRX001152, SRX5090173.05.

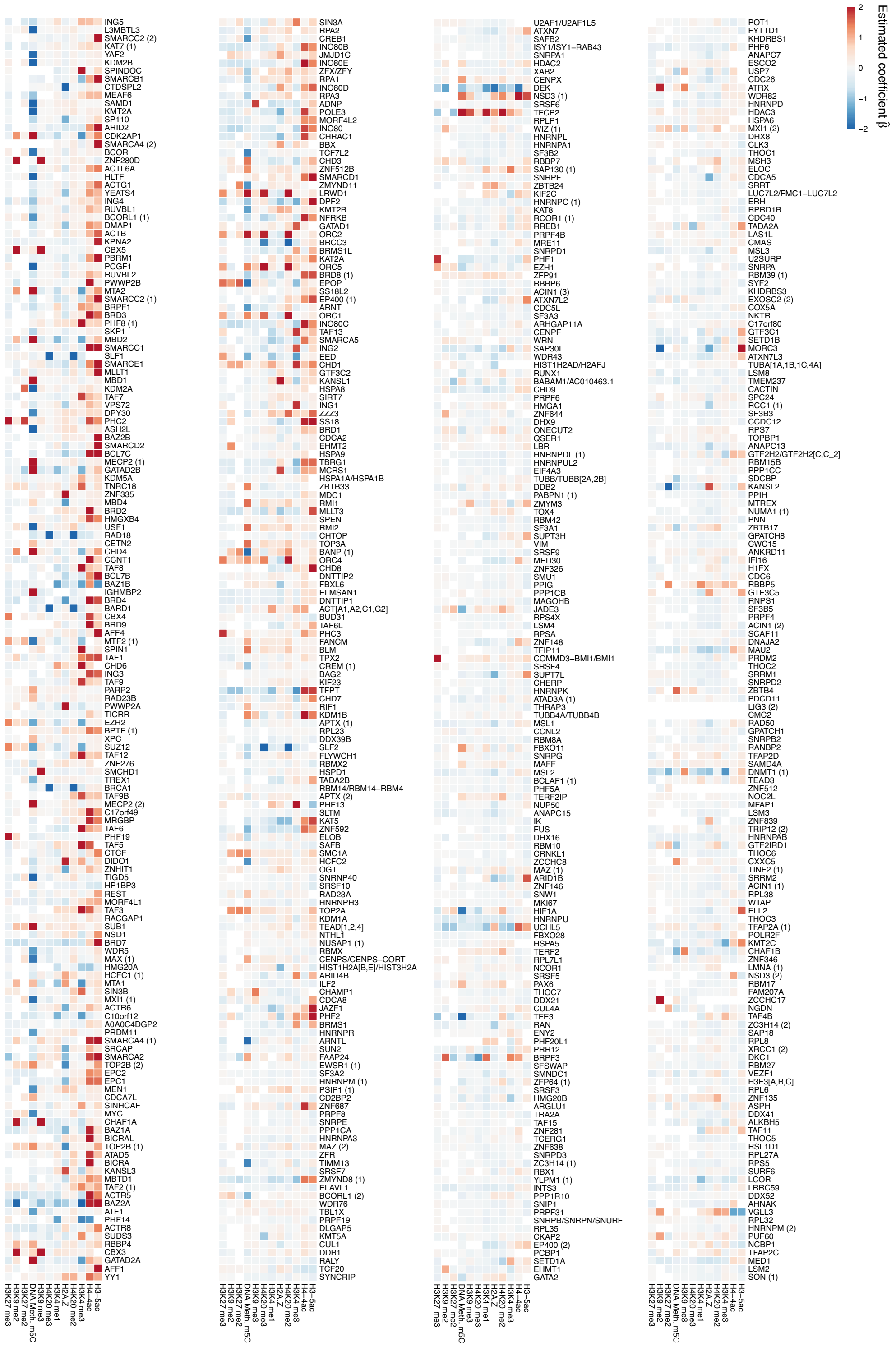


Fig S4. Heatmap of estimated main effects in the linear model. The proteins are ordered in descending order of predictive performance (R^2). A full list of all proteins is provided in Extended Fig. S4.

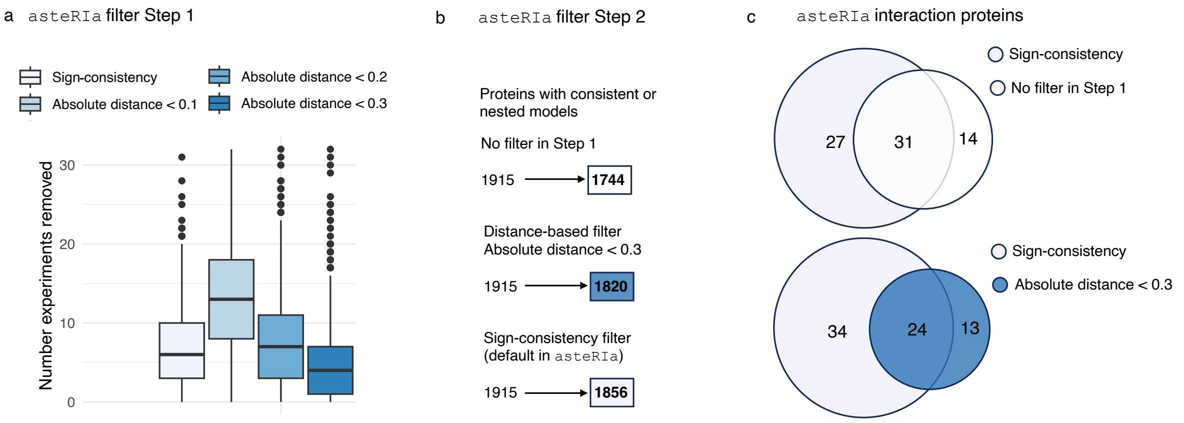


Fig S5. Sensitivity analysis of filter step 1 in asteRIa. **a**, Boxplots showing the number of experiments removed from the total of 1915 proteins for different filters: sign-consistency (default in asteRIa) and three distance-based filters. Retaining experiments where the absolute distance is < 0.1 or < 0.2 removes more experiments than requiring sign-consistency. **b**, Number of proteins filtered out in the model-based filtering in step 2 in asteRIa for the cases: no filter in step 1, distance-based filter (< 0.3) in step 1, and sign-consistency filter in step 1. **c**, Venn diagrams representing the number and overlap of proteins with identified interactions by asteRIa. The upper Venn diagram compares sign-consistency to no filter in step 1, the lower Venn diagram compares sign-consistency to the distance-based filter (< 0.3) in step 1.

Overview Extended Figures

- **Extended Fig. 3b** Extension of Fig. 3b. Full list of stability plots for all 1915 proteins.
- **Extended Fig. 3c** Extended Fig. 3c. Extension of Fig. 3c. Full list of scatter plots for all 55 proteins with robust interaction effects.
- **Extended Fig. 5b** Extension of Fig. 5b. Selection probability heatmaps and model coefficients for all 1915 protein complexes.
- **Extended Fig. S4** Extension of Fig. S4. Heatmap of estimated main effects in the linear model for all 1915 proteins.

A.2 Predictive modeling of microbial data with interaction effects

Contributing article

Stadler, M., Müller, C. L., Bien, J. (2024). Predictive modeling of microbial data with interaction effects. *bioRxiv, 2024-04.* doi: https://doi.org/10.1101/2024.04.29.591596

Replication code

The source data and code for reproducing all results of this study is available at https://github.com/marastadler/Microbial-Interactions.

Copyright information

Author contributions

M.S. developed and implemented the statistical approaches and data analyses. J.B. and C.L.M. supervised the work. J.B., C.L.M. and M.S. conceived the statistical models. J.B. and C.L.M. analyzed the results and provided feedback. M.S. wrote the manuscript with input from all co-authors.

Predictive modeling of microbial data with interaction effects

Mara Stadler[1,2], Jacob Bien[3], Christian L. Müller[1,2,4]

1 Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
2 Department of Statistics, Ludwig Maximilians University Munich, 80539 Munich, Germany
3 Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA
4 Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA

* Correspondence: mara.stadler@helmholtz-munich.de

The code for reproducing the analysis and figures is accessible on the following GitHub public repository: `https://github.com/marastadler/Microbial-Interactions`

Abstract

Microbial interactions are of fundamental importance for the functioning and the maintenance of microbial communities. Deciphering these interactions from observational data or controlled lab experiments remains a formidable challenge due to their context-dependent nature, i.e., their dependence on (a)biotic factors, host characteristics, and overall community composition. Here, we present a statistical regression framework for microbial data that allows the inclusion and parsimonious estimation of species interaction effects for an outcome of interest. We adapt the penalized quadratic interaction model to accommodate common microbial data types as predictors, including microbial presence-absence data, relative (or compositional) abundance data from microbiome surveys, and quantitative (absolute abundance) microbiome data. We study the effect of including hierarchical interaction constraints and stability-based model selection on model performance and propose novel interaction model formulations for compositional data. To illustrate our framework's versatility, we consider prediction tasks across a wide range of microbial datasets and ecosystems, including metabolite production in model communities in designed experiments and environmental covariate prediction from marine microbiome data. While we generally observe superior predictive performance of our interaction models, we also assess limits of these models in presence of extreme data sparsity and with respect to data type. On a large-scale gut microbiome cohort data, we identify sparse family-level interaction models that accurately predict the abundance of antimicrobial resistance genes, enabling the formulation of novel biological hypotheses about microbial community interactions and antimicrobial resistance.

1 Introduction

A fundamental objective in microbial ecology is to elucidate how species compositions and species-species interactions are related to the maintenance and functioning of a microbial community [1]. Interactions between microbial species come in many forms, including cross-feeding interactions through metabolite exchange, bacteriocin-induced growth-inhibitory interactions, and exchange of genetic material for genotype selection [2, 3]. Conceptually, microbial interactions can be described in terms of their net positive, negative, or neutral effect on their interaction partner, resulting in broad categories such as mutualistic, commensal, amensal, predatory/parasitic/exploitative, antagonistic or competitive interactions [4, 5, 6, 2]. Experimentally identifying and verifying such interactions within natural communities has remained a difficult task, owing to the sheer complexity of microbial ecosystems and limited technical capabilities to dissect such communities.

With the emergence of large-scale microbial survey data, computational approaches have become popular that use statistical regression and correlation methods to estimate sparse species-species association and co-occurrence networks from microbial abundance measurements [5, 7, 8, 9, 10, 11]. While these networks do not necessarily reflect true ecological relationships [12], they can provide valuable insights into the global structure of microbial communities across ecosystems [13, 14]. However, none of these methods allow to relate species-species associations or "interactions" to a community functional outcome of interest or to concomitant environmental or host-related covariates. Furthermore, most network approaches deliver context-independent (or averaged) pairwise associations, thus potentially missing species-species interactions that are relevant for a specific function of the community.

In this contribution, we provide a statistical regression framework for microbial data that allows the parsimonious inclusion of microbial interactions for predicting an outcome of interest, such as butyrate [15] or a concomitantly measured covariate [16]. Using the generic quadratic interaction regression model as a starting point, we adapt the model to accommodate all common microbial abundance data modalities (see Fig. 1 for an illustration). Important examples include data from designed *in-vitro* experimental studies on model microbial communities where microbial abundance comes in form of presence-absence (binary) data or absolute abundances, i.e., non-negative count or continuous data [17]. The majority of microbiome survey data, however, quantify taxon abundances by amplicon sequencing, thus providing primarily relative abundance (or compositional) data [18] in form of Operational Taxonomic Units (OTUs) or Amplicon Sequencing Variants (ASVs) [19]. Moreover, recent quantitative microbiome profiling techniques [20, 21, 22] combine absolute cell count measurements and relative amplicon data, thus providing absolute microbial abundance information, albeit with potential biases [23].

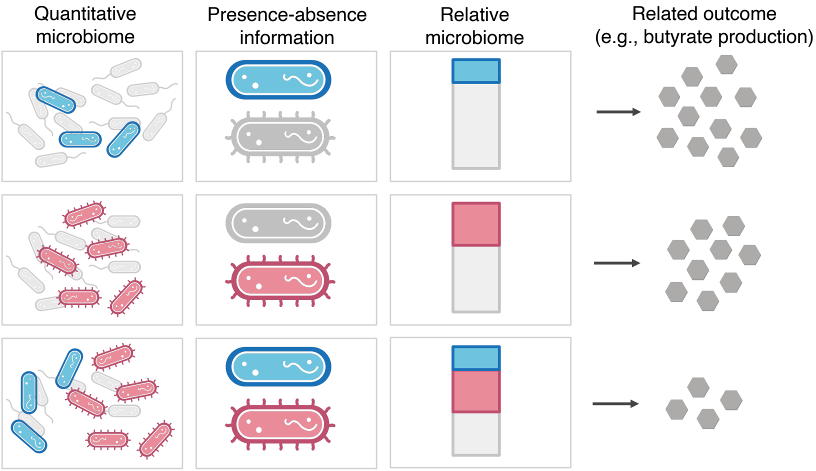

Quantitative microbiome
Presence-absence information
Relative microbiome
Related outcome (e.g., butyrate production)

Fig 1. Illustration of three data modalities in microbiome analysis and their combinatorial behavior with respect to an outcome, e.g., a community function (created with BioRender.com). The sketch depicts three distinct data modalities in the columns: (i) quantitative microbiome data, representing absolute counts; (ii) presence-absence information of microbial species; and (iii) relative abundance data, also known as compositions. Each row illustrates a simplified scenario. In the first scenario, blue microbes are present while red are absent, resulting in a large (production of) outcome (e.g., butyrate). In the second scenario, red microbes are present while blue are absent, leading to another large (production of) outcome. In the third scenario, both blue and red groups of microbes are present, yet only minimal amounts of the outcome are produced, indicating an antagonistic combinatorial effect between the two groups.

Our framework unifies and generalizes several seemingly disjoint approaches in the literature of microbial ecology and microbiome data science. For example, several recent studies in microbial ecology use presence-absence and absolute abundance data from designed experiments on small microbial communities to predict community functions [24, 15], such as, e.g., butyrate production, or overall host fitness [25] using the quadratic (and higher-order) interaction model. Our framework is readily available for such studies and gives statistical guidelines how to choose model complexity, how hierarchical constraints can increase model interpretability, and how to analyze higher dimensional datasets.

On the other hand, for regression tasks based on high-dimensional large-scale amplicon sequencing data, many statistical approaches consider the *linear* log-contrast model [26], which is the standard linear model for compositional data, as the baseline model. To deal with the high dimensionality (where typically the number of features p is larger than number of samples n), penalized and structured regression models have been proposed [27, 28, 29, 30, 31, 16]. Here, we extend these linear (main effects) model to include species interaction effects. Specifically, starting with Aitchison's (low- dimensional) proposal [26], we introduce three models with quadratic interactions that work with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic log-ratio model. To achieve parsimonious models in the high-dimensional setting, we employ ℓ_1 penalization and illustrate via semi-synthetic data simulations when quadratic interactions are identifiable, given the excess sparsity of typical microbiome data.

To achieve stable and interpretable interaction models [32], we follow [33] and incorporate hierarchical interaction modeling [34, 35, 36] and stability-based model selection [37, 38] into our framework. The hierarchy assumption enforces constraints on interaction features, requiring that they can only be included in the model if both features (strong hierarchy) or at least one feature (weak hierarchy) are already present as main effects. Stability-based model selection ensures that interactions are only included if they can be consistently and reproducibly identified across different subsets of the data which will likely help reduce the number of testable biological hypotheses.

We demonstrate the versatility of our framework by analyzing datasets that encompass all three data modalities across various ecosystems, including synthetic microbial communities, human gut microbiomes, and marine microbial ecosystems. Notably, our application of the quadratic interaction model on a quantitative microbiome data from the Metacardis study [39] reveals its effectiveness in accurately estimating the abundance of antimicrobial resistance genes (ARGs) from microbial taxa abundances. Furthermore, our analysis of a microbiome dataset containing presence-absence information rediscovers a stable interaction effect, specifically the inhibitory role of *D. piger* on the butyrate producer *A. caccae* [24]. For the newly introduced *sparse quadratic log-contrast model* tailored for relative microbiome data, we provide both semi-synthetic data simulations to demonstrate the model's ability to accurately detect interaction effects and, following [16], re-analyze Tara ocean data [40], highlighting superior predictive performance of sparse interaction modeling compared to their linear counterparts. We conclude by providing a comparative analysis of the quadratic interaction models across the three data modalities using the Metacardis ARG prediction task, illustrating commonalities and differences across the resulting predictive models. The latter analysis gives further guidance for the practitioner regarding merits and pitfalls of quadratic interaction models. Our framework for quadratic interaction modeling is freely available as reproducible R code at `https://github.com/marastadler/Microbial-Interactions`.

2 Methods

2.1 Interaction modeling strategy

Given the abundance information of p microbial taxa $X = (X_1, ..., X_p)$, the baseline model for uncovering (joint) additive effects of the microbial taxa on an outcome $Y \in \mathbb{R}^n$ (e.g., butyrate production), is the linear model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{1}$$

where β_0 is the intercept term, β_j is the effect of taxon j on Y, and ϵ models the technical and biological noise term.

In many prediction tasks, relying on a linear (main effect) model alone is insufficient to capture the complexity of dynamics within microbial communities. A common approach to introduce a more intricate yet interpretable model is the inclusion of quadratic terms. Here, we extend the baseline model by introducing a generic quadratic interaction model, incorporating all pairwise interactions between microbial taxa, namely

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \epsilon, \tag{2}$$

where $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$ is a symmetric matrix of interactions. We assume $\Theta_{jj} = 0$ in this model formulation. However, the general principles still apply if this constraint is removed.

In the following section, we instantiate the interaction model to accommodate distinct data types and denote the microbial abundance information by A for count information (absolute or relative) and B for presence-absence information (see Fig. 1).

2.1.1 Interaction model for quantitative microbiome data

Whenever microbial abundance information is given as absolute counts, the model is equal to the generic model 2 and does not require further transformation of the input data or any constraints on the model coefficients. Throughout this work, we denote the absolute count input data by $A \in \mathbb{R}_+^{n \times p}$. Assuming that Y depends on the actual amounts of taxon abundances, the quadratic interaction model is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j A_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} A_j A_k + \epsilon, \tag{3}$$

where the model parameters follow the description provided in model 2.

2.1.2 Interaction model for presence-absence microbiome data

If the microbial abundance information is represented as presence-absence data, given by a binary matrix, we denote the microbial abundance information as $B \in \{0, 1\}^{n \times p}$, where 1 indicates the presence of a microbial taxon, and 0 indicates its absence. One common alternative encoding is $B \in \{-1, 1\}^{n \times p}$, where the absence is encoded as -1. The choice of encoding does not affect the ability of the model to fit the data, it only changes the interpretation of the coefficient. Assuming that Y depends on the presence-absence information of microbial taxa, the quadratic interaction model is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j B_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} B_j B_k + \epsilon, \tag{4}$$

where the model parameters follow the description provided in model 2. For $B \in \{0, 1\}$, β_0 is the baseline effect when all features, here referring to microbial taxa, are absent, and β_j for $j = 1, ..., p$ represents the effect of the presence of B_j when all other taxa are absent. The interaction term, Θ_{jk}, accounts for the additional effect when both features B_j and B_k are present. For $B \in \{-1, 1\}$, β_0 signifies the overall mean (assuming a completely balanced design). For more details on the interpretation and the linear transformations of model coefficients between these two encodings, see the Supplementary Material. When describing Y as a fitness or phenotypic landscape, the different encodings in the interaction model are often associated with Fourier and Taylor expansions, allowing the parameters to describe landscape properties, like ruggedness [15, 41, 42].

2.1.3 Interaction modeling for relative microbiome data

When the microbial count information in X is provided as (sparse) compositions rather than as absolute counts, one way of modeling interaction effects between microbial taxa includes converting X to a binary matrix $B = \mathbb{1}_{\{X>0\}}$ that carries the presence-absence information of microbial taxa.

However, the compositional information might hold valuable insights beyond that provided by presence-absence data. We introduce three methods for modeling quadratic interactions with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic log-ratio model. The three models differ in terms of interpretability, dimensionality, and optimization, and the choice of which model to use depends on the underlying data and the biological question.

Alr transformed quadratic model While comparing the relative count information in compositional data is not biologically meaningful, describing the response as a linear combination of log-ratios derived from the original compositions is a valid comparison. One popular way of building log-ratios is by choosing a common reference feature p, such that the transformed count is given by $C_j = \log(A_j/A_p),\ j = 1, ..., p-1$ [26]. This transformation is known as the additive log-ratio transformation (alr)-transformation. The alr transformation allows modeling an outcome Y based on the $(p-1)$-dimensional compositional input data, assuming that Y depends on the composition of X, not on the actual amount, by a model linear in the features

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \epsilon \tag{5}$$

and an extension to interaction effects, given by

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j=1}^{p-1} \sum_{k=1}^{p-1} \Theta_{jk} C_j C_k + \epsilon. \tag{6}$$

For $p := p - 1$ this formulation is equal to the generic model 2 (with Θ not being symmetric). While this very general model formulation allows for the interpretation of the effects with respect to a specific reference feature p, extensions to expressions in the p-dimensional space [26] and log-ratio models that allow pairwise comparisons between features [30] have been proposed. Both approaches can be translated back to the model formulation in 5 and 6, respectively, and will be discussed in the following two paragraphs.

Constrained quadratic log-contrast model As shown in [26], a more convenient symmetric expression of the linear alr transformed model 5, that does not require a reference feature, can be derived by reformulating the equation as a p-dimensional problem including a zero-sum constraint, given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \epsilon, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0, \tag{7}$$

where the main (log) effect coefficients $\beta_j, j = 1, ..., p$ sum up to zero. As illustrated in [26], the extension to the quadratic log-contrast model can be represented as

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \frac{1}{2} \sum_{j \neq p} \Theta_{jk} \log(A_j / A_k)^2, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j = 0, \tag{8}$$

where the main (log) effect coefficients $\beta_j, j = 1, ..., p$ sum up to zero, with $\beta \in \mathbb{R}^p$, and the interaction effect coefficients Θ_{jk} correspond to the quadratic (log-ratio) interaction effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$. In the Supplementary information we show how to formulate the alr transformed model as constrained quadratic log-contrast model.

Quadratic log-ratio model Another way of accounting for compositionality in regression models is to build log-ratios between all possible pairs of features in $A \in \mathbb{R}_+^{n \times p}$. This approach is referred to the (all-pairs) log-ratio model [30], which is given by

$$Y = \beta_0 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j / A_k) + \epsilon, \tag{9}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k.

In the same way as in 8 the log-ratio model can be extended to a quadratic version, the quadratic log-ratio interaction model (qlr), namely,

$$Y = \beta_0 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j/A_k) + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log(A_j/A_k)^2 + \epsilon, \tag{10}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k, with $\beta \in \mathbb{R}^{p(p-1)/2}$ and the interaction effect coefficient Θ_{jk} corresponds to the quadratic (log-ratio) effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$. There exists a linear transformation between the main effect coefficients β_j in model 7 and model 8 and the main effects coefficients $\beta_{j,k}$ in model 9 and model 10, $\beta_j = -\sum_{k=1}^{j-1} \beta_{k,j} + \sum_{k=j+1}^{p} \beta_{j,k}$, implying that the zero-sum constraint on $\beta \in \mathbb{R}^p$ is inherently met in the linear and quadratic log-ratio model. While the models are mathematically equivalent, their interpretations are different and the choice might depend on the particular data application.

2.2 Penalized model estimation

Microbial datasets typically include a large number of features p and interactions between features $p(p-1)/2$ compared to the number of observations n. Moreover, even in scenarios where $n > p(p1)/2$, we assume that a parsimonious model is most appropriate, focusing only on the selection of few features and interactions that are relevant for the outcome.

To facilitate penalized model estimation, we employ regularized maximum-likelihood estimation incorporating ℓ_1-norm (lasso) penalization for both linear and interaction coefficients, as proposed by [43]. We introduce a generic optimization problem, consisting of an objective function $\rho(l, \beta_0, \beta, \Theta)$ and a (potential) constraint set on the model parameters $c(\beta_0, \beta, \Theta)$ that facilitates parameter estimation for all (linear and interaction) models introduced in Section 2.1. The objective function takes the general form

$$\rho(l, \beta_0, \beta, \Theta) = l(\beta_0, \beta, \Theta) + \lambda \left\|\beta\right\|_1 + \frac{\lambda}{2} \left\|\Theta\right\|_1 . \tag{11}$$

Here, $\lambda > 0$ serves as a tuning parameter, regulating the sparsity levels of the coefficients β and Θ, respectively. The loss function $l(\beta_0, \beta, \Theta)$ is specific to each model. Consequently, the generic optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{12}$$

This optimization problem is subsequently instantiated by specific loss functions and constraints.

Sparse quadratic interaction model for quantitative and presence-absence microbiome data The loss function $l(\beta_0, \beta, \Theta)$ for the sparse quadratic interaction model, also all-pairs lasso, for the interaction models for absolute count data or presence-absence data, introduced in 3 and 4, is defined as

$$l^{\mathrm{qi}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k \right\|_2^2,$$

with $X := A \in \mathbb{R}_+^{n \times p}$ for absolute count data and $X := B \in \{0,1\}^{n \times p}$ (or $B \in \{-1,1\}^{n \times p}$) for presence-absence data. This model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\operatorname{minimize}}\, \rho(l^{\mathrm{qi}}, \beta_0, \beta, \Theta). \tag{13}$$

In the linear model case the loss function in the optimization problem reduces to $l(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right\|_2^2$.

Sparse alr transformed quadratic model Given $A \in \mathbb{R}^{n \times p}$ is a matrix containing the relative abundance information of p microbial taxa, the loss function in 11 for the sparse alr transformed quadratic model, introduced in 6, is defined as

$$l^{\mathrm{qalr}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j=1}^{p-1} \sum_{k=1}^{p-1} \Theta_{jk} C_j C_k \right\|_2^2,$$

with $C_j = \log(A_j / A_p)$, $j = 1, ..., p-1$. The model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\operatorname{minimize}}\, \rho(l^{\mathrm{qalr}}, \beta_0, \beta, \Theta). \tag{14}$$

In the linear model case the loss function in the optimization problem reduces to $l^{\mathrm{alr}}(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j \right\|_2^2$.

Sparse quadratic log-contrast model The linear log-contrast model has been extended to the high-dimensional setting [31, 27, 28, 16], and is also known as the sparse log-contrast model. While this model has been used in various microbiome data analysis applications, the concept of introducing interactions in the log-contrast model has been defined in [26], but has not been extended to the high-dimensional setting or used in practical applications. Here, we translate the interaction model proposed in [26] to the high-dimensional setting. The loss function for the sparse quadratic log-contrast model (qlc) corresponding to the interaction model for compositional data, introduced in 8, is defined as

$$l^{\text{qlc}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) - \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log(A_j/A_k)^2 \right\|_2^2.$$

As this model comes with a zero-sum constraint on the main effect coefficients, the constraint set in 12 is given by

$$c(\beta_0, \beta, \Theta) = \left\{ \sum_{j=1}^{p} \beta_j = 0 \right\}.$$

Thus, the optimization problem for the sparse quadratic log-contrast model is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\text{qlc}}, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{15}$$

In the linear sparse log-contrast model defined in 7, the loss function reduces to $l^{\text{lc}}(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) \right\|_2^2$.
The main effect covariates, denoted by A_j for $j = 1, ..., p$ typically remain unscaled under the zero-sum constraint. However, the interaction features $\log(A_j/A_k)^2$ are not subject to the zero-sum constraint and we scale them. The ℓ_2-norm of these interaction features tends to increase with the ℓ_2-norm of their associated main effects (more specifically, the ℓ_2-norm of the main effects after transforming them with the centered log-ratio (clr) transformation). The clr divides each compositional part by the geometric mean of all parts, namely

$$\text{clr}(A) = \left(\log \frac{A_i}{g(A_i)} \right)_{i=1,\ldots,n} \quad \text{with } g(A_i) = \exp\left(\frac{1}{p} \sum_{j=1}^{p} \log(A_{ij}) \right).$$

Here, we introduce a way of scaling the interaction features that ensures equal penalization of the interaction features. Moreover, we adjust the scale of the interaction features to align with the norm of the average clr transformed ℓ_2-norms of all main effects. Mathematically, this can be expressed as follows: We denote each column of the interaction feature matrix as $A^I_{\cdot jk} = \log(A_j/A_k)^2$, with $A^I \in \mathbb{R}^{n \times p(p-1)/2}$, and its scaled version is given by

$$A^I_{\cdot jk} \left(\left\| A^I_{\cdot jk} \right\|_2 \right)^{-1} \frac{1}{p} \sum_{k=1}^{p} \left\| A^{\text{clr}}_k \right\|_2,$$

where $A^{\text{clr}} = \text{clr}(A) \in \mathbb{R}^{n \times p}$ is the clr transformed main effects matrix A and $\left\| A^{\text{clr}}_k \right\|_2$ is the ℓ_2-norm of the k-th column of A^{clr}.

Sparse quadratic log-ratio model The loss function of the sparse quadratic log-ratio (qlr) model corresponding to the interaction model for compositional data, introduced in 10, is defined as

$$l^{\mathrm{qlr}}(\beta_0, \beta, \Theta) = \frac{1}{2}\left\| Y - \beta_0 - \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k}\log(A_j/A_k) - \frac{1}{2}\sum_{j\neq k}\Theta_{jk}\log(A_j/A_k)^2 \right\|_2^2.$$

This model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. The optimization problem for the sparse quadratic log-ratio model is therefore given as

$$\underset{\beta_0,\beta,\Theta}{\text{minimize}}\, \rho(l^{\mathrm{qlr}}, \beta_0, \beta, \Theta). \tag{16}$$

In the sparse log-ratio model, that is linear in the features and corresponds to the model in 9, the loss function reduces to $l^{\mathrm{lr}}(\beta_0, \beta) = \frac{1}{2}\left\| Y - \beta_0 - \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_{j,k}\log(A_j/A_k) \right\|_2^2$. The $p(p-1)/2$-dimensional sparse log-ratio problem has been shown to be equivalent to the sparse log-contrast model problem for $\lambda^{\mathrm{qlr}} = 2\lambda^{\mathrm{qlc}}$ [30]. This equality can be directly translated to their quadratic extensions. As the dimensionality of the predictor space in the $p(p-1)/2$-dimensional log-ratio model becomes computationally inefficient for large p, the authors in [30] propose a two-stage procedure that involves a pre-selection step for covariates to reduce the predictor space before applying the log-ratio lasso. This two-step procedure can be directly applied to the $2 \cdot p(p-1)/2$-dimensional quadratic log-ratio lasso, introduced in 10, in scenarios where p is large.

2.3 Modeling hierarchical interactions

The quadratic interaction models, introduced before, can enhance predictive performance compared to models that are linear in the features, but they may not detect robust microbial interaction effects suitable for further functional analysis. To enhance model interpretability, we introduce the statistical concept of hierarchy in the context of quadratic models for microbiome data. The concept of hierarchy permits the inclusion of an interaction Θ_{jk} in the model *only if* both associated main effects are also present in the model (strong hierarchy), or if at least one of the associated main effects is included (weak hierarchy) [see [34], and references therein]. This hierarchy can be implemented by imposing constraints on the interaction effects $\Theta_j \in \mathbb{R}^p$ for $j = 1, ..., p$ namely

$$c(\beta_0, \beta, \Theta) = \left\{\Theta = \Theta^T,\ \left\|\Theta_j\right\|_1 \leq |\beta_j|\ \right\}. \tag{17}$$

By eliminating the symmetry constraint on Θ, the resulting model relaxes to weak hierarchy on the interaction features. While 17 is non-convex, we follow [34] who proposed a convex relaxation of the problem and provided an efficient implementation in the corresponding R package `hierNet` [44] (v1.9). The hierarchical constraint can be imposed within the generic optimization problem described in 12 and allows a direct application under the convex relaxation for (i) quantitative microbiome data (ii) or presence-absence information of microbial species with 13; and (iii) relative microbiome data, after performing the alr transformation with 14.

2.4 Model selection

An essential challenge in high-dimensional penalized regression is the selection of the regularization parameter λ. This parameter balances the sparsity of model coefficients with out-of-sample predictive performance [45, 46]. Standard methods for main effects and interaction models often involve techniques such as cross-validation [34] or Information Criteria like the Akaike (AIC) and the Bayesian Information Criterion (BIC) [47]. However, these methods tend to select more predictors and interactions than necessary [47]. If the main aim lies in detecting robust effects, one way of accounting for the potential limitations caused by cross-validation in penalized regression models is the concept of stability selection [37] for identifying a set of predictive features and interactions in microbial data. Stability selection has shown effectiveness across various scientific domains, ranging from network learning [48, 49] to data-driven partial differential equation identification [50, 51]. In the context of regression, stability selection involves iteratively learning sparse regression models from subsamples of the data (e.g., $n_s = \lfloor n/2 \rfloor$), recording the frequency of selected predictors across models, and selecting the most frequent predictors for the final model. A variant of stability selection, complementary pairs stability selection (CPSS) [38], is particularly advantageous for handling unbalanced experimental designs, as it ensures that individual subsamples are independent of each other. CPSS draws b subsamples as complementary pairs $\{(a_{2l-1}, a_{2l}) : l = 1, ..., b\}$, with $a_{2l-1} \cap a_{2l} = \emptyset$ from samples $\{1, ...n\}$ of size $\lfloor n/2 \rfloor$. Applying a variable selection procedure S (for instance choosing the k first predictors entering the penalized model in the regularization path or cross-validation) to each subsample allows defining a feature specific selection probability $\hat{\pi}_i$ for $i = 1, ..., p + p(p-1)/2$ that is given by

$$\hat{\pi}_i = \frac{1}{2b} \sum_{l=1}^{2b} \mathbb{1}_{\{i \in \hat{S}(a_l)\}}. \tag{18}$$

The final selection set, denoted as $\hat{S}^{\mathrm{CPSS}}$, consists of predictors i for which the estimated selection probability $\hat{\pi}_i$ exceeds a predefined threshold π_{thr}, that represents the minimum selection frequency required for a predictor to be included in the final set. We employ the `stabs` R package [52] (v0.6-4), which offers an efficient implementation of the CPSS procedure. This approach involves defining several hyperparameters, including the set of regularization parameters Λ, the threshold $\pi_{\mathrm{thr}} \in [0, 1]$, the number of initial predictors k entering the sparse model, and the number of complementary splits b. The CPSS procedure in [52] can be directly applied to linear models. As CPSS does not make a distinction between main and interaction effects, it can be directly applied to quadratic models. An integration of the

CPSS procedure within the hierarchical interaction modeling framework has been introduced in [33].

3 Results

3.1 Applications in quantitative and presence-absence microbial data

Antimicrobial resistance genes (ARGs) play a crucial role in the survival and evolution of individual microbial species. The extensive use of antimicrobials has increased the development of resistance in pathogens, leading to an increased presence of ARGs. The number of ARGs might be associated with the composition and abundance of certain microbes in the human gut [53].

To get a better understanding of how community composition and interactions might be related to ARGs, we use quantitative microbial count information derived as mOTUs (metagenomic operational taxonomic units) from quantitative microbiome profiling for a subset of $n = 690$ individuals. Specifically, we take the abundance data from the MetaCardis cohort [39] for which metadata information is available. We aggregate the mOTUs on genus level and illustrate the modeling strategy by considering the 30 most abundant genera in our model. We denote the underlying data by $A \in \mathbb{R}^{n \times p}$ and the number of ARGs by $y \in \mathbb{R}^n$. Given that the microbial counts in A are quantitative, we fit the sparse interaction model for absolute count data defined in 3 and 13, respectively, for 10 train test splits by using 5-fold cross-validation (CV). Our results suggest that some genera and interactions between them can explain the prevalence of ARGs (see Fig. 2a, right panel). In Fig. 2a (left panel), we visualize the estimated coefficients that exhibit a non-zero median over 10 train test splits. Next to some minor main and interaction effects, Bacteroides and Escherichia show a substantial effect on the increase of the number of ARGs, while Prevotella shows a decrease. Moreover, we identify a positive interaction effect between Prevotella and Faecalibacterium, which is contrary to their individual negative effects, indicating an antagonistic association. These findings suggest that the presence and co-presence of certain bacterial species in the gut microbiota can influence the prevalence of ARGs.

In a second example we investigate the contribution of certain bacteria as well as their pairwise interplay on butyrate production, a short-chain fatty acid beneficial to human health, within an in-vitro community given the presence-absence information of bacteria within a synthetic community from [24]. Certain bacteria, known as butyrate-producers, have the ability to ferment dietary fibers into butyrate, contributing to gut health, immune function, and energy metabolism [54]. Understanding how bacteria interact in this context is essential for understanding the complexity of this process. Following [15], we use the presence-absence information, denoted by $B \in \{0,1\}^{n \times p}$ of $p = 25$ bacteria in $n = 1561$ experiments, to fit the sparse quadratic interaction model defined in 4 and 13, respectively, for 10 train test splits by using 5-fold cross-validation, to the butyrate production, denoted by $y \in \mathbb{R}^n$.

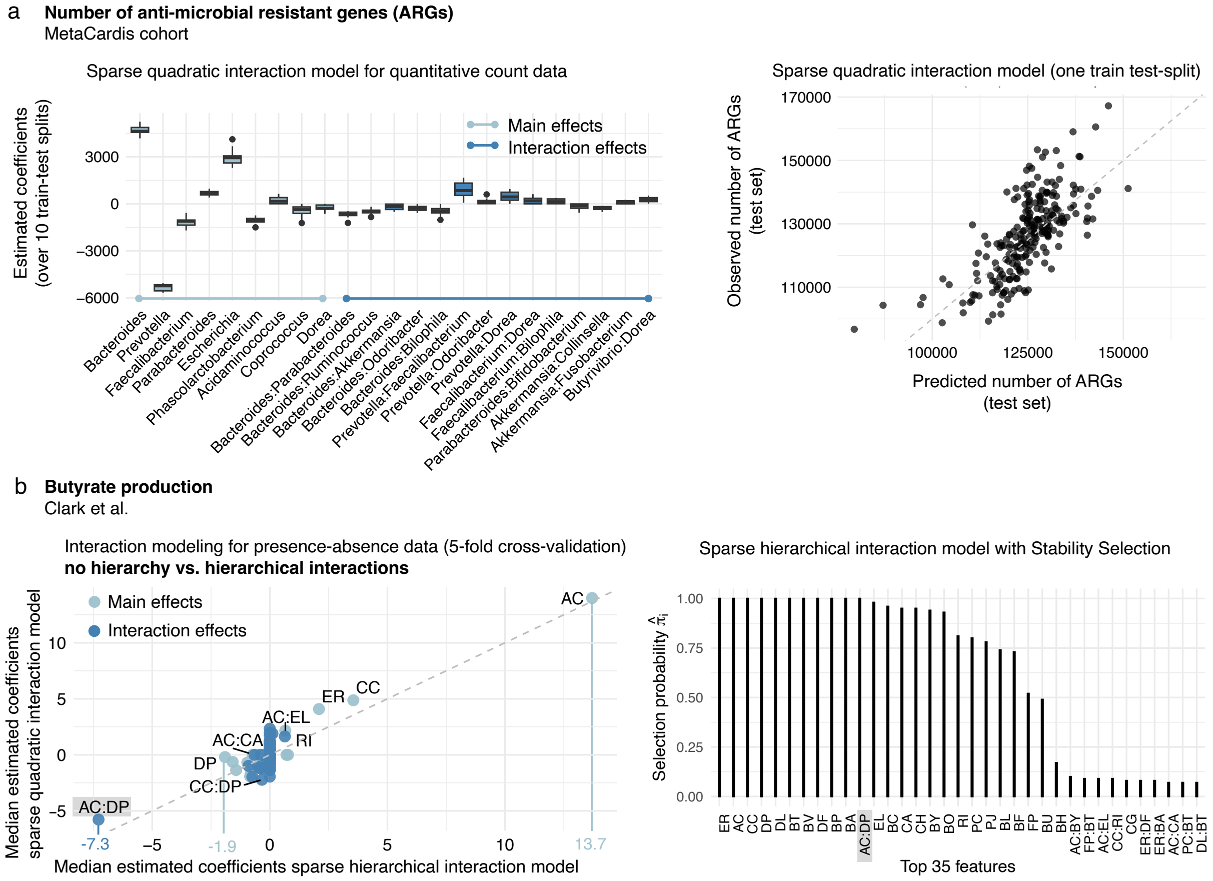


Fig 2. **a**, Prediction of the number of anti microbial resistance genes (ARGs) from absolute abundance information of microbial genera from the MetaCardis cohort. Left: distribution of estimated coefficients over 10 train test splits in the sparse quadratic interaction model 13. Only coefficients with a non-zero median estimated coefficient are shown; right: Scatterplot comparing the observed and predicted number of ARGs on a test data set for the sparse quadratic interaction model. **b**, Prediction of butyrate production from the abundance information of microbial species from [24]. Left: comparison of median estimated coefficients (over 10 train test splits) between the sparse quadratic interaction model (y axis) and the sparse hierarchical interaction model 17 with weak hierarchy (x axis); right: Top 35 selection probabilities from complementary pairs stability selection (CPSS) in the sparse hierarchical interaction model.

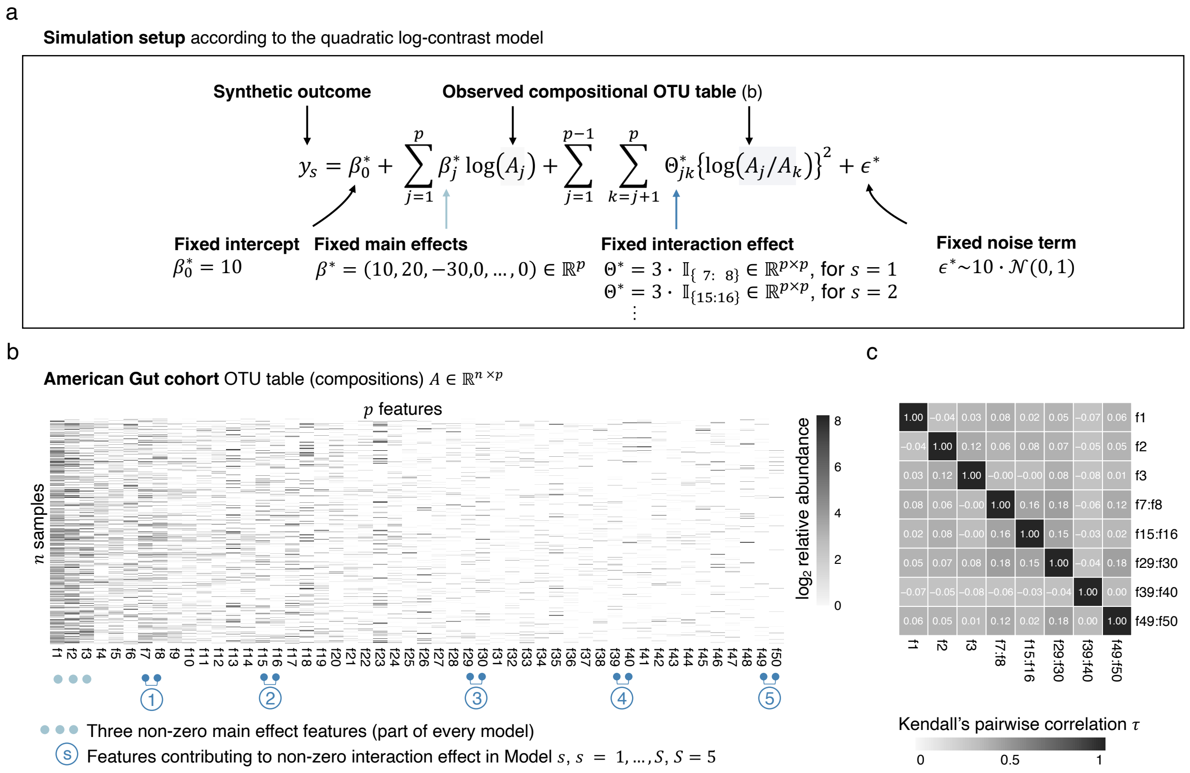
The model identifies a strong positive effect of *A. caccae* (AC) on butyrate production. AC is known to be a butyrate producer [55]. However, this effect experiences notable inhibition when AC is combined with *D. piger* (DP), reflected by a strong negative interaction effect (AC:DP) in our model, while DP itself shows only a modest negative impact on the butyrate

production. The inhibiting effect of DP on AC with respect to butyrate production has been shown in tri-cultures with *E. hallii* before [56] as well as in the context of hydrogen sulfide production by DP [24]. The model identifies a multitude of further minor main and interaction effects, including *R. intestinalis* (RI), *E. rectale* (ER), AC and *E. lenta* (AC:EL), and AC and *C. aerofaciens* (AC:CA). Under the assumption that at least one bacterium from each pair contributing to a pairwise interaction influences butyrate production individually, we apply the sparse hierarchical interaction model with weak hierarchy on the same data.
The model with hierarchy yields a substantially reduced set of selected effects while maintaining a similarly strong predictive performance (test set $R^2 = 0.72$ with weak hierarchy versus $R^2 = 0.78$ without hierarchy) (see Fig. 2b, left panel). While the effects of AC or the interaction between AC and DP stand out as clearly important predictors of butyrate production, regardless whether we fit the model with or without hierarchy or the choice of model selection procedure, the stability of smaller effects in the model, like ER or AC:ER, remains unclear. To further investigate, we combine the sparse hierarchical interaction model with stability selection, specifically complementary pairs stability selection (CPSS). The selection probabilities $\hat{\pi}_i$, where $i = 1, ..., p + p(p-1)/2$, indicate that some of the small effects are robust such as ER, *C. comes* (CC), and *D. longicatena* (DL) ($\hat{\pi}_i > 0.7$). However, all other interaction effects, such as AC:EL or AC:CA, fail to demonstrate stability across multiple subsamples.

3.2 Application of interaction modeling on relative microbiome data

3.2.1 Feasibility study of accurate interaction detection on semi-synthetic relative microbiome data

In this section, we generate semi-synthetic data to demonstrate the ability of the sparse quadratic log-contrast model (sparse qlc), defined in 8 and 15, to accurately detect interaction effects in relative microbiome data.
We elucidate the conditions under which accurate estimation is feasible by varying the degree of sparsity of the interaction features and the level of noise present in the data. In the semi-synthetic scenario, we leverage real-world relative microbial count data derived from 16S rRNA sequencing, $A \in \mathbb{R}_+^{n \times p}$, to generate S synthetic outcomes $y_s \in \mathbb{R}^n$ for $s = 1, \ldots, S$. The generation of synthetic outcomes from real data ensures that all inherent distribution properties of the dataset are retained.
In the first simulation, our goal is to understand how the sparsity level of an interaction feature affects the accuracy of the estimates. We use a subset of the data of the American Gut Project [57], processed in [16], comprising a selection of $p = 50$ OTUs, ranging from dense to sparse, and $n = 300$ subsamples (see Fig. 3b). We define $S = 5$ semi-synthetic scenarios, where $y_s \in \mathbb{R}^n$ for $s = 1, ..., S$ is given as the sum over a sparse linear combination



	f1	f2	f3	f7:f8	f15:f16	f29:f30	f39:f40	f49:f50
f1	1.00	–0.04	0.03	0.08	0.02	0.05	–0.07	0.06
f2	–0.04	1.00	0.12	0.06	0.08	0.07	–0.05	0.05
f3	0.03	0.12	1.00	–0.00	–0.00	0.08	–0.08	0.01
f7:f8	0.08	0.06	–0.00	1.00	0.16	0.18	–0.06	0.12
f15:f16	0.02	0.08	–0.00	0.16	1.00	0.15	–0.03	0.02
f29:f30	0.05	0.07	0.08	0.18	0.15	1.00	–0.04	0.18
f39:f40	–0.07	–0.05	–0.08	–0.06	–0.03	–0.04	1.00	0.00
f49:f50	0.06	0.05	0.01	0.12	0.02	0.18	0.00	1.00

Fig 3. Semi-synthetic simulation setup for varying feature sparsity levels. **a**, Simulation setup for generating a synthetic outcome y_s for $s = 1, \ldots, S$ based on the quadratic log-contrast model formulation. **b**, Heatmap of the OTU table carrying compositional information for a subset of $p = 50$ OTUs from the American Gut cohort sorted by sparsity in descending order. Non-zero main effects contributing to each of the $S = 5$ semi-synthetic scenarios (light blue) and features contributing to the non-zero interaction effect in model scenario s for $s = 1, \ldots, S$ (dark blue) are highlighted. **c**, Kendall's pairwise correlations τ between features that have non-zero effects in the models $s = 1, \ldots, S$. They should be as uncorrelated as possible ($|\tau| < .2$) to eliminate effects of correlated features.

of main and interaction effects in $A \in \mathbb{R}_+^{n \times p}$ according to model formulation 8 (see Fig. 3a). To account for the zeros in the data when building log-ratios we add a pseudo count of one to each entry in A, such that $A := A + 1$. Each outcome is characterized by a fixed intercept term $\beta_0^* = 10$, three non-zero main effects ($\beta_1 = 10$, $\beta_2 = 20$, and $\beta_3 = -30$), and a noise term $\epsilon^* = c_1 \cdot \mathcal{N}(0, 1)$, with $c_1 = 10$. Moreover, each outcome y_s for $s = 1, \ldots, S$ is characterized by a unique non-zero interaction effect $\Theta_{jk}^* = 3$ between two features A_j and A_k for $j = 1, \ldots, p - 1$ and $k = j + 1, \ldots, p$ which varies across interaction features of different sparsity levels, namely, 36% (f7:f8, $s = 1$), 52% (f15:f16, $s = 2$), 67% (f29:f30, $s = 3$), 74% (f39:f40, $s = 4$), and 88% (f49:f50, $s = 5$) zero entries (see Fig. 3b). To avoid undesired correlation effects between the predictive features in the model, we ensure that the interaction features are uncorrelated (absolute Kendall's pairwise correlation $|\tau| < 0.2$) with the main effects (see Fig. 3c). By fitting the sparse qlc model to the semi-synthetic data, we demonstrate that as interaction features become sparser, the accuracy of feature estimates with comparably small effect sizes (here $\Theta^*_{jk} = 3$) diminishes (see Tab. 1). Our simulations show that, while very sparse features with small true nonzero estimates are still selected by the sparse quadratic log-contrast model, they tend to be underestimated and are accompanied by many spuriously selected features with estimates of similar magnitude (see also Fig. S1).

Table 1. Median and Variance of the estimation error of the interaction coefficient $\sqrt{(\Theta^*_{jk} - \hat{\Theta}_{jk})^2}$ for $S = 5$ semi-synthetic scenarios.

	Sparsity interaction feature				
	36% (f7:f8)	52% (f15:f16)	67% (f29:f30)	74% (f39:f40)	88% (f49:f50)
Median estimation error	0.15	0.30	0.82	0.87	1.78
Variance estimation error	0.04	0.09	1.15	0.21	1.22

In a second simulation, we investigate the impact of noise in the data on the accurate estimation of main and interaction effects. Again, we utilize a subset of the data from the American Gut Project [57], preprocessed as described in [16]. In contrast to the previous scenario, our objective is to evaluate the influence of noise while mitigating the effects of sparsity. To achieve this, we aggregate the data at the phylum level which is the highest taxonomic level with $p = 10$ phyla. We generate $L = 4$ synthetic outcomes y_l, $l = 1, ..., L$ according to model formulation 8 by fixing main and interaction effects, while allowing the noise levels to vary. We fix the intercept term at $\beta^*_0 = 10$, assign six non-zero main effects that sum up to zero to the first six features as $\beta^* = (10, 20, 30, -10, -20, -30, 0, \dots, 0) \in \mathbb{R}^p$, and introduce three non-zero interaction effects (between A_1 and A_3, A_8 and A_{10}, and A_9 and A_{10}) as $\Theta^* = 10 \cdot \mathbb{1}_{1:3,\ 8:10,\ 9:10} \in \mathbb{R}^{p \times p}$. The noise terms ϵ^*_l follow a normal distribution with mean 0 and variance 1 with that is multiplied by a constant factor c_l, $l = 1, \dots, L$, given by $c = (10, 100, 200, 500)^T$, that varies the noise level. Thus, $\epsilon^*_l = c_l \cdot \mathcal{N}(0, 1)$. Consequently, the synthetic outcomes y_l, $l = 1, ..., L$ are given by

$$y_l = \beta^*_0 + \sum_{j=1}^{p} \beta^*_j \log(A_j) + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \Theta^*_{jk} \log(A_j/A_k)^2 + \epsilon^*_l. \tag{19}$$

To relate the noise terms to the signal part, we translate the noise levels to signal-to-noise ratios (SNR) denoted by snr_l for $l = 1, ..., L$ given by $\text{snr} = (178.01, 1.78, 0.45, 0.07)^T$. We fit both models, the sparse linear log-contrast model (sparse lc) and the sparse quadratic log-contrast (sparse qlc), to the newly generated data. Notably, features f2, f4, f5, and f6 show no contribution to interaction effects in our synthetic example and are accurately estimated by the (misspecified) sparse lc model (see Fig. 4a and b). However, for the features f1 and f3, which both have non-zero main effects as well as a common interaction effect, the sparse lc model accommodates the positive interaction effect between f1 and f3 within their main effect estimates, leading to an overestimation of the true positive main effect of f1 ($\beta_1^* = 10$) and an underestimation of the true positive main effect of f3 ($\beta_3^* = 30$). Moreover, the sparse lc model selects f8, f9 and f10 as relevant main effect features despite their lack of true nonzero main effects, in order to integrate their true underlying interaction effects. In contrast, the sparse qlc model captures the coefficients accurately. For both models, the overall performance deteriorates as the SNR decreases (see R^2 values in Fig. 4c and estimation error summary statistics in Tab. 2). Our simulations indicate that the sparse qlc model outperforms the sparse lc model when true interaction effects are present, provided that the noise level is not excessively high or the SNR is not too low (see Fig. 4c). In scenarios where noise levels are exceptionally high, the interaction model tends to overfit the data (see Fig. 4c, right plot).

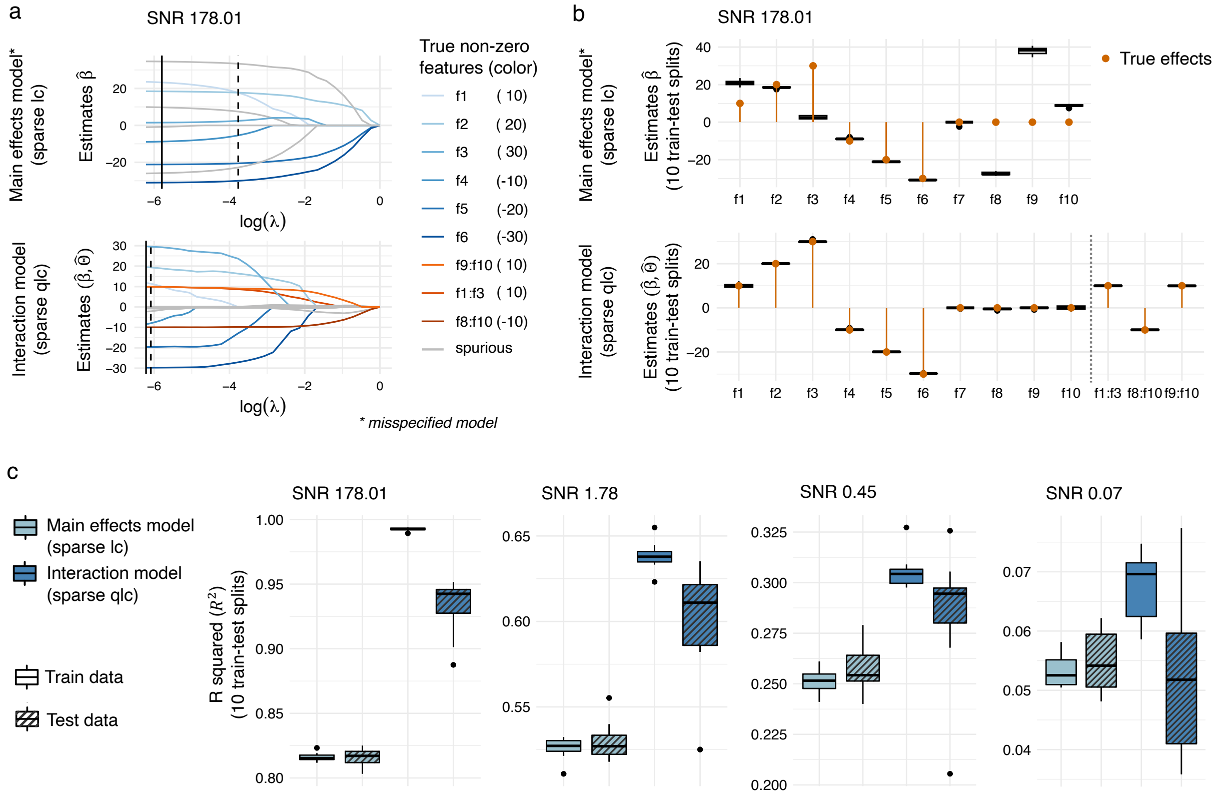


Fig 4. Influence of model misspecification and noise in semi-synthetic scenarios. **a**, Solution path for the misspecified main effects model (sparse lc) and the interaction model (sparse qlc) for the synthetic scenario $l = 1$ with a signal-to-noise ratio (SNR) of 178.01 for one train test split. **b**, Estimated coefficients distributions over 10 train test splits corresponding to the solution paths in a. For the interaction model only three non-zero interaction features are shown for visualization purposes. **c**, Comparison of model performance via R squared (R^2) for the main effects model and the interaction effects model on train and test data for four different SNRs.

Table 2. Median and Variance of the estimation error of the interaction coefficient $\sqrt{([\beta^*, \Theta^*] - [\hat{\beta}, \hat{\Theta}])^2}$ for $L = 4$ semi-synthetic scenarios with different Signal-to-noise ratios (SNR).

	SNR: 178.01	SNR: 1.78	SNR: 0.45	SNR: 0.07
Median	3.89	38.44	66.96	138.78
Variance	1.74	57.44	362.86	30.35

3.2.2 Interaction modeling on real-world relative microbiome data

Here, we perform sparse interaction modeling on environmental rather than experimental or host-associated microbiome data, highlighting salinity as a crucial factor in marine ecosystems. Variations in salinity play a critical role in shaping microbial community diversity and functionality in marine ecosystems, thereby influencing nutrient dynamics and ecosystem health [58, 59]. Using the marine data collection from Tara Oceans [60], which includes relative microbial abundance information derived as metagenomic OTUs (mOTUs) of ocean surface water and associated environmental covariates, we illustrate how sparse interaction modeling can substantially improve the predictive performance compared to models linear in the features. We aggregate the data on family level and learn a sparse quadratic log-contrast model (sparse qlc) as defined in 15 for ocean salinity from $n = 136$ samples and a subset of $p = 30$ most abundant families [61] and compare this to the corresponding sparse log contrast model that is linear in the features (sparse lc). Our comparison over 10 train test splits shows that modeling interaction effects rather than only main effects not only improves predictive accuracy (see Fig. 5a), but also allows predicting salinity concentrations beyond the interval $[33.8, 36.6]$ (see Fig. 5b). The model selects various features that are not consistently selected among all train test splits and no large main effects. However, it also identifies strong negative interaction effects between a family without annotation (f55) that belongs to the order SAR11 clade and the family Sphingomonadaceae as well as between two families without annotation (f13 and f96) that both belong to the order SAR11 clade. The negative interactions indicate that the quadratic effect of these groups (high abundances of both groups) comes with reduced salinity levels. For the families f8 and f60, that are also both part of SAR11 clade, the model identifies a positive interaction effect for all train test splits, indicating that the quadratic effect of these families comes with higher salinity levels. Several studies indicate a potential link between salinity in the ocean and the abundance of the SAR11 clade [62, 63, 64, 65]. Our findings suggest that interactions among subtypes of the SAR11 clade and between SAR11 clade families and other families may play a role in relation to the salinity level in the ocean.

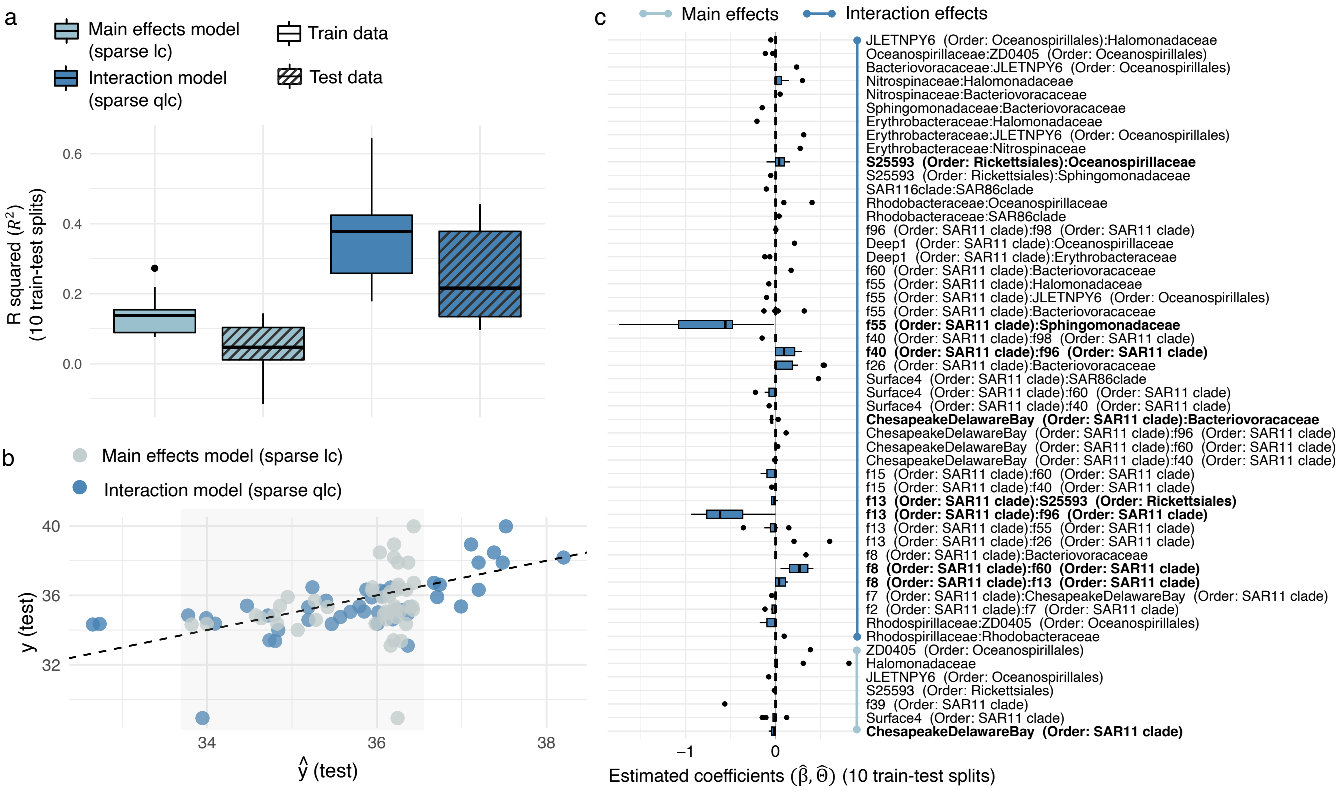


Fig 5. Summary plots of the Tara ocean data on family level for salinity prediction. **a,** Train and test set R squared R^2 distribution over 10 train test splits for the main effects sparse log contrast model (SLC) and the all-pairs log-contrast interaction lasso (SLC + int.). **b,** Scatterplot between the observed test set outcome y (salinity) and the prediction $\hat{y}$ from the main effects sparse log-contrast model (SLC) and the all-pairs log-contrast interaction lasso (SLC + int.). **c,** Distribution of estimated main effect and interaction effect coefficients in the the all-pairs log-contrast interaction lasso (SLC + int.) over 10 train test splits. Only features (main or interaction features) with a non-zero mean coefficient are shown. Features with a non-zero median are bold.

3.3 Abundance information can hold valuable insights beyond presence-absence information

The interpretation of the relationship between an outcome variable Y and species abundance information varies with the different data modalities discussed in this work. When using absolute counts, the effect of the actual abundance value can be derived, whereas presence-absence data provide insights into the effects of existence. Compositional data offer information on the impact of proportions between species abundances. While these modalities are typically analyzed based on availability, it remains unclear how effectively each data modality truly explains an outcome of interest. Furthermore, it is uncertain whether the same microbial taxa and interactions between taxa would be identified if another modality was used in the prediction task.

To illustrate this, we revisit the example on quantitative microbial abundance information from the MetaCardis cohort as discussed in Section 3.1. This data can be transformed into compositions or presence-absence information using $\boldsymbol{B} = \mathbb{1}_{A>0}$, allowing for a comparative analysis across the three data modalities for predicting the number of antimicrobial resistance genes (ARGs). We apply the sparse quadratic interaction model for absolute count data (defined in 3), the sparse quadratic interaction model for presence-absence data (defined in 4), the sparse quadratic log-contrast model (defined in 8) to the three versions of the abundance data. In this comparison, we employ the quadratic log-contrast model to analyze the relative information, as it offers the most straightforward interpretation for such comparisons.

We observe that transforming counts to relative abundances does not substantially affect the predictive performance, underscoring the importance of both absolute values and their proportions in describing the number of antimicrobial-resistant genes (ARGs), as shown in Fig. 6a. However, the presence-absence information alone does not adequately explain the number of ARGs. The genera, Prevotella, Bacteroides, and Escherichia, are identified as important predictors with consistent signs for the number of ARGs in both absolute counts and relative count information, as shown in Fig. 6b. This indicates that both the absolute abundance and the proportions relative to other taxa are significant. Escherichia, which is also part of the Proteobacteria phylum known to harbor a variety of ARGs [53, 66], and Bacteroides fragilis, a species within the Bacteroides genus, exhibits high antimicrobial resistance rates and possesses numerous mechanisms related to antimicrobial resistance [67]. These factors could explain the effects detected by our model. However, the presence-absence data do not select any of the main effects. Instead, this model frequently selects a multitude of interaction effects involving Prevotella. Overall, there is almost no agreement in the interaction effects identified across the different data modalities. In summary, our findings suggests that abundance information is overall more informative than presence-absence information and that whether we identify an interaction between two taxa is highly dependent on the underlying data modality.

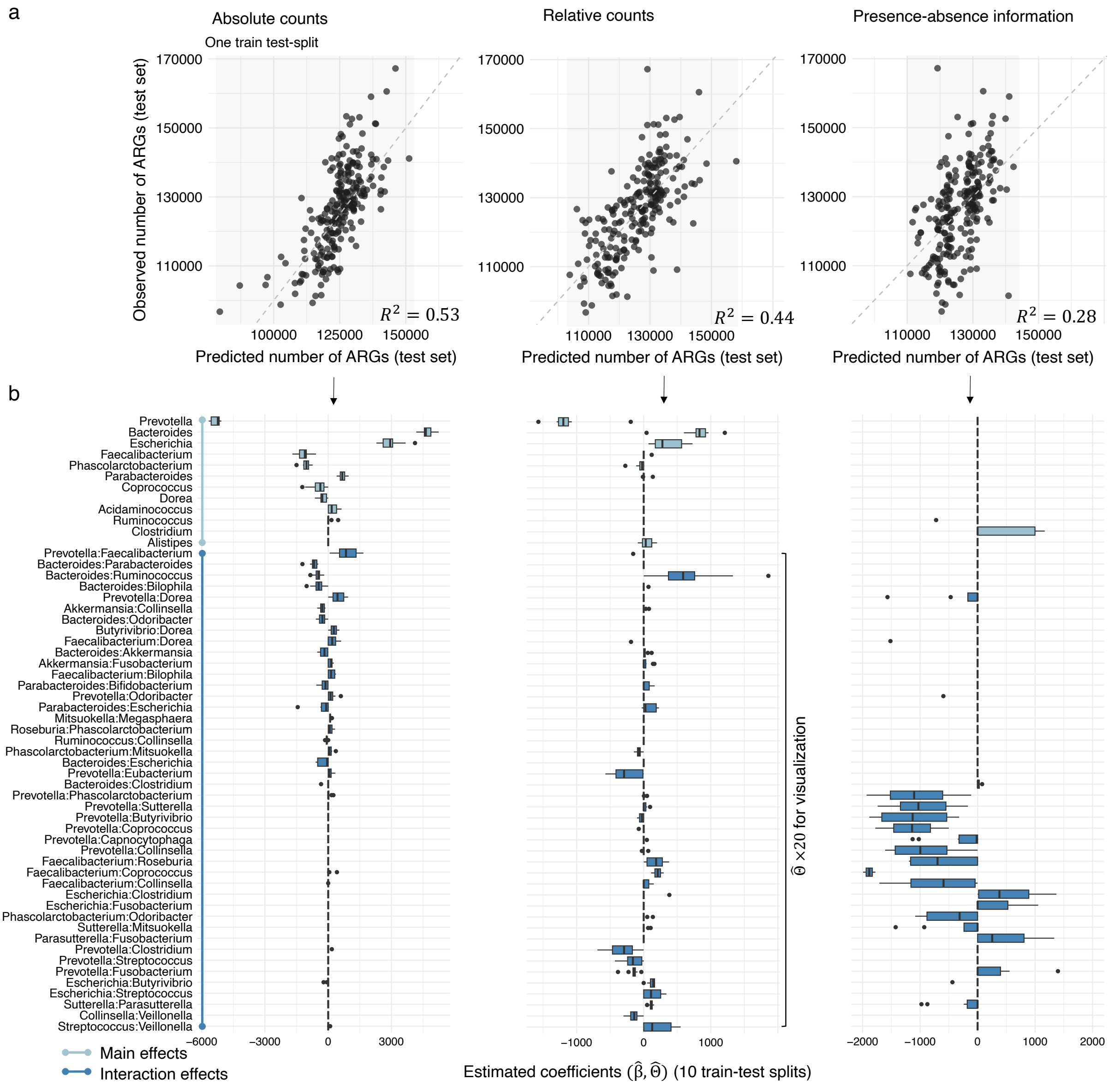


Fig 6. **a**, Scatterplots (one train test split) and test set R squared R^2 (average over 10 train test splits) for the comparison of the predicted number of anti microbial resistance genes (ARGs) and the observed number of ARGs on a test dataset based on the absolute count information of genera, the relative count information of genera and the presence-absence information of genera of the MetaCardis cohort. **b**, Distribution of coefficients over 10 train test splits for the superset of coefficients that are non-zero in one of the three data modalities.

4 Discussion

Identifying predictive and interpretable main and interaction effects between microbial taxa that can be related to ecological, host-associated or environmental features is a cornerstone of statistical data analysis of microbiome data. To this end, we have introduced a generic modeling approach for quadratic interactions, that is further instantiated to accommodate three distinct data types, in which microbial abundance information typically appears: (i) quantitative microbiome, (ii) presence-absence information, or (iii) relative count information. For these three data modalities we introduce a generic optimization problem, that allows penalized model estimation to estimate parsimonious models, where the number of features p (here the microbial taxa) and pairwise interactions often exceeds the number of measurements n. Our interaction modeling strategy combines existing approaches for sparse interaction modeling and introduces novel ways of modeling interaction effects in the compositional setting. In the realm of interpretability, we combine the quadratic interaction modeling framework with the concept of hierarchical interactions [34] and stability selection [38, 37] as optional extensions. We introduce three mathematically equivalent ways of modeling quadratic interactions with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic log-ratio model. The three models differ in terms of interpretability, dimensionality, and optimization. In the $\tilde{p} + \tilde{p}(\tilde{p} - 1)/2$-dimensional alr transformed quadratic model, where $\tilde{p} = p - 1$, the interpretation of coefficients is consistently tied to a reference feature, which is enforced to be part of the model in the regularized model case. This model can be optimized without constraints on the main effects and can be flexibly extended to hierarchical interactions. The $p + p(p - 1)/2$-dimensional quadratic log-contrast model requires a zero-sum constraint on the main effects but offers a more convenient expression that does not require the assignment of a reference feature. In this model, each main effect is interpreted with respect to all other features. The $2 \cdot p(p - 1)/2$-dimensional quadratic log-ratio model does not require a constraint on the main effect coefficients in the optimization, as this property is automatically met, and allows for the interpretation of the relative effects between all pairs of features. Notably, the interpretation of the interaction coefficients in both the quadratic log-contrast model and the quadratic log-ratio model are identical. Although the log-contrast model in the linear case enjoys the clear advantage of lower dimensionality compared to the log-ratio model, this distinction becomes less relevant when introducing interactions. Hence, the primary criterion for selecting between the models should be based on the preferred interpretation.

We demonstrate the broad applicability of our framework by analyzing microbial data covering all three data modalities across various ecosystems, including synthetic microbial communities, human gut microbiomes, and marine microbial ecosystems. We show how quadratic models can improve the predictive performance compared to linear models on real-world data and illustrate how main and interaction effects can be robustly estimated by integrating hierarchical interaction modeling and stability selection. Notably, using a synthetic community dataset from [24], we identified a strong inhibition of butyrate production by *A. caccae* when *D. piger* is present. We demonstrate that this is the only robust and consequently relevant combinatorial effect within this community. Additionally, we generate semi-synthetic data to demonstrate the ability of the sparse quadratic interaction model for relative microbiome data to accurately detect interaction effects. Based on this, we showcase how sparse an interaction feature can be in order to achieve an accurate estimation in our model. Further, we demonstrate, for varying noise levels, how a misspecified main effects model tends to inaccurately estimate effects when true interactions are present. Finally, we perform a comparative analysis of the quadratic interaction models for the three data modalities discussed in this study, demonstrating that absolute microbial counts and compositions achieve similarly strong predictive performance when predicting the number of antimicrobial resistance genes (ARGs). In contrast, transforming abundance data into presence-absence information results in a decline in predictive accuracy. Our analysis suggests that the identified (main and) interaction effects highly depend on the underlying data type. Moreover, the identified effects of individual genera and interactions between them are an interesting finding in themselves. Many of the relevant features correspond to the definition of enterotypes [68], as they were defined for this specific data in [69], suggesting a potential link between enterotypes and ARGs.

As more and larger data sets become available, our models can be extended to higher-order interactions or more complex, arbitrary non-linear effects. Moreover, our results suggest that different data modalities carry different information, and it might be meaningful to include multiple layers of information in statistical prediction tasks with interaction effects.

In summary, we believe that our framework and its implementation in R provide a valuable tool to study robust interaction effects in microbiome data of different modalities and distinct ecosystems, ranging from synthetic communities and community function studies in microbial ecology to observational microbiome data linking the microbiome to the host or environment.

References

[1] A. Konopka. “What is microbial community ecology?” In: *The ISME journal* 3.11 (2009), pp. 1223–1230.

[2] E. J. Culp and A. L. Goodman. “Cross-feeding in the gut microbiome: Ecology and mechanisms”. In: *Cell Host & Microbe* 31.4 (2023), pp. 485–499.

[3] R. M. Braga, M. N. Dourado, and W. L. Araújo. “Microbial interactions: ecology in a molecular perspective”. In: *brazilian journal of microbiology* 47 (2016), pp. 86–98.

[4] W. Z. Lidicker Jr. “A clarification of interactions in ecological systems”. In: *Bioscience* 29.8 (1979), pp. 475–477.

[5] K. Faust and J. Raes. “Microbial interactions: from networks to models”. In: *Nature Reviews Microbiology* 10.8 (2012), pp. 538–550.

[6] N. Weiland-Bräuer. “Friends or foes—microbial interactions in nature”. In: *Biology* 10.6 (2021), p. 496.

[7] J. Friedman and E. J. Alm. “Inferring correlation networks from genomic survey data”. In: (2012).

[8] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. “Sparse and compositionally robust inference of microbial ecological networks”. In: *PLoS computational biology* 11.5 (2015), e1004226.

[9] J. Tackmann, J. F. M. Rodrigues, and C. von Mering. “Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data”. In: *Cell systems* 9.3 (2019), pp. 286–296.

[10] G. Yoon, I. Gaynanova, and C. L. Müller. “Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data”. In: *Frontiers in genetics* 10 (2019), p. 516.

[11] S. Peschel, C. L. Müller, E. Von Mutius, A.-L. Boulesteix, and M. Depner. “NetCoMi: network construction and comparison for microbiome data in R”. In: *Briefings in bioinformatics* 22.4 (2021), bbaa290.

[12] F. G. Blanchet, K. Cazelles, and D. Gravel. “Co-occurrence is not evidence of ecological interactions”. In: *Ecology Letters* 23.7 (2020), pp. 1050–1063.

[13] B. Ma, Y. Wang, S. Ye, S. Liu, E. Stirling, J. A. Gilbert, K. Faust, R. Knight, J. K. Jansson, C. Cardona, et al. “Earth microbial co-occurrence network reveals interconnection pattern across microbiomes”. In: *Microbiome* 8.1 (2020), pp. 1–12.

[14] T. Zamkovaya, J. S. Foster, V. de Crécy-Lagard, and A. Conesa. “A network approach to elucidate and prioritize microbial dark matter in microbial communities”. In: *The ISME journal* 15.1 (2021), pp. 228–244.

[15] A. Skwara, K. Gowda, M. Yousef, J. Diaz-Colunga, A. S. Raman, A. Sanchez, M. Tikhonov, and S. Kuehn. "Statistically learning the functional landscape of microbial communities". In: *Nature Ecology & Evolution* 7.11 (2023), pp. 1823–1833.

[16] J. Bien, X. Yan, L. Simpson, and C. L. Müller. "Tree-aggregated predictive modeling of microbiome data". In: *Scientific Reports* 11.1 (2021), p. 14505.

[17] A. S. Weiss, A. G. Burrichter, A. C. Durai Raj, A. von Strempel, C. Meng, K. Kleigrewe, P. C. Münch, L. Rössler, C. Huber, W. Eisenreich, et al. "In vitro interaction network of a synthetic gut bacterial community". In: *The ISME journal* 16.4 (2022), pp. 1095–1109.

[18] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. "Microbiome datasets are compositional: and this is not optional". In: *Frontiers in microbiology* 8 (2017), p. 294209.

[19] B. J. Callahan, P. J. McMurdie, and S. P. Holmes. "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis". In: *The ISME journal* 11.12 (2017), pp. 2639–2643.

[20] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, et al. "Quantitative microbiome profiling links gut community variation to microbial load". In: *Nature* 551.7681 (2017), pp. 507–511.

[21] A. Tkacz, M. Hortala, and P. S. Poole. "Absolute quantitation of microbiota abundance in environmental samples". In: *Microbiome* 6 (2018), pp. 1–13.

[22] C. Jian, P. Luukkonen, H. Yki-Järvinen, A. Salonen, and K. Korpela. "Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling". In: *PloS one* 15.1 (2020), e0227285.

[23] G. Galazzo, N. Van Best, B. J. Benedikter, K. Janssen, L. Bervoets, C. Driessen, M. Oomen, M. Lucchesi, P. H. van Eijck, H. E. Becker, et al. "How to count our microbes? The effect of different quantitative microbiome profiling approaches". In: *Frontiers in cellular and infection microbiology* 10 (2020), p. 403.

[24] R. L. Clark, B. M. Connors, D. M. Stevenson, S. E. Hromada, J. J. Hamilton, D. Amador-Noguez, and O. S. Venturelli. "Design of synthetic human gut microbiome assembly and butyrate production". In: *Nature communications* 12.1 (2021), p. 3254.

[25] A. L. Gould, V. Zhang, L. Lamberti, E. W. Jones, B. Obadia, N. Korasidis, A. Gavryushkin, J. M. Carlson, N. Beerenwinkel, and W. B. Ludington. "Microbiome interactions shape host fitness". In: *Proceedings of the National Academy of Sciences* 115.51 (2018), E11951–E11960.

[26] J. Aitchison and J. Bacon-Shone. "Log contrast models for experiments with mixtures". In: *Biometrika* 71.2 (1984), pp. 323–330.

[27] W. Lin, P. Shi, R. Feng, and H. Li. “Variable selection in regression with compositional covariates”. In: *Biometrika* 101.4 (2014), pp. 785–797.

[28] P. Shi, A. Zhang, and H. Li. “Regression analysis for microbiome compositional data”. In: (2016).

[29] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. “Balances: a new perspective for microbiome analysis”. In: *MSystems* 3.4 (2018), pp. 10–1128.

[30] S. Bates and R. Tibshirani. “Log-ratio lasso: scalable, sparse estimation for log-ratio models”. In: *Biometrics* 75.2 (2019), pp. 613–624.

[31] P. L. Combettes and C. L. Müller. “Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications”. In: *Statistics in Biosciences* 13.2 (2021), pp. 217–242.

[32] B. Yu. “Stability”. In: *Bernoulli* 19.4 (2013), pp. 1484–1500.

[33] M. Stadler, S. Lukauskas, T. Bartke, and C. L. Mueller. “asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers”. In: *bioRxiv* (2024), pp. 2024–03.

[34] J. Bien, J. Taylor, and R. Tibshirani. “A lasso for hierarchical interactions”. In: *The Annals of Statistics* 41.3 (June 2013).

[35] M. Hamada and C. J. Wu. “Analysis of designed experiments with complex aliasing”. In: *Journal of quality technology* 24.3 (1992), pp. 130–137.

[36] J. L. Peixoto. “Hierarchical variable selection in polynomial regression models”. In: *The American Statistician* 41.4 (1987), pp. 311–313.

[37] N. Meinshausen and P. Bühlmann. “Stability Selection”. In: *Journal of the Royal Statistical Society, Series B* 72 (2010), pp. 417–473.

[38] R. D. Shah and R. J. Samworth. “Variable selection with error control: Another look at stability selection”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75.1 (2013), pp. 55–80.

[39] S. K. Forslund, R. Chakaroun, M. Zimmermann-Kogadeeva, L. Markó, J. Aron-Wisnewsky, T. Nielsen, L. Moitinho-Silva, T. S. Schmidt, G. Falony, S. Vieira-Silva, et al. “Combinatorial, additive and dose-dependent drug–microbiome associations”. In: *Nature* 600.7889 (2021), pp. 500–505.

[40] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, et al. “Structure and function of the global ocean microbiome”. In: *Science* 348.6237 (2015), p. 1261359.

[41] F. J. Poelwijk, V. Krishna, and R. Ranganathan. “The context-dependence of mutations: a linkage of formalisms”. In: *PLoS computational biology* 12.6 (2016), e1004771.

[42] E. D. Weinberger. “Fourier and Taylor series on fitness landscapes”. In: *Biological cybernetics* 65.5 (1991), pp. 321–330.

[43] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[44] J. Bien and R. Tibshirani. *hierNet: A Lasso for Hierarchical Interactions.* R package version 1.9. 2020.

[45] J. Lederer and C. Müller. “Don’t fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX”. In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 29. 1. 2015.

[46] Y. Wu and L. Wang. “A survey of tuning parameter selection for high-dimensional regression”. In: *Annual review of statistics and its application* 7 (2020), pp. 209–226.

[47] N. Hao, Y. Feng, and H. H. Zhang. “Model selection for high-dimensional quadratic regression via regularization”. In: *Journal of the American Statistical Association* 113.522 (2018), pp. 615–625.

[48] H. Liu, K. Roeder, and L. Wasserman. “Stability approach to regularization selection (stars) for high dimensional graphical models”. In: *Advances in neural information processing systems* 23 (2010).

[49] B. Bodinier, S. Filippi, T. H. Nøst, J. Chiquet, and M. Chadeau-Hyam. “Automated calibration for stability selection in penalised regression and graphical models”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* (2023), qlad058.

[50] S. Maddu, B. L. Cheeseman, I. F. Sbalzarini, and C. L. Müller. “Stability selection enables robust learning of differential equations from limited noisy data”. In: *Proceedings of the Royal Society A* 478.2262 (2022), p. 20210916.

[51] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton. “Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control”. In: *Proceedings of the Royal Society A* 478.2260 (2022), p. 20210904.

[52] B. Hofner and T. Hothorn. *stabs: Stability Selection with Error Control.* R package version 0.6-4. 2021.

[53] K. Lee, S. Raguideau, K. Sirén, F. Asnicar, F. Cumbo, F. Hildebrand, N. Segata, C.-J. Cha, and C. Quince. “Population-level impacts of antibiotic usage on the human gut microbiome”. In: *Nature Communications* 14.1 (2023), p. 1191.

[54] V. Singh, G. Lee, H. Koh, T. Unno, and J.-H. Shin. “Butyrate producers,“The Sentinel of Gut”: Their intestinal significance with and beyond butyrate, and prospective use as microbial therapeutics”. In: *Frontiers in microbiology* 13 (2023), p. 1103836.

[55] A. Schwiertz, G. L. Hold, S. H. Duncan, B. Gruhl, M. D. Collins, P. A. Lawson, H. J. Flint, and M. Blaut. “Anaerostipes caccae gen. nov., sp. nov., a new saccharolytic, acetate-utilising, butyrate-producing bacterium from human faeces”. In: *Systematic and applied microbiology* 25.1 (2002), pp. 46–51.

[56] P. Marquet, S. H. Duncan, C. Chassard, A. Bernalier-Donadille, and H. J. Flint. “Lactate has the potential to promote hydrogen sulphide formation in the human colon”. In: *FEMS Microbiology Letters* 299.2 (2009), pp. 128–134.

[57] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, et al. “American gut: an open platform for citizen science microbiome research”. In: *Msystems* 3.3 (2018), pp. 10–1128.

[58] H. S. Tee, D. Waite, G. Lear, and K. M. Handley. “Microbial river-to-sea continuum: gradients in benthic and planktonic diversity, osmoregulation and nutrient cycling”. In: *Microbiome* 9 (2021), pp. 1–18.

[59] L. Wang, C. Lian, W. Wan, Z. Qiu, X. Luo, Q. Huang, Y. Deng, T. Zhang, and K. Yu. “Salinity-triggered homogeneous selection constrains the microbial function and stability in lakes”. In: *Applied Microbiology and Biotechnology* 107.21 (2023), pp. 6591–6605.

[60] S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, et al. “Tara Oceans: towards global ocean ecosystems biology”. In: *Nature Reviews Microbiology* 18.8 (2020), pp. 428–445.

[61] R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, et al. “Metagenomic 16S rDNA I llumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities”. In: *Environmental microbiology* 16.9 (2014), pp. 2659–2671.

[62] D. P. Herlemann, J. Woelk, M. Labrenz, and K. Jürgens. “Diversity and abundance of “Pelagibacterales”(SAR11) in the Baltic Sea salinity gradient”. In: *Systematic and applied microbiology* 37.8 (2014), pp. 601–604.

[63] B. J. Campbell, S. J. Lim, and D. L. Kirchman. “Controls of SAR11 subclade abundance, diversity, and growth in two Mid-Atlantic estuaries”. In: *bioRxiv* (2022), pp. 2022–05.

[64] Y. Tada, R. Makabe, N. Kasamatsu-Takazawa, A. Taniguchi, and K. Hamasaki. “Growth and distribution patterns of Roseobacter/Rhodobacter, SAR11, and Bacteroidetes lineages in the Southern Ocean”. In: *Polar biology* 36 (2013), pp. 691–704.

[65] S. Kraemer, A. Ramachandran, D. Colatriano, C. Lovejoy, and D. A. Walsh. “Diversity and biogeography of SAR11 bacteria from the Arctic Ocean”. In: *The ISME Journal* 14.1 (2020), pp. 79–90.

[66] N. O. Eltai, A. A. Al Thani, S. H. Al Hadidi, K. Al Ansari, and H. M. Yassine. “Antibiotic resistance and virulence patterns of pathogenic Escherichia coli strains associated with acute gastroenteritis among children in Qatar”. In: *BMC microbiology* 20 (2020), pp. 1–12.

[67] S. Valdezate, F. Cobo, S. Monzón, M. J. Medina-Pascual, Á. Zaballos, I. Cuesta, S. Pino-Rosa, and P. Villalón. “Genomic background and phylogeny of cfi A-positive Bacteroides fragilis strains resistant to meropenem-EDTA”. In: *Antibiotics* 10.3 (2021), p. 304.

[68] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, et al. “Enterotypes of the human gut microbiome”. In: *nature* 473.7346 (2011), pp. 174–180.

[69] S. Vieira-Silva, G. Falony, E. Belda, T. Nielsen, J. Aron-Wisnewsky, R. Chakaroun, S. K. Forslund, K. Assmann, M. Valles-Colomer, T. T. D. Nguyen, et al. “Statin therapy is associated with lower prevalence of gut microbiota dysbiosis”. In: *Nature* 581.7808 (2020), pp. 310–315.

Supplementary information

Binary encoding of covariates in regression models

Whether to encode the input data in a regression model as $B \in \{0,1\}^{n \times p}$ or as $B \in \{-1,1\}^{n \times p}$ has an impact on the interpretations of the model coefficients. In the main effects model case with $p = 1$, for $B \in \{0,1\}^{n \times p}$, the outcome Y would be given by

$$Y = \begin{cases} \beta_0 & \text{if } B_1 = 0 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1. \end{cases}$$

The interpretation of β_0 in this case is the effect of 'absence' and the interpretation of β_1 is the difference between the effect of 'presence' and the effect of 'absence'. The interaction model, illustrated for $p = 2$, is given by

$$Y = \begin{cases} \beta_0 & \text{if } B_1 = 0 \text{ and } B_2 = 0 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1 \text{ and } B_2 = 0 \\ \beta_0 + \beta_2 & \text{if } B_1 = 0 \text{ and } B_2 = 1 \\ \beta_0 + \beta_1 + \beta_2 + \Theta_{12} & \text{if } B_1 = 1 \text{ and } B_2 = 1. \end{cases}$$

Here, β_0 is the effect of co-absence, and β_j is the effect of the difference of co-absence and presence of B_j, for $j = 1, 2$. The interaction term Θ_{12} is the additional effect when both features are 1.
If $B \in \{-1,1\}^{n \times p}$, the outcome Y in the main effects model is given by

$$Y = \begin{cases} \beta_0 - \beta_1 & \text{if } B_1 = -1 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1. \end{cases}$$

The interpretation here is that β_0 is the mean effect of the two group means, and $2\beta_1$ is the difference of the two conditions in mean.
In the interaction model Y is given by

$$Y = \begin{cases} \beta_0 - \beta_1 - \beta_2 + \Theta_{12} & \text{if } B_1 = -1 \text{ and } B_2 = -1 \\ \beta_0 + \beta_1 - \beta_2 - \Theta_{12} & \text{if } B_1 = 1 \text{ and } B_2 = -1 \\ \beta_0 - \beta_1 + \beta_2 - \Theta_{12} & \text{if } B_1 = -1 \text{ and } B_2 = 1 \\ \beta_0 + \beta_1 + \beta_2 + \Theta_{12} & \text{if } B_j = 1 \text{ and } B_k = 1. \end{cases}$$

Now, β_0 represents the mean of the four group means (if the design is completely balanced this is the overall mean). The main effect coefficients β_j, $j \in \{1,2\}$ are the average difference effects between the two conditions the respective feature can take. The interaction effect

$2\Theta_{jk}$ explains the difference between the two conditions when either both features are present or absent and when only one of the two features is present.

Moreover, there exists a linear transformation between the coefficients of both encodings. We denote all coefficients in the 0 and 1 encoding as $\tilde{\beta}$ and the coefficients in the -1 and 1 encoding as β. The transformation between both encodings in the quadratic interaction model for $p = 2$ is given by

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 - \beta_1 - \beta_2 + \Theta_{12} \\
\tilde{\beta}_1 &= 2(\beta_1 - \Theta_{12}) \\
\tilde{\beta}_2 &= 2(\beta_2 - \Theta_{12}) \\
\tilde{\Theta}_{12} &= 4\Theta_{12}.
\end{aligned}$$

For $p \geq 2$ this can be translated to a general form as by

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 - \sum_{j=1}^{p} \beta_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \Theta_{jk} \\
\tilde{\beta}_j &= 2\beta_j - 2\sum_{\substack{k=1 \\ k \neq j}}^{p} \Theta_{jk}, \text{ for } j = 1, ..., p \\
\tilde{\Theta}_{jk} &= 4\Theta_{jk}, \text{ for } j = 1, ..., p-1, \; k = j+1, ..., p.
\end{aligned}$$

The transformation from one encoding to the other can be derived by replacing the input matrix in the model according to this equation: $B^{\{-1,1\}} = 2B^{\{0,1\}} - 1$.

Alr transformed model versus constrained log contrast model

Main effects only:

$$ALR = \sum_{j=1}^{p-1} \beta_j \log \frac{X_j}{X_p} = \sum_{j=1}^{p-1} \beta_j \log X_j - \left(\sum_{j=1}^{p-1} \beta_j \right) \log X_p$$

Thus, we can define $\beta_p := -\sum_{j=1}^{p-1} \beta_j$ and then write this as $\sum_{j=1}^{p} \beta_j \log X_j$ s.t. $\sum_{j=1}^{p} \beta_j = 0$,

which corresponds to the log contrast model.

Model with interactions:

$$
\begin{aligned}
ALR &= \sum_{j=1}^{p-1} \beta_j \log \frac{X_j}{X_p} + \sum_{1 \le j,k \le p} \Theta_{jk} \log \frac{X_j}{X_p} \log \frac{X_k}{X_p} \\
&= \sum_{j=1}^{p-1} \beta_j \log X_j + \left(- \sum_{j=1}^{p-1} \beta_j \right) \log X_p \\
&+ \sum_{1 \le j < k < p} \Theta_{jk} \left(\log X_j \log X_k - \log X_j \log X_p - \log X_k \log X_p + \log^2 X_p \right) \\
&= \sum_{j=1}^{p} \beta_j \log X_j \quad (\text{taking } \sum_{j=1}^{p} \beta_j := 0) \\
&+ \sum_{1 \le j,k \le p} \Theta_{jk} \log X_j \log X_k - 2 \log X_p \sum_{j=1}^{p-1} \left(\sum_{k=1}^{p-1} \Theta_{jk} \right) \log X_j + \left(\sum_{j,k} \Theta_{jk} \right) \log^2 X_p
\end{aligned}
$$

For $j < p, \Theta_{jp} = \Theta_{pj} := - \sum_{k=1}^{p-1} \Theta_{jk}$

For $\Theta_{pp} := \sum_{1 \le j,k \le p} \theta_{jk}$.

Note: $\sum_{j=1}^{p} \Theta_{jp} = \sum_{j=1}^{p-1} \left(- \sum_{k=1}^{p-1} \Theta_{jk} \right) + \left(\sum_{1 \le j,k < p} \Theta_{jk} \right) = 0$

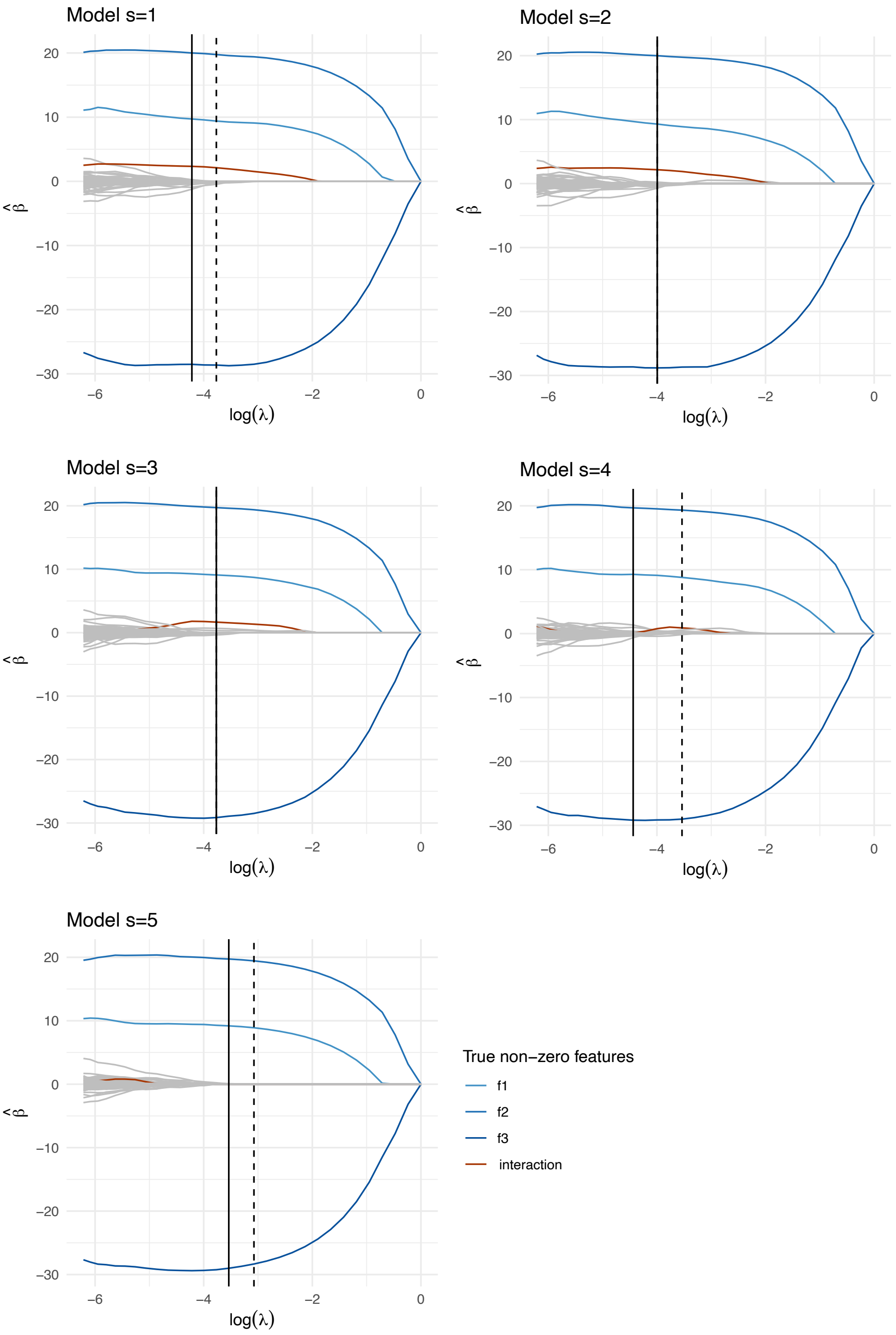


Fig S1. Solution path of the interaction model (sparse qlc) for the $S = 5$ semi-synthetic simulation setups for varying feature sparsity levels.

A.3 A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach

Contributing article

Stadler, M.*, Kupczyk, E.*, Buckett, L., Zhang, X., Müller, C.L. (2024). A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach. *Draft manuscript.* *joint first co-authorship

Replication code

The source data and code for reproducing all results of this study is available at https://github.com/marastadler/Drug-interactions-HCS.

Author contributions

M.S. developed and implemented the statistical approaches and data analyses. C.L.M. supervised the work. C.L.M. and M.S. conceived the statistical models. E.K. performed computational analyses and performed data pre-processing. X.Z. performed lab experiments. E.K. and L.B. will analyze the results and provide feedback.

A statistical framework for robust drug interaction estimation with a high-content screening cell painting approach

Mara Stadler, Erwin Kupczyk, Lance Bucket, Xin Zhan, Christian L. Müller

Introduction

High Content Screening (HCS) generates large quantities of cell morphological data (e.g., nucleus size or cell shape) derived from microscopy images under various chemical conditions or drug combinations [1], making it a popular tool for drug discovery [2]. The morphological features are typically presented as summary statistics across multiple single cells [3]. A key question includes how certain drugs and combinations of drugs influence the morphology of cells. Not all morphological features may carry relevant information, and some might be redundant.

In this study, we develop a combinatorial drug design involving 20 distinct compounds across 408 experiments to analyze cell morphological features within HCS experiments. From this data, a generic computational framework is developed to examine stable drug interaction effects. This framework incorporates a robustified version of hierarchical interaction modeling, combined with stability-based model selection from [4]. Instead of reducing the feature space of morphological features before conducting further statistical analyses, as suggested in previous studies [5, 6, 7], the framework presented in this study examines both main and interaction effects across all features and subsequently employs a post-estimation clustering approach. For each of the inferred clusters, prototypical morphological features are statistically determined after being associated with the combinatorial drug design. In summary, this contribution introduces a generalizable computational tool (available at `https://github.com/marastadler/Drug-interactions-HCS`) that uncovers combinatorial drug effects and efficiently reduces the morphological feature space in HCS studies.

Materials and methods

Robust hierarchical modeling

Our objective is to predict characteristics that describe the morphology of single cells given as $p_2 = 68$ summary statistics over multiple cells derived from HCS cell painting, denoted by $(Y_i)_{1 \leq i \leq p_2}$, based on a binary combinatorial design of $p_1 = 20$ compounds $(B_j)_{1 \leq j \leq p_1}$ across $n = 408$ experiments (see Fig. 1).

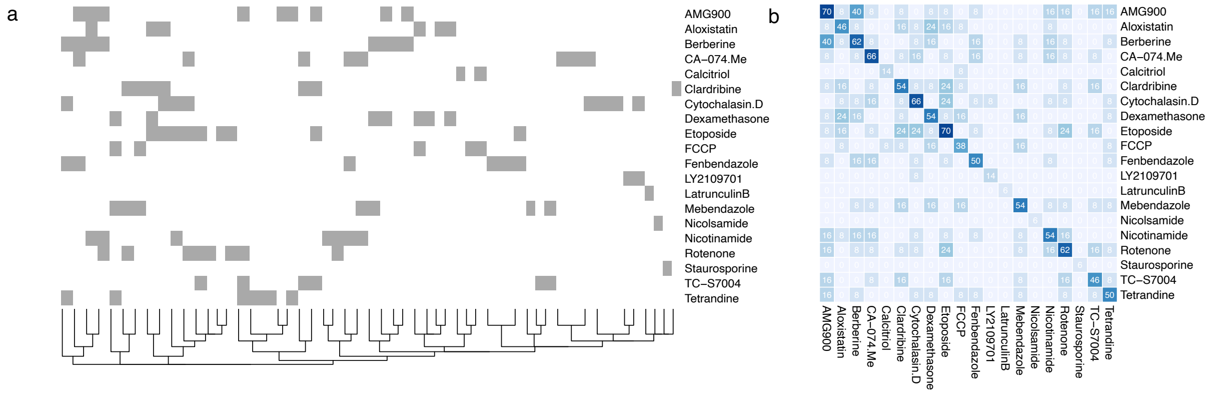


Fig 1. **a**, Heatmap representing the experimental design with $p_1 = 20$ compounds across $n = 408$ experiments. **b**, Heatmap representing how often compounds co-occur within the same experiment.

Utilizing the binary design matrix B, the interaction model for revealing the additive effects of compounds on a morphological feature $y = Y_i \in \mathbb{R}^n, i = 1, \dots, p_2$, is represented by the linear model

$$y = \beta_0 + \sum_{j=1}^{p_1} \beta_j B_j + \epsilon, \tag{1}$$

where $\beta_0 \in \mathbb{R}^n$ is a feature-specific intercept, β_j is the effect of compound j on the ith morphological feature y, and ϵ models the noise component.
We extend the linear model by incorporating all pairwise interactions among compounds. For each morphological feature represented as $y = Y_i$, $i = 1, \dots, p_2$, the fundamental model is expressed as follows

$$y = \beta_0 + \sum_{j=1}^{p_1} \beta_j B_j + \frac{1}{2} \sum_{j=1}^{p_1} \sum_{k=1}^{p_1} \Theta_{jk} B_j B_k + \epsilon, \tag{2}$$

with Θ_{jk} denoting the interaction effect between compound j and k. The sign of the interaction effect, Θ_{jk}, facilitates the characterization of compound interplay. For instance, a positive interaction coefficient $\Theta_{jk} > 0$ implies that compounds j and k exhibit a synergistic binding effect when both $\beta_j > 0$ and $\beta_k > 0$. To derive parsimonious models, we employ regularized maximum-likelihood estimation with ℓ_1-norm (lasso) penalization [8] on the linear and interaction coefficients. The objective is to minimize the log-likelihood function of the model, denoted as $l(\beta_0, \beta, \Theta) = \left\| y - \beta_0 - B\beta - \frac{1}{2} B \Theta B^T \right\|_2^2$. The lasso problem for all pairs is formulated as follows:

$$\min_{\beta_0, \beta, \Theta} \; l(\beta_0, \beta, \Theta) + \lambda \left\| \beta \right\|_1 + \frac{\lambda}{2} \left\| \Theta \right\|_1, \tag{3}$$

Here, $\lambda > 0$ serves as a tuning parameter, regulating the sparsity levels of the coefficients β and Θ, respectively. To enhance interpretability and adhere to the statistical principle of weak hierarchy, we allow for the presence of an interaction in the model only if at least one associated linear effect is also included [9]. Mathematically, this weak hierarchy principle is expressed as

$$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0 \text{ or } \hat{\beta}_k \neq 0,$$

signifying that interaction effects are present only if either one of the linear effects or both enter the model. This constraint is implemented through a hierarchical interaction optimization problem

$$\begin{aligned} &\min_{\beta,\Theta} \; l(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1 \\ &\text{s.t.} \quad \left\|\Theta_j\right\|_1 \leq |\beta_j| \, . \end{aligned} \tag{4}$$

To address the non-convex nature of the problem in (4), we adopt the convex relaxation proposed by [9] and utilize the efficient implementation provided by the R package `hierNet` [10] (v1.9). Employing `hierNet`, we model each morphological feature $y = Y_i$, where $i = 1, ..., p_2$, with hierarchical interactions. In addition to reducing the number of spurious interaction effects, the weak hierarchy constraint introduces the concept of "practical sparsity" favoring models that effectively "reuse" measured variables.

Robust learning with Huber loss

In this section, we introduce a method to improve the robustness of each outcome $y = Y_i$, $i = 1, ..., p_2$, in a regression model. The Huber loss function, first introduced in [11], serves as a robust alternative to the squared error loss (L2). It is widely used in (penalized) regression analysis because of its resistance to outliers [12]. The function merges the L2 loss for smaller residuals with an absolute loss (L1) for larger residuals, reducing the impact of outliers on the estimation of parameters. The convex nature of the Huber loss ensures the presence of a unique minimum, which is beneficial for optimization challenges in statistical learning. In order to improve the robustness within the hierarchical interaction modeling framework, we incorporate the Huber loss as a robust alternative. The formulation of the Huber loss function, defined for $r = l(\beta_0, \beta, \Theta)^{1/2}$, is expressed as

$$l_{\text{Huber}}(\beta_0, \beta, \Theta, \delta) = \begin{cases} \frac{1}{2} r^2 & \text{for } |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \tag{5}$$

where δ is an adjustable parameter that dictates the shift from quadratic (L2) to linear (L1) behavior in the loss function [11]. This Huber loss can be seamlessly integrated into the interaction modeling approach, replacing the L2 loss in the optimization framework.

Stability-based Model Selection for Hierarchical Interactions

A major challenge in penalized regression lies in determining an appropriate regularization parameter λ. This parameter balances the sparsity (i.e., interpretability) of the model coefficients with the out-of-sample predictive performance of the model [13, 14]. A standard technique to determine an optimal λ involves cross-validation [9] as well as Information Criteria such as the Akaike Information Criterion (AIC) and the extended Bayesian Information Criterion (EBIC) [15]. While these approaches tend to select more predictors than necessary [15], we follow [4] and incorporate the principle of stability [16]. Specifically, we introduce stability selection [17] as a method to detect a stable set of predictive main and interaction effects based on a certain selection method (e.g., the k-largest features). In regression modeling, stability selection iteratively fits sparse models from b random subsamples of fixed size (e.g., $n_s = \lfloor n/2 \rfloor$). This allows defining so-called selection probabilities $\hat{\pi} \in [0,1]^{p_1+p_1(p_1-1)/2}$ that describe the frequency of all selected features across all subsamples. Ultimately, the features above a certain threshold $\pi_{\text{thr}} \in [0,1]$ define the set of predictors in the final regression model. Here, we employ complementary pairs stability selection (CPSS) [18] which is a variant of stability selection. CPSS draws subsamples as complementary pairs of samples $\{1, ...n\}$. This approach proves particularly advantageous when dealing with unbalanced experimental designs and limited sample sizes, ensuring that individual samples are evaluated equally often. As default, we use Λ as the internal λ-path in [10], $\pi_{\text{thr}} = .6$, $k = 15$, and $b = 50$, resulting in 100 subsamples.

Consistent feature reduction

The analysis of high-dimensional single-cell level microscopy data obtained through High-Content Screening (HCS) encounters challenges due to the vast number of morphological features. Various strategies have been suggested to address this issue and streamline the analysis of cellular phenotypes in HCS datasets. These approaches encompass techniques such as factor analysis, correlation elimination, and interactive selection, all aimed at reducing the multitude of parameters associated with the data [19]. Moreover, multiple benchmarking studies for feature selection methods for compressing image information in HCS have been performed [6, 20].

Our statistical interaction-detection framework operates on a feature-wise basis, which ensures that the outcomes remain independent of the initial set of morphological features under analysis. This independence provides the flexibility to conduct the analysis on all available morphological features without the need to reduce the feature space prior to the analysis. Instead, we introduce a post-estimation clustering strategy. This strategy performs hierarchical clustering with prototypes via minimax linkage, as introduced by [21]. We perform the clustering on a common feature effect matrix that combines all main and interaction effects for each morphological feature Y_i, $i = 1, ..., p_2$. This matrix is given by $[\hat{\beta}, \hat{\Theta}] \in \mathbb{R}^{p_1+p_1(p_1-1)/2 \times p_2}$. The clustering approach structures the morphological features based on their identified responses to individual and pairwise drug effects. It also assigns prototypical features to each cluster that best represent the cluster, providing a condensed view of the results.

Results

Our interaction modeling strategy reveals that certain cell morphological features respond similarly to both individual and combinatorial drug effects, suggesting redundancy in the features analyzed. In Fig. 2, we present a clustered representation of the estimated main effects $\hat{\beta}$ (left) and interaction effects $\hat{\Theta}$ (right) using hierarchical clustering with prototypes via minimax linkage. Each cluster is represented by a prototypical morphological feature, highlighted with an asterisk (*), providing a condensed view of the data.

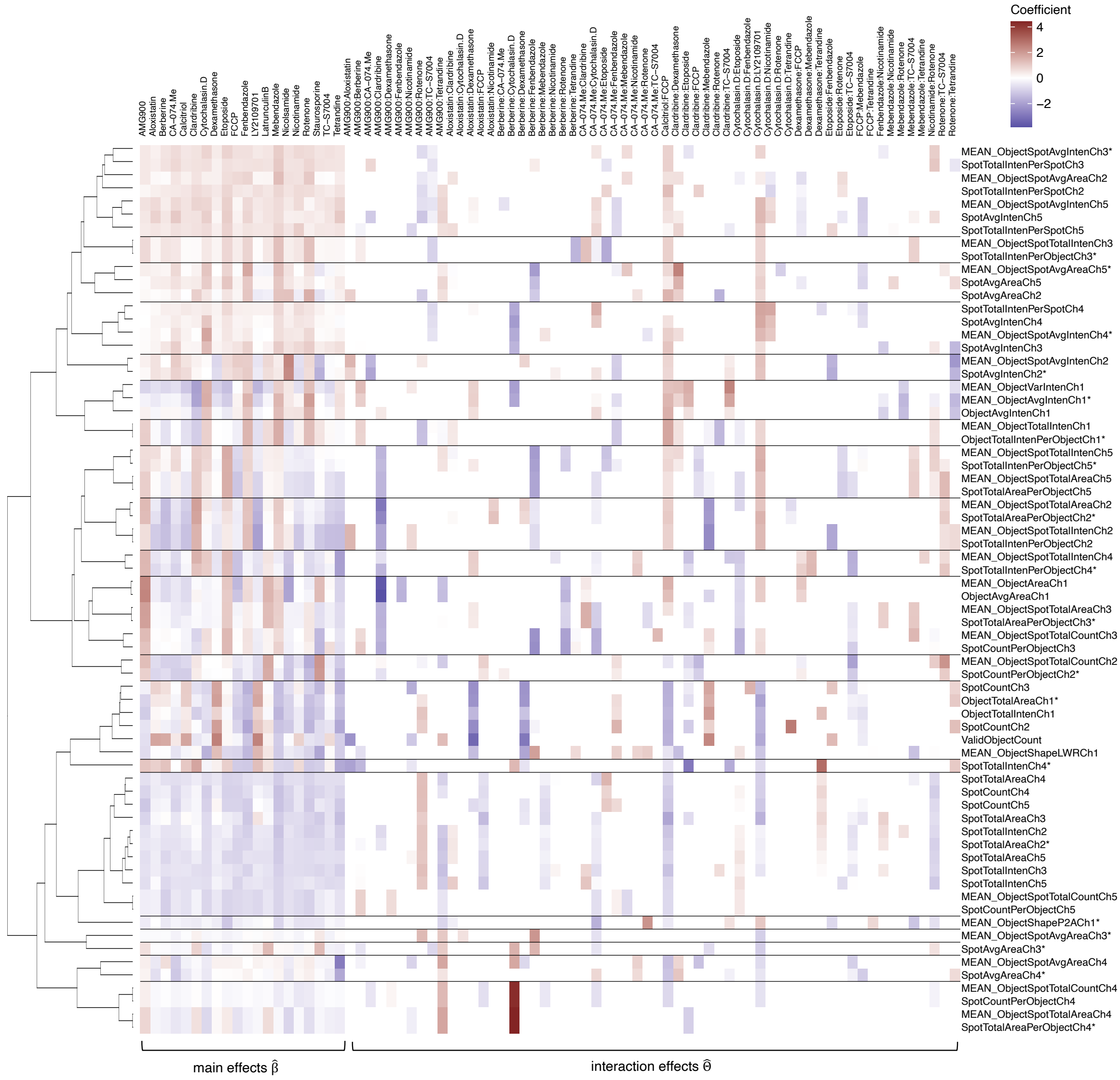


Fig 2. Clustered heatmap representation of estimated main effects $\hat{\beta}$ (left) and interaction effects $\hat{\Theta}$ (right). Each row represents a cell morphological feature Y_i, $i = 1, \ldots, p_2$, derived from the interaction model under a weak hierarchy with a robust Huber loss function. Only features (columns) with at least one non-zero effect are displayed. Prototypical features for each cluster are highlighted with an asterisk (*).

In Fig. 3, we visualize the results through scatterplots that contrast the individual effects, $\hat{\beta}_j$ and $\hat{\beta}_k$, with their corresponding interaction effects, $\hat{\Theta}_{jk}$. While labeling all dots is impracticable, this joint representation provides an overall impression, indicating that large main effects of the same sign tend to exhibit antagonistic interaction effects. Furthermore, certain prototypical features, obtained from the hierarchical clustering approach, and their corresponding interaction effects are labeled.

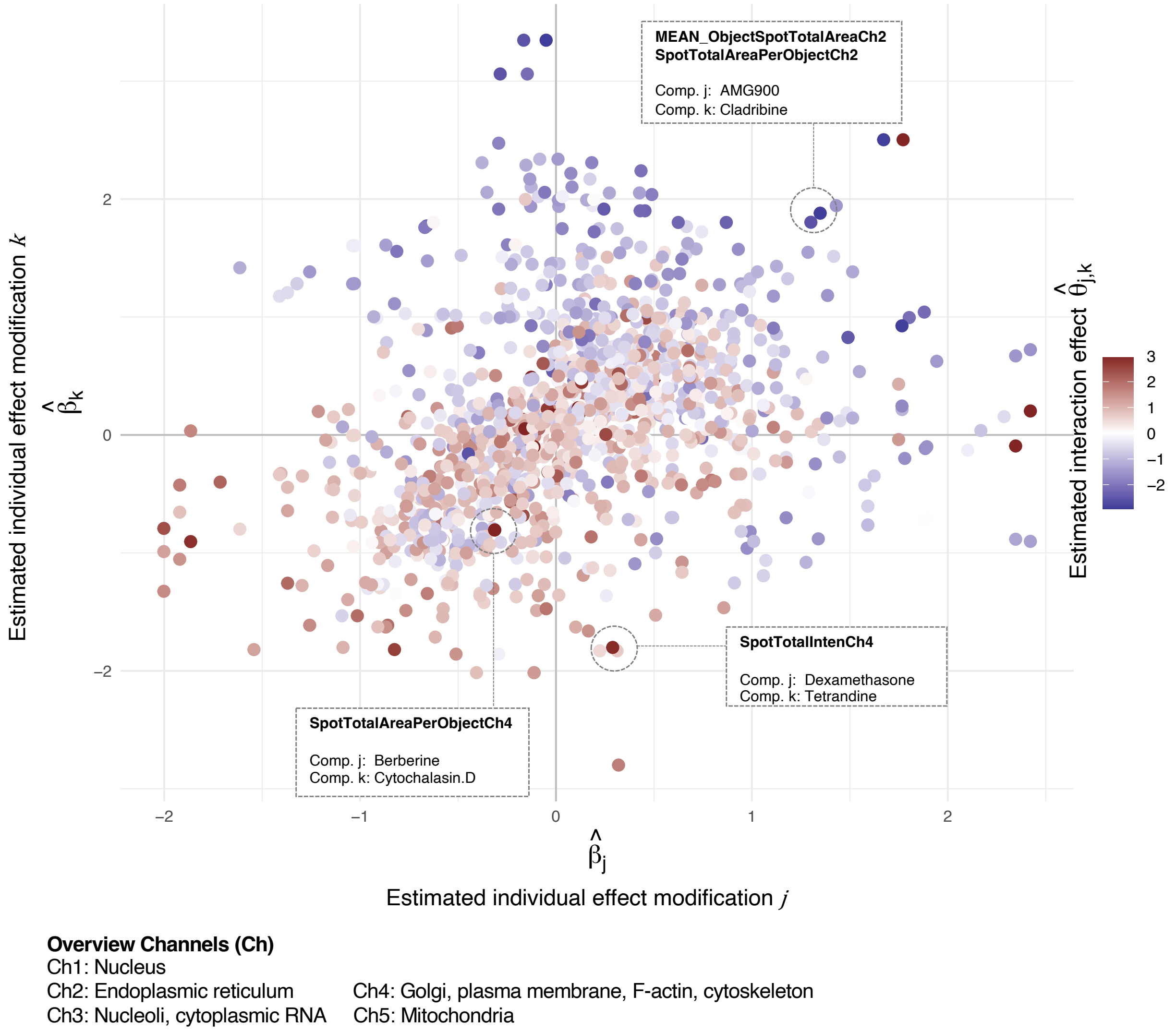


Fig 3. Scatterplot of drug effects on features describing cellular morphology with unspecific linear effects $\hat{\beta}_j$ and $\hat{\beta}_k$ on the x- and y-axes, and the corresponding interaction effect $\hat{\Theta}_{jk}$ represented by color. Some prototypical morphological features are labeled for an illustration.

References

[1] M. Bickle. “The beautiful cell: high-content screening in drug discovery”. In: *Analytical and bioanalytical chemistry* 398 (2010), pp. 219–226.

[2] K. A. Giuliano, J. R. Haskins, and D. L. Taylor. “Advances in high content screening for drug discovery”. In: *Assay and drug development technologies* 1.4 (2003), pp. 565–577.

[3] Y. E. Pearson, S. Kremb, G. L. Butterfoss, X. Xie, H. Fahs, and K. C. Gunsalus. “A statistical framework for high-content phenotypic profiling using cellular feature distributions”. In: *Communications Biology* 5.1 (2022), p. 1409.

[4] M. Stadler, S. Lukauskas, T. Bartke, and C. L. Mueller. “asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers”. In: *Nucleic Acids Res. (accepted)* (2024).

[5] A. Kümmel, P. Selzer, M. Beibel, H. Gubler, C. N. Parker, and D. Gabriel. “Comparison of multivariate data analysis strategies for high-content screening”. In: *Journal of biomolecular screening* 16.3 (2011), pp. 338–347.

[6] F. Reisen, X. Zhang, D. Gabriel, and P. Selzer. “Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery”. In: *Journal of biomolecular screening* 18.10 (2013), pp. 1284–1297.

[7] D. Siegismund, M. Fassler, S. Heyse, and S. Steigele. “Benchmarking feature selection methods for compressing image information in high-content screening”. In: *SLAS technology* 27.1 (2022), pp. 85–93.

[8] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[9] J. Bien, J. Taylor, and R. Tibshirani. “A lasso for hierarchical interactions”. In: *The Annals of Statistics* 41.3 (June 2013).

[10] J. Bien and R. Tibshirani. *hierNet: A Lasso for Hierarchical Interactions.* R package version 1.9. 2020.

[11] P. J. Huber. “Robust estimation of a location parameter”. In: *Breakthroughs in statistics: Methodology and distribution.* Springer, 1992, pp. 492–518.

[12] Y. Liu, P. Zeng, and L. Lin. “Degrees of freedom for regularized regression with Huber loss and linear constraints”. In: *Statistical Papers* 62.5 (2021), pp. 2383–2405.

[13] J. Lederer and C. Müller. “Don’t fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX”. In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 29. 1. 2015.

[14] Y. Wu and L. Wang. “A survey of tuning parameter selection for high-dimensional regression”. In: *Annual review of statistics and its application* 7 (2020), pp. 209–226.

[15] N. Hao, Y. Feng, and H. H. Zhang. “Model selection for high-dimensional quadratic regression via regularization”. In: *Journal of the American Statistical Association* 113.522 (2018), pp. 615–625.

[16] B. Yu. “Stability”. In: *Bernoulli* 19.4 (2013), pp. 1484–1500.

[17] N. Meinshausen and P. Bühlmann. “Stability Selection”. In: *Journal of the Royal Statistical Society, Series B* 72 (2010), pp. 417–473.

[18] R. D. Shah and R. J. Samworth. “Variable selection with error control: Another look at stability selection”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75.1 (2013), pp. 55–80.

[19] A. K”ummel, P. Selzer, M. Beibel, H. Gubler, C. N. Parker, and D. Gabriel. “Comparison of Multivariate Data Analysis Strategies for High-Content Screening”. In: *Journal of Biomolecular Screening* 16.3 (2011), pp. 338–347.

[20] D. Siegismund, M. Fassler, S. Heyse, and S. Steigele. “Benchmarking feature selection methods for compressing image information in high-content screening”. In: *SLAS Technology* 27.1 (2022), pp. 85–93.

[21] J. Bien and R. Tibshirani. “Hierarchical clustering with prototypes via minimax linkage”. In: *Journal of the American Statistical Association* 106.495 (2011), pp. 1075–1084.

B Contribution as co-author

B.1 Decoding chromatin states by proteomic profiling of nucleosome readers

Contributing article

Lukauskas, S., Tvardovskiy, A., Nguyen, N. V., **Stadler, M.**, Faull, P., Ravnsborg, T.,..., Bartke, T. (2024). Decoding chromatin states by proteomic profiling of nucleosome readers. *Nature, 1-9.* doi: https://doi.org/10.1038/s41586-024-07141-5

Replication code

The source code developed for this study for data processing and analyses (https://github.com/lukauskas/publications-lukauskas-2024-marcs) and for the interactive web interface (https://github.com/lukauskas/marcs) are available at GitHub.

Copyright information

Author contributions

S.L., A.T., N.V.N. and T.B. designed the study and planned experiments and analyses. A.T. and N.V.N. performed experiments. S.L. performed computational analyses and designed the MARCS online resource. M.S. performed computational analyses. P.F., T.K.B. and T.R. performed MS measurements. B.Ö.A. and S.D. prepared reagents and performed quality control. K.B. helped with tissue culture experiments. S.L., A.T., P.F., H.F., R.G.H.L., T.K.B., T.R., S.M.H., O.N.J., M.V. and A.P.S. analyzed MS data. S.L., N.V.N., A.T., R.S., C.L.M., P.A.D. and T.B. analyzed and interpreted data. P.A.D. and T.B. supervised experiments and data analysis. T.B. coordinated the study. S.L., N.V.N., A.T. and T.B. wrote the manuscript with input from all of the authors.

Article

Decoding chromatin states by proteomic profiling of nucleosome readers

https://doi.org/10.1038/s41586-024-07141-5

Received: 10 May 2021

Accepted: 31 January 2024

Published online: 6 March 2024

Open access

Check for updates

Saulius Lukauskas[1,2,3,17], Andrey Tvardovskiy[1,17], Nhuong V. Nguyen[2,4,17], Mara Stadler[1,5,6], Peter Faull[2,7,15], Tina Ravnsborg[8], Bihter Özdemir Aygenli[1], Scarlett Dornauer[1], Helen Flynn[7], Rik G. H. Lindeboom[9,10], Teresa K. Barth[11,16], Kevin Brockers[1], Stefanie M. Hauck[11], Michiel Vermeulen[9,10], Ambrosius P. Snijders[7], Christian L. Müller[5,6,12], Peter A. DiMaggio[3], Ole N. Jensen[8], Robert Schneider[1,13,14] & Till Bartke[1,2,4] ✉

DNA and histone modifications combine into characteristic patterns that demarcate functional regions of the genome[1,2]. While many 'readers' of individual modifications have been described[3–5], how chromatin states comprising composite modification signatures, histone variants and internucleosomal linker DNA are interpreted is a major open question. Here we use a multidimensional proteomics strategy to systematically examine the interaction of around 2,000 nuclear proteins with over 80 modified dinucleosomes representing promoter, enhancer and heterochromatin states. By deconvoluting complex nucleosome-binding profiles into networks of co-regulated proteins and distinct nucleosomal features driving protein recruitment or exclusion, we show comprehensively how chromatin states are decoded by chromatin readers. We find highly distinctive binding responses to different features, many factors that recognize multiple features, and that nucleosomal modifications and linker DNA operate largely independently in regulating protein binding to chromatin. Our online resource, the Modification Atlas of Regulation by Chromatin States (MARCS), provides in-depth analysis tools to engage with our results and advance the discovery of fundamental principles of genome regulation by chromatin states.

Almost all genetic material of eukaryotic cells is stored in the nucleus in the form of chromatin, a nucleoprotein complex comprising DNA, histones and other structural and regulatory factors. DNA and histones carry chemical modifications that have central roles in chromatin regulation by either directly affecting chromatin structure or by recruiting reader proteins that mediate downstream events through specialized binding domains[4,6]. Chromatin modifications rarely occur in isolation but exist in specific combinations on histones or nucleosomes, often also involving histone variants[7–12]. As these combinations are highly correlated and predictable[13,14], they form the basis for the definitions of 'chromatin states' that are used to annotate functional regions in the genome such as enhancers, promoters, gene bodies and heterochromatin[1,2].

Most chromatin regulators contain several modification-binding domains, indicating that recognizing multiple modifications is an integral function of many nuclear proteins[15]. However, although readers of individual modifications are often well understood[3–5], only few factors recognizing multiple modifications are known[16–24]. Thus, how complex combinatorial modification patterns underlying chromatin states are interpreted is largely unclear.

To obtain a comprehensive understanding of how chromatin readers decode different chromatin states, we have implemented a multidimensional mass spectrometry (MS)-based chromatin profiling strategy combining large-scale nucleosome affinity purification[25] and chromatin immunoprecipitation (ChIP)–MS approaches with computational methods for the integrative analysis of high volumes of proteomics and next-generation sequencing (NGS) data. We performed over 80 affinity purification experiments with semisynthetic dinucleosomes containing modification signatures and DNA linkers representing promoter, enhancer or heterochromatin states[1,10,26], and identified close to 2,000 nucleosome-interacting proteins, including transcription, replication, remodelling and DNA repair factors. Systematically quantifying their binding to the different modification states enabled the discovery of co-regulated proteins and complex chromatin modification read-outs driven by particular nucleosomal features, thereby revealing basic principles of how chromatin readers decode the chromatin landscape.

[1]Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany. [2]MRC Laboratory of Medical Sciences (LMS), London, UK. [3]Department of Chemical Engineering, Imperial College London, London, UK. [4]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, London, UK. [5]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [6]Department of Statistics, Ludwig Maximilian University Munich, Munich, Germany. [7]Proteomic Sciences Technology Platform, The Francis Crick Institute, London, UK. [8]VILLUM Center for Bioanalytical Sciences and Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark. [9]Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Oncode Institute, Radboud University Nijmegen, Nijmegen, The Netherlands. [10]The Netherlands Cancer Institute, Amsterdam, The Netherlands. [11]Metabolomics and Proteomics Core, Helmholtz Zentrum München, Munich, Germany. [12]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA. [13]Faculty of Biology, Ludwig Maximilian University Munich, Martinsried, Germany. [14]German Center for Diabetes Research (DZD), Neuherberg, Germany. [15]Present address: Northwestern Proteomics Core Facility, Northwestern University, Chicago, IL, USA. [16]Present address: Clinical Protein Analysis Unit (ClinZfP), Biomedical Center (BMC), Faculty of Medicine, Ludwig Maximilian University Munich, Martinsried, Germany. [17]These authors contributed equally: Saulius Lukauskas, Andrey Tvardovskiy, Nhuong V. Nguyen. ✉e-mail: till.bartke@helmholtz-munich.de

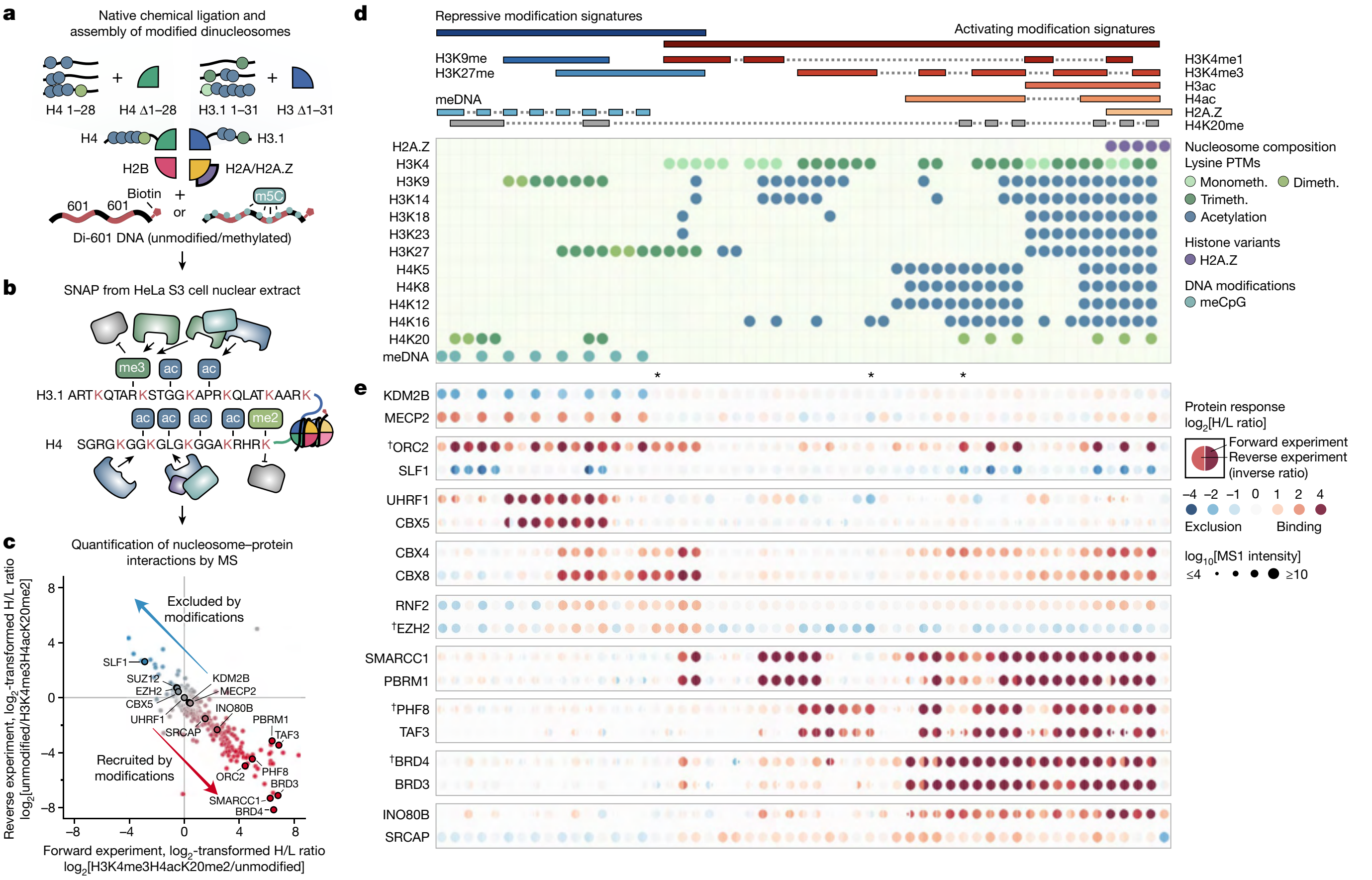


Fig. 1 | Large-scale identification of chromatin readers by SILAC dinucleosome affinity purifications. a, Generation of modified dinucleosomes. Modified histones H3.1 and H4 were prepared by native chemical ligations of N-terminal tail peptides (H3, amino acids 1–31; H4, amino acids 1–28) to truncated histone cores (H3.1Δ1–31T32C or H4Δ1–28I29C, respectively). Note that this introduces H3T32C and H4I29C mutations that might affect protein binding to nearby modifications. Ligated histones were refolded into octamers and assembled into dinucleosomes using a biotinylated DNA containing two nucleosome-positioning sequences (di-601)[47]. For some experiments, CpG-methylated DNA (m5C) or H2A.Z were used. **b**, SNAP purifications. Modified nucleosomes were immobilized on streptavidin beads and incubated with nuclear extracts from HeLa S3 cells grown in isotopically light (R_0K_0) or heavy ($R_{10}K_8$) SILAC medium. **c**, Protein responses to modified nucleosomes. For each SNAP experiment, bound proteins were identified and quantified using MS, and the forward (*x* axis) and reverse (*y* axis) SILAC ratios (H/L ratio) were plotted on a logarithmic ($\log_2$) graph. **d**, A library of modified dinucleosomes. A header specifies the modification status of each nucleosome. Nucleosomes are arranged in columns, with the respective modifications displayed in rows. Modifications of specific lysine residues in histone H3 and H4 and the presence of DNA methylation (meCpG) or H2A.Z are colour coded as indicated. Nucleosomes are ordered to imitate clustering by increasingly active chromatin states. Monometh., monomethylation; PTMs, post-translational modifications. **e**, Visualization of protein binding responses to the 55 modified dinucleosomes profiled by SNAP. The $\log_2$[H/L] ratios for each protein in each SNAP experiment are shown as circles, with the right half representing the forward and the left half the reverse $\log_2$[H/L ratio]. Recruitment (red) and exclusion (blue) are indicated. The reverse H/L ratio was inverted to display both ratios on the same scale. Circle sizes denote the total MS1 peak intensities on a $\log_{10}$ scale. The asterisks indicate experiments that are shown in Extended Data Fig. 1b–d. The dagger symbols (†) indicate proteins that are highlighted in Extended Data Fig. 1b–e.

To make our data easily accessible, we have developed computational tools to analyse and visualize the nucleosome-binding data and we have implemented them in the interactive online resource MARCS (https://marcs.helmholtz-munich.de/). Our results bridge the gap between chromatin states and chromatin readers, and we anticipate that MARCS will become a valuable resource to drive future chromatin research forward as numerous other observations emerge.

Proteomic profiling of chromatin readers

To systematically profile the interactomes of chromatin modifications in the nucleosomal context, we performed SILAC nucleosome affinity purification (SNAP)[25]. We assembled nucleosomes from biotinylated DNA and histone octamers containing site-specifically modified histones H3.1 and H4 prepared by native chemical ligation[27] (Fig. 1a) and used them in forward and reverse SILAC nucleosome pull-down experiments in HeLa S3 cell nuclear extracts (Fig. 1b and Extended Data Fig. 1a). The label swap enables unbiased identification of proteins that are reproducibly either recruited or excluded by the modification(s). Moreover, the SILAC heavy/light (H/L) ratios also indicate a relative strength of recruitment or exclusion of a protein by the modifications (Fig. 1c). After optimizing our SNAP methodology (Supplementary Information) for a large-scale comparison of interactomes of different chromatin states, we used single-end biotinylated dinucleosomes in all SNAP experiments.

To understand how distinct chromatin states marked by combinations of modifications are read by binding proteins, we created a library of nucleosomes incorporating biologically relevant modification signatures, including mono- and tri-methylation of lysine 4 of histone H3 (H3K4me1/3), di- and tri-methylation of lysines 9 and 27 of histone H3 (H3K9me2/3 and H3K27me2/3), di- and tri-methylation of lysine 20 of histone H4 (H4K20me2/3), varying degrees of acetylation of lysines (Kac), the histone variant H2A.Z or CpG-methylated DNA. This design of the nucleosome library enabled us to capture the interactomes of major repressive and activating chromatin states (Fig. 1d), including enhancer, promoter and different heterochromatin states. A detailed list of modified histones, octamers and nucleosomes and corresponding quality controls is provided in the Supplementary Information.

In total we performed SILAC-linked affinity purifications with 55 dinucleosomes. The forward and reverse experiments were generally very reproducible, and we achieved high detection coverage for most of the identified proteins. After correction for batch effects and imputation of missing values (Supplementary Information), we catalogued the responses of 1,915 proteins to the various modification states (Supplementary Table 1), covering a large part of the known chromatin proteome. Collectively, the SNAP experiments not only characterize protein binding to the nucleosomal modifications but also offer systematic insights into the behaviour of chromatin readers through analysis of the changes in the H/L ratios across the entire dataset.

MARCS maps chromatin-binding responses

Comparing the $\log_2$-transformed H/L ratios of individual proteins across SNAP experiments revealed characteristic nucleosome-binding behaviours (Extended Data Fig. 1b–d). To facilitate the analysis and exploration of many SNAP experiments (Extended Data Fig. 1e), we implemented the interactive online visualization resource MARCS (https://marcs.helmholtz-munich.de).

Figure 1d,e shows an exemplary set of heat maps generated using MARCS. The clustered heat map of all proteins is provided in Supplementary Table 2. Our data capture a broad range of responses by chromatin readers to repressive and activating modification states and thereby reveal two principle modes of interaction: simple responses to single modifications as exemplified by the recruitment of MECP2 or exclusion of KDM2B by DNA methylation (Fig. 1e); and complex binding patterns indicating binding to multiple modifications or synergistic responses as illustrated by the origin recognition complex (ORC) that shows recruitment to H3K9, H3K27 or H4K20 methylations, with further stimulation by DNA methylation (ORC2 in Fig. 1e). Importantly, while these examples constitute internal controls by consistently showing known and expected binding behaviours, our broad and unbiased profiling of chromatin states also enables the identification of interactions with modified nucleosomes in new contexts. For example, we find that the INO80 chromatin remodelling complex[28] and polycomb repressive complex 1 (PRC1)[29] are enriched on nucleosomes displaying active modification signatures, including acetylations of the histone H3 and H4 N-terminal tails (INO80B for INO80 in Fig. 1e; CBX4 and CBX8 for PRC1 in Fig. 1e and Extended Data Fig. 2a,b).

Unbiased prediction of binding features

Inspection of the heat maps further revealed that many proteins exhibit broad nucleosome binding responses that cannot be explained by one single feature, that is, a particular histone modification, DNA methylation or the H2A.Z variant alone. To describe such complex binding behaviours, we deconvoluted the SNAP binding profiles into individual nucleosomal features driving these associations. We achieved this by comparing $\log_2$[H/L ratio] values between related nucleosomes that differ by only one single feature. For example, four pairs of dinucleosomes are informative of the effect of H3K4me3 on protein binding (Fig. 2a). A consistent increase or decrease in the $\log_2$[H/L ratio] across these nucleosome pairs can be attributed only to H3K4me3, irrespective of other modifications that the chromatin reader may recognize. Repeatedly sampling this effect across multiple nucleosome pairs, in addition to the H3K4me3 dinucleosome-purification experiment itself (Extended Data Fig. 3a), enables statistical evaluation and calculation of a ‘feature effect estimate’ expressed as the H3K4me3-dependent change in the $\log_2$[H/L ratio] for a particular protein (Fig. 2b). This way, we were able to resolve the responses of chromatin readers to 15 different modification features resulting from 82 pairs of nucleosomes (Fig. 2b, Extended Data Figs. 3b–d and 5a and Supplementary Table 3). The feature effect estimates enable us to quantitatively describe the chromatin-binding behaviours of several hundred proteins and provide a breakdown of complex binding profiles into a set of key features that either positively or negatively regulate their association with the modified nucleosomes (Extended Data Fig. 2c,d). We have implemented this decomposition of binding profiles into ‘chromatin feature motifs’ in the MARCS online resource. Importantly, an integrative analysis of public ENCODE[30] ChIP followed by sequencing (ChIP–seq) datasets covering a subset of identified nucleosome-interacting proteins and relevant chromatin features demonstrates that the binding behaviours observed in our in vitro dinucleosome system recapitulate the binding behaviours found in cellular chromatin (Extended Data Fig. 4a–j and Supplementary Table 4).

Notably, the number of proteins responding to each of the 15 features is highly variable, with euchromatic features such as H3ac or H4ac recruiting or excluding many more proteins than heterochromatic ones such as H3K9me2/3 or H3K27me2/3 (Fig. 2c). However, this might be biased by the extract preparation method, which preferentially releases euchromatic proteins. Furthermore, many proteins are regulated by more than one feature (Fig. 2d,e) indicating that they either respond to multiple modifications independently or recognize composite modification signatures. Clustering of individual protein binding behaviours revealed that they can be grouped into 40 major binding responses, largely defined by multisubunit protein complexes (Fig. 2e and Supplementary Table 5). For example, multiple factors such as the INO80, MLL3/4, NuA4 or TFIID complexes show highly specific responses to the different ‘promoter state’ features H3K4me3, H3ac, H4ac and H2A.Z. Whereas binding of, for example, the INO80 remodeller[28] is stimulated by H2A.Z in addition to H3 and H4 acetylation (Extended Data Fig. 5a–c), the NuA4 histone acetyltransferase complex responds similarly to H3 and H4 acetylation, but not H2A.Z (Fig. 2e). This complex regulation of INO80 by a H3ac/H4ac–H2A.Z axis was not directly apparent from the original SNAP data (Extended Data Fig. 5d), illustrating how the feature effect estimates can be used to decode nucleosome-binding determinants across entire chromatin states.

Absence of distinctive H3K4me1 readers

Another notable result from the feature effect analysis was the differential binding of proteins to H3K4 methylations (Fig. 3a). For the promoter mark H3K4me3, we identified 45 strongly recruited proteins (positive effect to $\log_2$[H/L ratio] ≥ 1 at a false-discovery rate (FDR) of 1%), including known H3K4me3 readers such as TFIID[31] and PHF8[32], and 31 strongly excluded proteins (Fig. 2b and Supplementary Table 3), such as polycomb repressive complex 2 (PRC2)[33]. By contrast, the enhancer mark H3K4me1 enriched only one protein, BRPF3 (Extended Data Fig. 3c). Consistent with these findings, our integrative ChIP–seq data analysis revealed no proteins showing strong association with H3K4me1, while many proteins preferentially localized to H3K4me3-marked genomic loci (Extended Data Fig. 4c,d). This was further supported by a label-free quantitative ChIP–MS analysis of H3K4me1- and H3K4me3-enriched mononucleosomes (Extended Data Fig. 6a–c). Although many proteins were significantly enriched in both H3K4me1 and H3K4me3 ChIPs compared with bulk nucleosome purifications, the vast majority of these proteins preferentially associated with H3K4me3- but not H3K4me1-modified chromatin (Extended Data Fig. 6d–h and Supplementary Table 6). This suggests the absence of a distinctive H3K4me1 interactome, supporting the notion that H3K4me1 is not a main driver of protein recruitment to enhancer chromatin states.

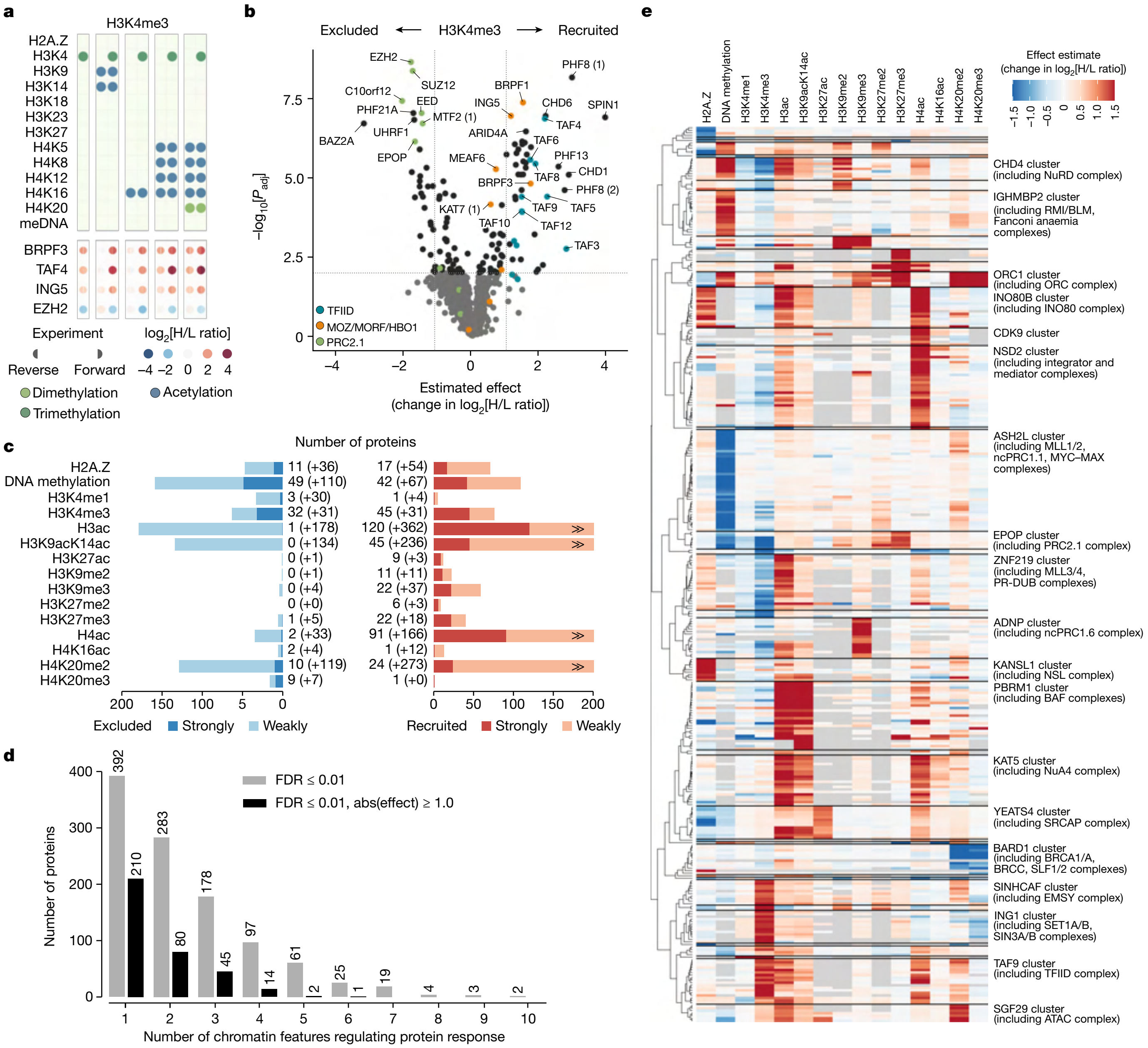


Fig. 2 | Feature effect estimates reveal binding responses of chromatin readers to different nucleosomal features. a, Nucleosomes informative of protein responses to H3K4me3. The four pairs of dinucleosomes that differ only by H3K4me3, alongside the self-informative H3K4me3 dinucleosome (top), and the binding responses of four representative proteins in the corresponding SNAP experiments (bottom) are shown. **b**, Feature effect estimates of proteins showing H3K4me3-dependent nucleosome binding. The change in the $\log_2$[H/L ratio] attributable to H3K4me3 (x axis) is plotted against the P value (limma, two-sided, Benjamini–Hochberg adjusted) on a $-\log_{10}$ scale (y axis). The vertical lines highlight an effect to fold change of 1, and the horizontal line signifies the FDR threshold of 0.01. Selected protein complexes are highlighted. Duplicate protein identifiers, for example, PHF8 (1), mark distinct UniProt IDs with the same gene name (Trembl versus SwissProt versions); for annotations, see Supplementary Table 1. **c**, The number of interactors responsive to different chromatin features. Owing to their frequent co-occurrence, blocks of acetylation, such as H3K9acK14ac, H3K9acK14acK18acK23acK27ac (H3ac) and H4K5acK8acK12acK16ac (H4ac) were treated as single features. Proteins with statistically significant (limma, FDR ≤ 0.01) effect estimates ≥ 1 classify as strongly recruited, or strongly excluded if their estimate is ≤ −1. Changes in $\log_2$[H/L ratio] < 1 are considered to be weakly recruited or excluded. **d**, The number of chromatin features regulating protein binding responses. The grey bars tally the number of proteins with statistically significant feature effects (limma, FDR ≤ 0.01). The black bars additionally tally proteins with strong feature effects (absolute effect ≥ 1). **e**, Clustered heat map of feature effect estimates of proteins strongly responding to at least one feature as shown in **c**. Individual estimates are colour coded. Entries without an estimate due to insufficient data are marked in grey. Prototype proteins representing the binding response of each cluster are shown on the right. Notable protein complexes are highlighted.

MARCS recovers protein interaction networks

Closer analysis of binding profiles of protein complexes indicated that their subunits showed highly similar binding behaviours (for example, the H2A.Z-responsive INO80, SRCAP and NSL complexes; Extended Data Fig. 5d), underscoring that their native compositions remained intact during the affinity purifications. This prompted us to reconstruct a network of proteins co-regulated by similar chromatin states and use this to predict protein–protein interactions. To this end, we trained and tested several network inference algorithms (Extended

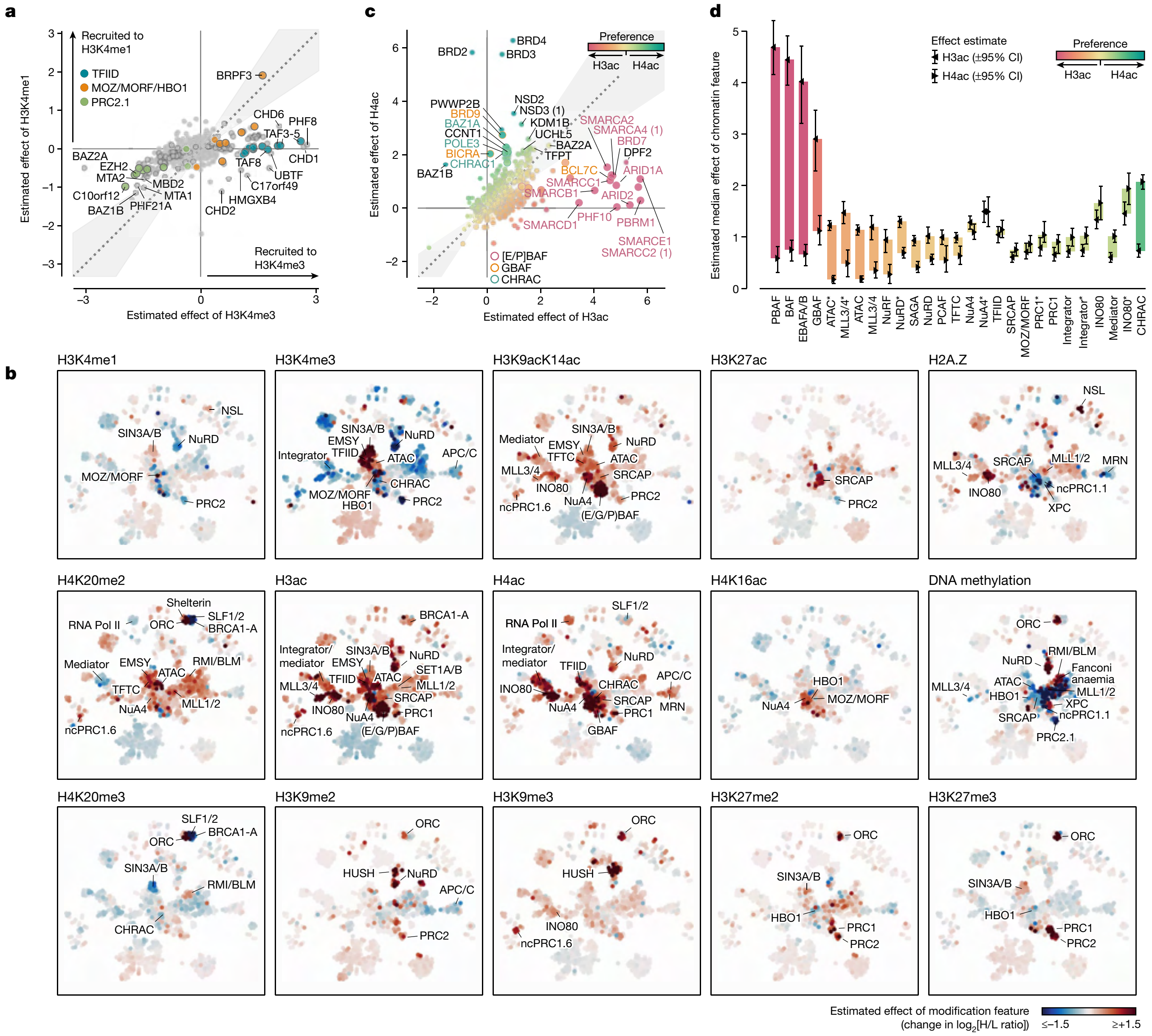


Fig. 3 | Differential binding of proteins to H3K4 methylation and H3/H4 acetylation states. a, Comparison of H3K4me3- versus H3K4me1-responsive proteins. H3K4me3- or H3K4me1-dependent changes in the $\log_2$[H/L ratio] are plotted on the *x* and *y* axes, respectively. Proteins with statistically significant estimates (limma, two-sided, Benjamini–Hochberg-adjusted FDR ≤ 0.01) are circled with a grey border. The grey area marks ±0.2 radians away from the $x = y$ line. Selected protein complexes are highlighted. While H3K4me1 recruits only BRPF3 but no other interactors, it still excludes, for example, the PRC2 complex, albeit not as strongly as H3K4me3. **b**, CLR-predicted network overlayed with chromatin feature effects. The heat maps reveal the degree and specificity of protein recruitment or exclusion by the different features. Protein complexes with statistically significant regulation (CAMERA, FDR ≤ 0.01, median effect ≥ 0.3; Supplementary Table 8) were annotated for each feature after manual curation. A zoomable version is provided in the MARCS resource. **c**, Comparison of proteins responding to H3 versus H4 acetylation. Changes in the $\log_2$[H/L ratio] attributable to H3ac or H4ac are plotted on the *x* and *y* axes, respectively. Data representation as in **a**. Proteins are coloured by the difference between their H3ac and H4ac responses. BAF and CHRAC complex subunits are highlighted with coloured borders and labels. **d**, The preference of protein complexes for H3 or H4 acetylation. Markers indicate the median effect of the H3ac versus the H4ac feature across all complex subunits with protein response measurements (the number of measurements per complex/feature is shown in Supplementary Fig. 1). The error bars represent the empirical 95% confidence interval (CI) of this median effect estimated from 100,000 random samples of subunit effects, accounting for their variance. The coloured bars highlight the difference between these median estimates for H3ac and H4ac. Complexes are ordered from H3ac to H4ac preference. The asterisks denote estimates for exclusive complex subunits.

Data Fig. 7a) against BioGRID[34]. In this analysis, the context-likelihood of relatedness (CLR) algorithm[35,36] performed best based on the highest area under the precision-recall curve (Extended Data Fig. 7b). CLR also scored interactions reported by multiple publications and validated by co-crystal structures and co-purifications highest (Extended Data Fig. 7c,d), confirming the reliability of the predicted network.

Within the resulting network (Supplementary Table 7), key chromatin regulatory complexes formed clusters (Extended Data Fig. 7e) that, at increased stringencies, resolved into separate complexes and high-confidence binary interactions (Extended Data Fig. 8). Importantly, the normalized mutual information (MI) estimates between pairs of proteins in our integrative ChIP–seq analysis increased in line with increasing confidence of the predicted interactions (Extended Data

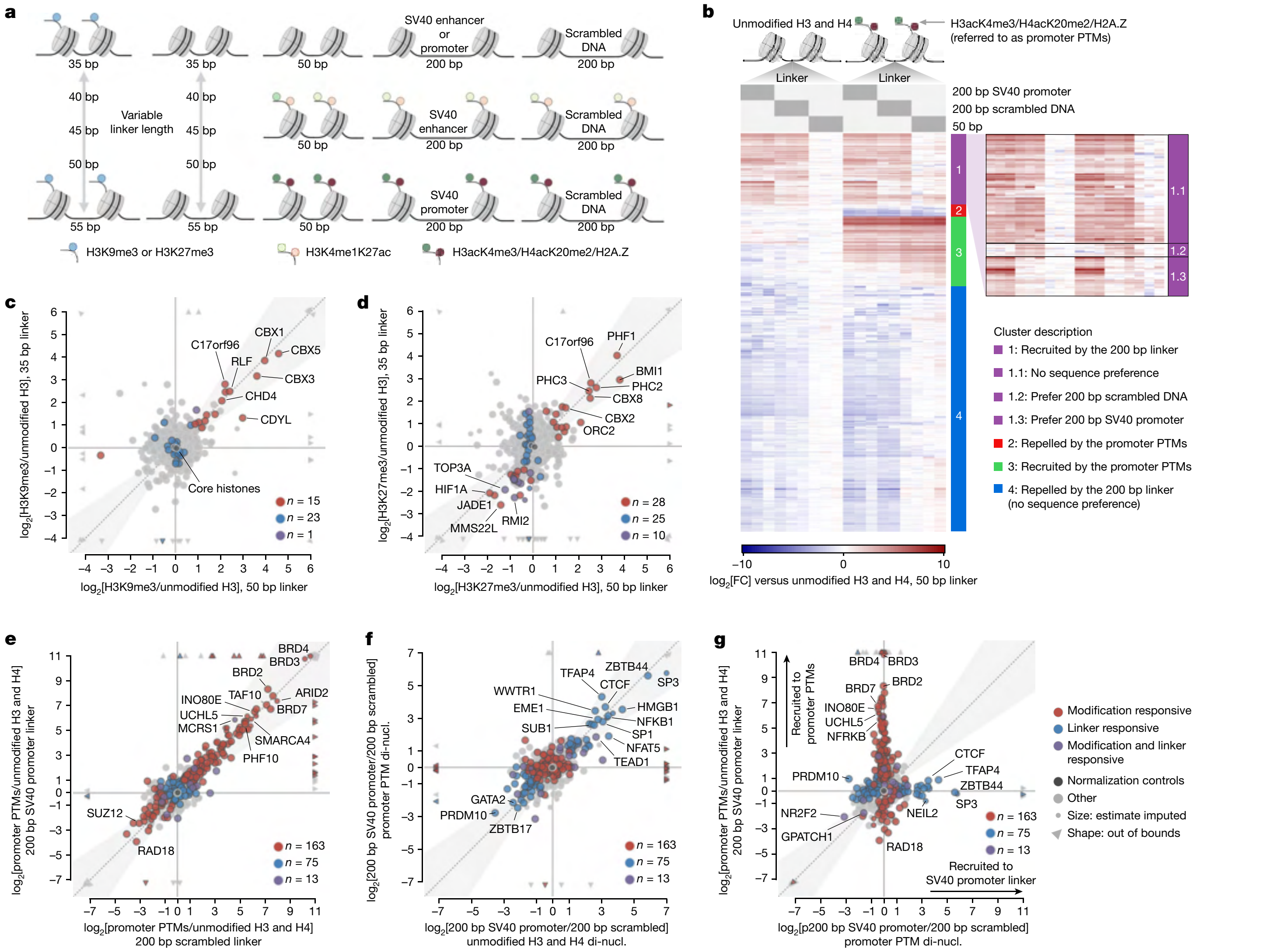


Fig. 4 | Nucleosomal modifications and linker DNA constitute orthogonal routes of protein engagement with chromatin. a, Schematic of dinucleosomes used in label-free MS-based pull-downs for evaluating the effect of linker DNA length and sequence on protein binding to active (right) and repressive (left) chromatin states. **b**, Clustered heat map depicting protein binding responses to dinucleosomes incorporating different combinations of 200 bp scrambled DNA or SV40 promoter sequence-based linkers and promoter PTMs (H3K4me3K9acK14acK18acK23acK27ac in combination with H4K5acK8acK12acK16acK20me2 and H2A.Z). Data are shown as the $\log_2$-transformed fold change ($\log_2$[FC]) in the normalized protein abundances compared with unmodified dinucleosomes with a 50 bp linker. **c**, Comparison of H3K9me3-binding responses on dinucleosomes with 35 bp and 50 bp linkers. Proteins responding to H3K9me3, linker length or both were determined using limma statistics and are highlighted in red, blue or purple, respectively. Only binding responses fulfilling the following two criteria are depicted: (1) $\log_2$[FC] > 1 or $\log_2$[FC] < −1 compared with unmodified dinucleosomes with 50 bp linker; (2) Benjamini–Hochberg-adjusted $P \leq 0.05$. The $x = y$ line indicates where binding responses to H3K9me3 dinucleosomes incorporating 35 bp and 50 bp linkers are identical. The grey area marks ±0.2 radians away from the $x = y$ line. Core histones (normalization controls) are indicated in dark grey. The smaller datapoints indicate response estimates based on single data points. The triangles indicate points outside the data axes. **d**, Comparison of H3K27me3-binding responses on dinucleosomes with 35 bp and 50 bp linkers. Data representation in **d**–**g** is as described in **c**. **e**, Comparison of protein binding responses to promoter PTMs on dinucleosomes with 200 bp scrambled DNA and SV40-promoter-sequence-based linkers. **f**, Comparison of sequence-specific protein binding responses to the SV40 promoter linker in unmodified dinucleosomes (di-nucl.) and dinucleosomes decorated with promoter PTMs. **g**, Comparison of protein binding responses to SV40 promoter linker and promoter PTMs.

Fig. 7f), indicating that the CLR-predicted network correctly enriches in vivo chromatin interactions. We leverage the identified local protein interactions to implement similarity predictions in the MARCS resource and augment these with a curated list of protein complexes (Supplementary Table 8), incorporating information from other resources such as EpiFactors[37] and the Complex Portal[38].

The CLR algorithm, being based on MI, treats mutually exclusive interactions similarly to correlated ones. Overlaying the chromatin feature effect estimates for each protein onto the network reveals how their arrangement into tight subnetworks is driven by the chromatin modification responses (Fig. 3b). Among other regulations, these data reveal differential binding of many factors to H3 and H4 acetylations, as different subnetworks show distinct binding responses to H3K27ac, H4K16ac, and the combined H3K9acK14ac, H3ac and H4ac features, suggesting a finely orchestrated regulation of active chromatin states by differential acetylation. Whereas, for example, the CHRAC chromatin remodelling complex shows preferential binding to H4ac, BAF (SWI/SNF) remodellers show a strong preference for H3ac (Fig. 3c,d), mainly driven by H3K9acK14ac (Fig. 3b). Furthermore, while many proteins respond to multiple acetylations in

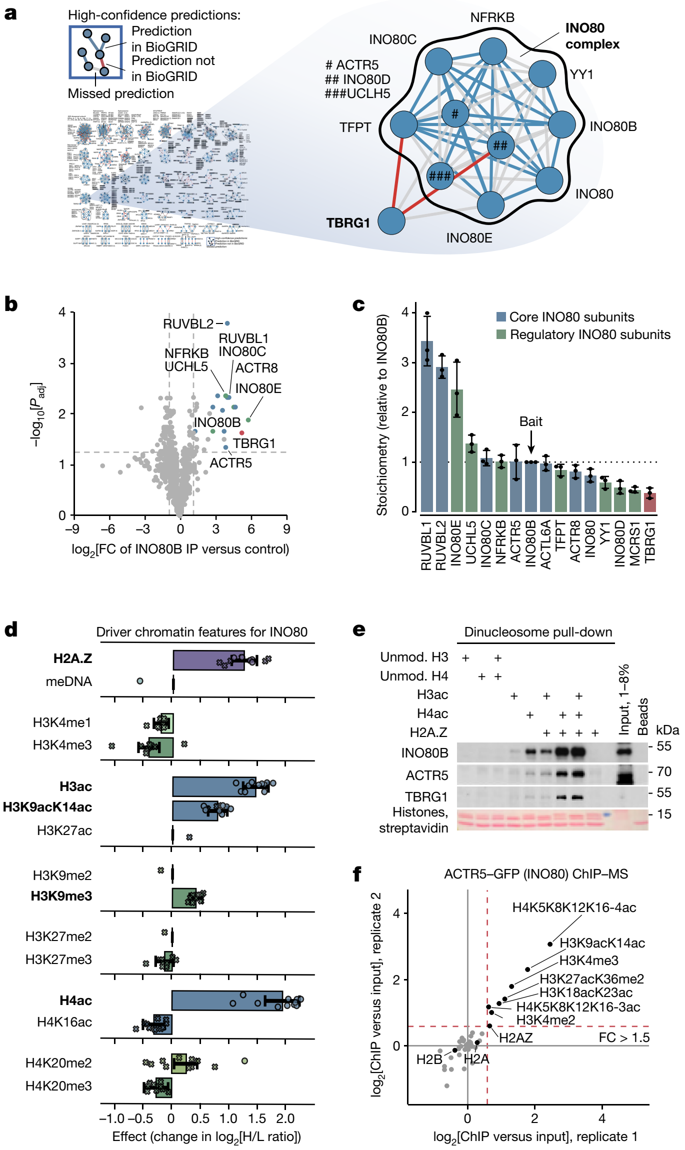


Fig. 5 | The INO80 complex recognizes a multivalent nucleosome-modification signature. a, CLR-predicted TBRG1–INO80 interaction. TBRG1–INO80 interactions were reported in several screens[48–50] and deposited at BioGRID but never validated. **b**, TBRG1 interacts with INO80. Volcano plot of proteins that are significantly enriched (*t*-test, two-sided, Benjamini–Hochberg-adjusted FDR ≤ 0.05) in $n = 3$ biologically independent INO80B-V5 immunoprecipitations (Extended Data Fig. 5h) followed by label-free MS. **c**, Composition of the INO80 complex. The relative stoichiometries between TBRG1 and INO80 were calculated using quantitative MS data from the INO80B-V5 immunoprecipitation experiments shown in **b**. $n = 3$. Data are the mean ± s.d. of the stoichiometry values. **d**, Features driving the INO80 nucleosome-binding response. Individual effect estimates (change in log$_2$[H/L ratio]) for INO80-exclusive subunits are shown as dots (estimate significantly non-zero, limma, two-sided, Benjamini–Hochberg-adjusted FDR ≤ 0.01) or crosses (estimate not statistically significant). The bars highlight the median effect across all complex subunits with protein response measurements ($n = 11$, except for DNA methylation, H3K27ac, H3K9me2 and H3K27me2, for which $n = 1$ and no estimate was derived). The error bars represent the empirical 95% CI of this median effect estimated from 100,000 random samples of subunit effects, accounting for their variance. The bold font indicates features with enrichments greater than expected by chance (CAMERA, Benjamini–Hochberg-adjusted FDR ≤ 0.01; Supplementary Table 8). **e**, Targeted dinucleosome pull-downs confirm INO80 binding to nucleosomes containing hyperacetylated H3 (H3ac), H4 (H4ac) and/or H2A.Z. Binding was detected by immunoblotting against INO80B and ACTR5. TBRG1 follows the INO80-binding pattern. The HeLa S3 cell nuclear extract used was a mixture of three independent preparations. Different amounts of the mixed extract were loaded as inputs for the different immunoblots. Experiments were independently repeated three times with similar results. Unmod., unmodified. **f**, Quantitative label-free LC–MS-based analysis of histone modifications and H2A.Z in mononucleosomes co-purified with ACTR5 from MNase-digested HeLa cell chromatin. The relative PTM or H2A.Z abundance over input chromatin is plotted as the log$_2$[FC] for $n = 2$ independent biological experiments.

the H3 and H4 tails, only few factors respond to H3K27ac or H4K16ac alone (Fig. 3b). This breakdown of the SNAP data into local interaction networks of co-regulated proteins and their responses to specific chromatin features provides important insights into how chromatin states are decoded by chromatin readers.

Modifications and linkers act independently

Apart from covalent modifications, characteristic features of chromatin states also include linker DNA length, typically ranging from 35–55 bp in most chromatin domains[39] to over 200 bp in nucleosome-depleted regions (NDRs). To investigate the effects of linker DNA on chromatin recognition by nuclear proteins, we performed an additional set of affinity purifications using dinucleosomes incorporating different DNA linkers (Fig. 4a and Supplementary Information). Notably, the binding of heterochromatin as well as active promoter modification readers was generally not affected by variations in linker length nor linker sequence (Fig. 4b–e, Extended Data Fig. 9a–g and Supplementary Table 9), highlighting the robustness of the protein binding responses captured in MARCS. Likewise, the binding of sequence-specific transcription factors recognizing DNA motifs in the 200 bp long SV40 promoter linker was insensitive to the active promoter modifications on the adjacent nucleosomes (Fig. 4f,g and Extended Data Fig. 9d,g). Similarly, incorporating a 200 bp long SV40 enhancer linker had no prominent effect on H3K4me1 and H3K4me1K27ac enhancer state readout (Extended Data Fig. 10a–c and Supplementary Table 9), and transcription factor recognition of the SV40 enhancer sequence was not affected by the H3 modifications (Extended Data Fig. 10d,e). Nucleosomal modifications and DNA linkers therefore appear to act largely independently in recruiting proteins to chromatin. Notably, many proteins, including multiple spliceosome subunits, showed diminished binding when increasing the linker length from 50 to 200 bp, regardless of the linker sequence or modification status of the adjacent nucleosomes (Fig. 4b and Extended Data Figs. 9l,m,o and 10a,f–h), underscoring the regulatory potential of nucleosome spacing on chromatin engagement irrespective of the underlying modification landscape.

Multivalent chromatin engagement by INO80

Our combined analyses can be used to identify chromatin binding behaviours and nuclear regulators with unknown functions. As a proof of principle, we selected INO80, an ATP-dependent nucleosome remodeller and exchange factor for the histone variant H2A.Z that is involved in transcription, replication and DNA repair[28], for which several interesting observations emerged from our data (Extended Data Fig. 5d). First, our high-confidence CLR network predicted an interaction with transforming growth factor beta regulator 1 (TBRG1), a putative tumour suppressor and p53 activator[40] (Fig. 5a and Extended Data Fig. 8). Consequently, we were able to co-purify TBRG1 together with INO80 in co-immunoprecipitation (co-IP) experiments from INO80B-V5 knock-in cell lines (Fig. 5b and Extended Data Fig. 5e–h). Label-free MS-based estimation of the TBRG1:INO80B ratio indicated that TBRG1 is present in the complex at substoichiometric levels comparable to the regulatory subunits MCRS1, INO80D and YY1 (Fig. 5c).

Second, while the INO80 complex was unresponsive to variations in the linker DNA (Fig. 4e–g and Extended Data Fig. 9c,d,f,g), our feature effect estimates predicted binding to a multivalent nucleosomal modification signature consisting of acetylations in the H3 and H4 N-terminal tails and the histone variant H2A.Z (Fig. 5d and Extended Data Fig. 5b,c). Confirming our prediction, we found in targeted pull-downs (Fig. 5e) that H3ac had a small positive effect on INO80 recruitment, which was more pronounced in the case of H4ac. Notably, while no effect of H2A.Z alone was detectable by western blotting, the presence of H2A.Z greatly enhanced INO80 binding when combined with H4ac, and to a lesser extent with H3ac (Fig. 5e). Consistent with the in vitro results, mononucleosomes co-purified with INO80 from micrococcal nuclease (MNase)-digested HeLa chromatin through the subunit ACTR5 were enriched in H4ac and H3ac as well as H2A.Z (Fig. 5f and Extended Data Fig. 5i–k). These results confirm that the INO80 remodelling complex indeed binds to nucleosomes decorated by the predicted multivalent chromatin modification signature in human cells and suggest a role of histone acetylation and H2A.Z in stimulating INO80 recruitment to specific genomic loci (Extended Data Fig. 5l).

These independent experimental validations highlight the reliability of our analyses and predictions, and underscore the value of our data to identify previously undescribed protein interactions and complex binding events involving the concerted interplay between multiple chromatin modification features.

Discussion

Here we have combined large-scale quantitative nucleosome affinity purification approaches and computational analysis methods to understand how chromatin states are read and interpreted by nuclear machineries. Our approach has enabled us to delineate direct effects of composite modification signatures of promoter, enhancer and heterochromatin states on chromatin engagement by several hundred chromatin readers and to uncover interconnected networks of nuclear proteins targeting similar chromatin states. Deconvoluting the responses of chromatin factors to 15 different modification features unravels how complex modification signatures are sensed by chromatin-binding proteins. Combining these responses to individual modification features into modification response profiles, akin to DNA-binding-motif logos of transcription factors[41], enables the comprehensive prediction of chromatin regulators that recognize complex modification patterns. Similarly, it enables the systematic identification of nucleosomal features modulating the binding of various nuclear proteins to their genomic target loci. Predicted responses to multiple features point towards a synergistic interplay between the components, as we show for the INO80 remodeller (Fig. 5e,f).

While an interplay between distinct nucleosomal modifications is clearly visible for many proteins, it generally seems not to involve linker DNA as we observe no apparent synergy even between active modifications and NDRs often coupled in vivo. However, this might reflect the static nature of the interactions in our pull-downs, in which the absence of ATP and the presence of HDAC inhibitors prevent enzymatic activities that are known to be involved in highly dynamic regulatory circuits, such as nucleosome remodelling and rapid histone acetylation turnover. In the case of multistep enzymatic processes, such as chromatin remodelling by INO80, the reported interactions might therefore reflect particular intermediate states of a dynamic reaction cycle, probably representing one of the first engagement steps of the complex with chromatin. Likewise, although we saw no prominent effects of different linkers on protein binding to modifications and vice versa, a dynamic interplay between the two cannot be excluded. The testable transcription-factor-binding sites in the linkers were located distant from the nucleosome-bound DNA regions, and histone modifications were unlikely to directly modulate their accessibility. In the presence of ATP, nucleosomal modifications can potentially modulate chromatin remodelling activities that could in turn expose nucleosomal DNA sequences, therefore facilitating, for example, the binding of pioneer transcription factors[42] thereby enabling the establishment or maintenance of NDRs.

Notably, modifications that are characteristic of distinct chromatin states vary greatly in their regulatory potential, as promoter-associated H3K4me3 and hyperacetylated H3 and H4 tails affect the binding of many nuclear factors, while enhancer-associated H3K4me1 and H3K27ac appear largely inert in targeting proteins to chromatin. Consistent with previous findings[43,44], this suggests that modifications found at enhancers may act, for example, by preventing the binding of repressive factors to the underlying regulatory loci[45], rather than by directly recruiting proteins.

Our study unifies two complementary views of chromatin—the modification-centric view that defines chromatin states based on chromatin marks[1,2], and the protein-centric view that defines the chromatin states by their protein constituents[46]. By combining both aspects, our experiments reveal major principles of how complex modification patterns define and regulate functional chromatin states. Our data are easily accessible through the interactive online resource MARCS (https://marcs.helmholtz-munich.de) with the aim to serve as a platform for both hypothesis generation and validation, and thereby act as a catalyst for future chromatin research. We encourage researchers to thoroughly explore the data as there are many discoveries to be made.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07141-5.

1. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
2. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
3. Musselman, C. A., Lalonde, M.-E., Côté, J. & Kutateladze, T. G. Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* **19**, 1218–1227 (2012).
4. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
5. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
6. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications—cause and consequence of genome function. *Nat. Rev. Genet.* **23**, 563–580 (2022).
7. Garcia, B. A., Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Pervasive combinatorial modification of histone H3 in human cells. *Nat. Methods* **4**, 487–489 (2007).
8. Pesavento, J. J., Bullock, C. R., LeDuc, R. D., Mizzen, C. A. & Kelleher, N. L. Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. *J. Biol. Chem.* **283**, 14927–14937 (2008).
9. Voigt, P. et al. Asymmetrically modified nucleosomes. *Cell* **151**, 181–193 (2012).
10. Young, N. L. et al. High throughput characterization of combinatorial histone codes. *Mol. Cell Proteom.* **8**, 2266–2284 (2009).
11. Tvardovskiy, A., Schwämmle, V., Kempf, S. J., Rogowska-Wrzesinska, A. & Jensen, O. N. Accumulation of histone variant H3.3 with age is associated with profound changes in the histone methylation landscape. *Nucleic Acids Res.* **45**, 9272–9289 (2017).
12. Shema, E. et al. Single-molecule decoding of combinatorially modified nucleosomes. *Science* **352**, 717–721 (2016).
13. Liu, C. L. et al. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3**, e328 (2005).
14. Rando, O. J. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.* **22**, 148–155 (2012).
15. Ruthenburg, A. J., Li, H., Patel, D. J. & Allis, C. D. Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell Biol.* **8**, 983–994 (2007).
16. Li, B. et al. Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science* **316**, 1050–1054 (2007).
17. Tsai, W.-W. et al. TRIM24 links a non-canonical histone signature to breast cancer. *Nature* **468**, 927–932 (2010).
18. Eustermann, S. et al. Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. *Nat. Struct. Mol. Biol.* **18**, 777–782 (2011).
19. Ruthenburg, A. J. et al. Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions. *Cell* **145**, 692–706 (2011).

20. Su, W.-P. et al. Combined interactions of plant homeodomain and chromodomain regulate NuA4 activity at DNA double-strand breaks. *Genetics* **202**, 77–92 (2016).
21. Borgel, J. et al. KDM2A integrates DNA and histone modification signals through a CXXC/PHD module and direct interaction with HP1. *Nucleic Acids Res.* **45**, gkw979 (2016).
22. Jurkowska, R. Z. et al. H3K14ac is linked to methylation of H3K9 by the triple Tudor domain of SETDB1. *Nat. Commun.* **8**, 2057 (2017).
23. Bartke, T. & Groth, A. A chromatin-based signalling mechanism directs the switch from mutagenic to error-free repair of DNA double strand breaks. *Mol. Cell. Oncol.* **6**, 1605820 (2019).
24. Xie, S. & Qian, C. The growing complexity of UHRF1-mediated maintenance DNA methylation. *Genes* **9**, 600 (2018).
25. Bartke, T. et al. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470–484 (2010).
26. Sidoli, S. et al. Middle-down hybrid chromatography/tandem mass spectrometry workflow for characterization of combinatorial post-translational modifications in histones. *Proteomics* **14**, 2200–2211 (2014).
27. Muir, T. W. Semisynthesis of proteins by expressed protein ligation. *Annu. Rev. Biochem.* **72**, 249–289 (2003).
28. Poli, J., Gasser, S. M. & Papamichos-Chronakis, M. The INO80 remodeller in transcription, replication and repair. *Philos. Trans. R. Soc. B* **372**, 20160290 (2017).
29. Geng, Z. & Gao, Z. Mammalian PRC1 complexes: compositional complexity and diverse molecular mechanisms. *Int. J. Mol. Sci.* **21**, 8594 (2020).
30. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
31. Vermeulen, M. et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).
32. Kleine-Kohlbrecher, D. et al. A functional link between the histone demethylase PHF8 and the transcription factor ZNF711 in X-linked mental retardation. *Mol. Cell* **38**, 165–178 (2010).
33. Schmitges, F. W. et al. Histone methylation by PRC2 is inhibited by active chromatin marks. *Mol. Cell* **42**, 330–341 (2011).
34. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2018).
35. Faith, J. J. et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
36. Meyer, P. E., Lafitte, F. & Bontempi, G. minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* **9**, 461 (2008).
37. Medvedeva, Y. A. et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015**, bav067 (2015).
38. Meldal, B. H. M. et al. The complex portal—an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.* **43**, D479–D484 (2014).
39. Voong, L. N. et al. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* **167**, 1555–1570 (2016).
40. Tompkins, V. S. et al. A novel nuclear interactor of ARF and MDM2 (NIAM) that maintains chromosomal stability. *J. Biol. Chem.* **282**, 1322–1333 (2006).
41. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
42. Sinha, K. K., Bilokapic, S., Du, Y., Malik, D. & Halic, M. Histone modifications regulate pioneer transcription factor cooperativity. *Nature* https://doi.org/10.1038/s41586-023-06112-6 (2023).
43. Sankar, A. et al. Histone editing elucidates the functional roles of H3K27 methylation and acetylation in mammals. *Nat. Genet.* **54**, 754–760 (2022).
44. Zhang, T., Zhang, Z., Dong, Q., Xiong, J. & Zhu, B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.* **21**, 45 (2020).
45. Bleckwehl, T. et al. Enhancer-associated H3K4 methylation safeguards in vitro germline competence. *Nat. Commun.* **12**, 5771 (2021).
46. Filion, G. J. et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224 (2010).
47. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
48. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
49. Pardo, M. et al. Myst2/Kat7 histone acetyltransferase interaction proteomics reveals tumour-suppressor Niam as a novel binding partner in embryonic stem cells. *Sci. Rep.* **7**, 8157 (2017).
50. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

Methods

Experimental procedures

Preparation of recombinant canonical histones. Recombinant human canonical histone proteins were expressed in *Escherichia coli* BL21(DE3)-CodonPlus-RIL cells (Agilent Technologies) from pET21b(+) (Novagen) vectors and purified by denaturing gel filtration and ion-exchange chromatography as previously described[25,51].

Preparation of recombinant histone H2A.Z. A codon-optimized sequence encoding human H2A.Z (H2AFZ, UniProtKB: P0C0S5) was purchased from GenScript and cloned into the NdeI/XhoI sites of the pET24a(+) vector (Novagen). H2A.Z was then expressed in *E. coli* BL21(DE3)-CodonPlus-RIL cells (Agilent Technologies) and purified as previously described for canonical H2A[25].

Preparation of truncated histones for native chemical ligations. Truncated human H3Δ1–31T32C protein for ligations of modified histone H3 was expressed in *E. coli* BL21(DE3)-CodonPlus-RIL cells (Agilent Technologies) and purified as previously described[52]. Truncated human H4Δ1–28I29C protein for ligations of modified histone H4 was expressed from pET24b(+) vectors (Novagen) in *E. coli* BL21(DE3)-CodonPlus-RIL cells (Agilent Technologies). The insoluble protein was extracted from inclusion bodies with unfolding buffer (20 mM Tris (pH 7.5), 7 M guanidine hydrochloride, and 100 mM dithiothreitol (DTT)) for 1 h at room temperature, and the cleared supernatant was loaded onto a Sephacryl S-200 gel filtration column (Cytiva) in SAU-1000 buffer (20 mM sodium acetate (pH 5.2), 7 M urea, 1 M NaCl, and 1 mM ethylenediaminetetraacetic acid (EDTA)) without any reducing agents. Positive fractions were combined and further purified by reversed-phase chromatography. Truncated H3Δ1–31T32C was purified over a Resource RPC column (Cytiva) using a gradient of 0–65% B (buffer A: 0.1% trifluoroacetic acid in water; B: 90% acetonitrile, 0.1% trifluoroacetic acid) over 20 column volumes. Truncated H4Δ1–28I29C was purified over a PerkinElmer Aquapore RP-300 (C8) column (250 mm × 4.6 mm inner diameter) using a gradient of 0–65% B (buffer A: 0.1% trifluoroacetic acid in water; B: 90% acetonitrile, 0.1% trifluoroacetic acid) over 20 column volumes. The fractions containing pure H3Δ1–31T32C or H4Δ1–28I29C were pooled and lyophilized.

Preparation of modified histone H3 and histone H4 by native chemical ligation. For the preparation of modified histone H3, N-terminal H3 peptides (amino acids 1–31) were ligated to truncated H3Δ1–31T32C and, for the preparation of modified histone H4, N-terminal H4 peptides (amino acids 1–28) were ligated to truncated H4Δ1–28I29C using native chemical ligation. All peptides contained a C-terminal benzyl thioester. All histone H4 peptides were N-terminally acetylated. Ligations were performed in 550 µl of degassed ligation buffer (200 mM KPO_4, 2 mM EDTA, 6 M guanidine hydrochloride) containing 1 mg of modified/unmodified histone tail thioester peptide (purchased from Cambridge Peptides or Almac Sciences), 4 mg of truncated histone, 20 mg 4-mercaptophenylacetic acid and 25 mg Tris(2-carboxyethyl) phosphine as reducing agent at a pH of 7.5. The reactions were incubated overnight at 40 °C and quenched by addition of 60 µl 1 M DTT and 700 µl 0.5% acetic acid. After precipitation clearance by centrifugation, the ligation reactions were directly loaded and purified onto a reversed-phase chromatography column (PerkinElmer Aquapore RP-300 (C8) 250 mm × 4.6 mm inner diameter). Modified histone H3 was purified using a gradient of 45–55% B (buffer A: 0.1% trifluoroacetic acid in water; B: 90% acetonitrile, 0.1% trifluoroacetic acid) over 10 column volumes. Modified histone H4 was purified using a gradient of 35–45% B (buffer A: 0.1% trifluoroacetic acid in water; B: 90% acetonitrile, 0.1% trifluoroacetic acid) over 10 column volumes. Positive fractions containing ligated full-length histone H3 or histone H4 were then combined and lyophilized.

Nucleosome assembly. Histone octamers were refolded from the purified histones and assembled into nucleosomes with biotinylated DNA through salt deposition dialysis as previously described[25,51]. Biotinylated nucleosomal DNAs containing either one (mononucleosomes) or two 601 nucleosome-positioning sequences[47] separated by a 50-base-pair (bp) linker (dinucleosomes), or four 601 nucleosome-positioning sequences (tetranucleosomes), were prepared as described previously[25]. CpG-methylated DNA was prepared using the M.SssI methyltransferase and complete methylation was confirmed by restriction digest (Supplementary Information). Dinucleosomes and tetranucleosomes were assembled in the presence of mouse mammary tumour virus A (MMTVA) competitor DNA (prepared in the same way as 601 DNA) and a slight excess of octamers as described for longer chromatin arrays to ensure saturation of the 601 repeats[53]. The reconstituted nucleosomes were then immobilized on streptavidin Sepharose High Performance beads (Cytiva) through the biotinylated DNA, washed to remove MMTVA competitor DNA and MMTVA nucleosomes (in the case of dinucleosomes and tetranucleosomes), and directly used for SILAC or label-free nucleosome affinity purifications. Correct assembly and immobilization of nucleosomes was verified by native polyacrylamide gel electrophoresis (Supplementary Information). Nucleosomes for pull-downs in which only modifications on histone H3 were tested were assembled with octamers containing recombinant histone H4 purified from *E. coli* instead of ligated H4. Likewise, nucleosomes for pull-downs in which only modifications on histone H4 were tested contained recombinant H3 and not ligated histone H3. Matched unmodified control nucleosomes were assembled with unmodified ligated H3 and recombinant H4, or recombinant H3 and unmodified ligated H4 accordingly. Nucleosomes containing only CpG methylation (H27M) were assembled with ligated unmodified H3 and recombinant H4, and nucleosomes containing only H2A.Z (H36) and no other modifications were assembled with recombinant (and therefore unmodified) H3 and H4 produced in *E. coli*.

Generation of 601 dinucleosomes incorporating different linker DNAs. Plasmid constructs for the preparation of biotinylated 601 dinucleosome DNAs containing different linker lengths (35 bp, 40 bp, 45 bp, 50 bp and 55 bp linkers) between the two 601 nucleosome-positioning sequences were generated by annealing forward and reverse primers of corresponding length and ligating them into pUC19-di601_NcoI/NheI_5xGal4 (pTB891, gene synthesis by Genscript) digested with NcoI and NheI restriction enzymes (Thermo Fisher Scientific), thereby exchanging the '5×Gal4 linker' against the different linker fragments. Plasmid constructs for the preparation of biotinylated 601 dinucleosome DNAs containing 200 bp linkers consisting of either the SV40 enhancer or the SV40 promoter were generated by PCR amplification of the SV40 enhancer and promoter sequences from pGL3-control (Promega) and cloning the resulting fragments into the vector backbone of pUC19-di601_NcoI/NheI_5xGal4 through NcoI and NheI, thereby exchanging the '5×Gal4 linker' against the 200 bp SV40 enhancer or promoter sequences. For all of the constructs, the dinucleosome sequences were then amplified from one copy to eight copies per plasmid as described previously[25,51].

The biotinylated 601 dinucleosome DNAs containing 200 bp linkers with randomized DNA sequences were generated from a library of single-stranded 200 bp scrambled linker oligonucleotides (custom synthesis by Biolegio) containing 192 bp of randomized DNA sequence flanked by 5′ NcoI and 3′ NheI restriction sites and 5′ bGHR and 3′ pCIfor primer-binding sites. The single-stranded oligo was converted to double-stranded DNA by annealing it to the pCIfor primer (Sigma-Aldrich) and performing a primer extension of pCIfor. The primer extensions were performed using Taq DNA polymerase in a 96-well plate format with 96 × 50 µl reactions. Each 50 µl reaction contained 1 µg of the 200 bp scrambled linker oligonucleotide (250 nM), 340 ng pCIfor primer (1 µM, fourfold molar excess over the 200 bp scrambled linker oligonucleotide), 200 µM dNTPs and 2.5 U Taq polymerase (New England Biolabs) in 1× ThermoPol buffer (New England Biolabs). Using a thermocycler, the oligonucleotides were denatured for 5 min at 95 °C, annealed for 1 min at 58 °C and the primer extension reaction was then allowed to proceed for 5 min at 68 °C. The reactions were pooled and the remaining single-stranded DNA was removed by direct addition of 2,000 U of exonuclease I (New England Biolabs) per ml reaction volume and incubation for 30 min at 37 °C. The resulting double-stranded DNA was purified using the QIAquick PCR purification kit (Qiagen) according to the manufacturer's instructions (20× columns, total yield of 75 µg in 1 ml buffer EB). The double-stranded 200 bp scrambled linker DNAs were digested with NcoI and NheI (Thermo Fisher Scientific) using 5 µl of FastDigest enzyme per µg DNA, concentrated using the QIAquick PCR purification kit (10× columns, total elution volume of 500 µl buffer EB) and separated by 2.5% agarose gel electrophoresis. The 200 bp band containing the scrambled linker fragments was excised from the gel and purified using the QIAquick gel extraction kit (Qiagen) according to the manufacturer's instructions (eight columns, total yield of 11.64 µg in 300 µl buffer EB). The purified NcoI/NheI-digested 200 bp scrambled linker fragments were subsequently ligated into the NcoI/NheI-digested, dephosphorylated (Quick CIP, New England Biolabs) and agarose-gel-purified vector backbone of pUC19-di601_NcoI/NheI_5×Gal4, thereby exchanging the '5×Gal4 linker' against the library of 200 bp scrambled linker fragments. Ligations were assembled using 50 µg of NcoI/NheI-linearized pUC19-di601 vector backbone, 11.64 µg of NcoI/NheI-digested 200 bp scrambled linker inserts (approximately 3.5-fold molar excess of inserts over the 3 kb vector backbone) and 200 µl (400,000 cohesive end units) of T4 DNA Ligase (New England Biolabs) in a total volume of 4 ml of 1× T4 DNA ligase reaction buffer, and incubated overnight at 16 °C. After the ligation, ATP was added to the reaction to a final concentration of 1 mM and unligated linear DNA was digested by addition of 1,000 U of exonuclease V (New England Biolabs) and incubation for 50 min at 37 °C. Circular plasmid DNA that was protected from the exonuclease V digestion was then purified and concentrated using the QIAquick PCR purification kit (10 columns, elution in 30 µl buffer EB per column). The total yield of ligated circular plasmid DNA was 6.5 µg in 280 µl. The ligated plasmids represent a library of pUC19 vectors in which each vector contains one copy of a 601 dinucleosome DNA each incorporating a different 200 bp linker of random sequence between the two 601 nucleosome-positioning sequences. The plasmid library was amplified by electroporation into 10-beta electrocompetent *E. coli* cells (New England Biolabs) according to the manufacturer's instructions using 2 µl (47 ng) of library DNA and 25 µl of competent cells per electroporation. Cells were recovered in 1 ml of outgrowth medium and selected on 24.5 cm^2 BioAssay LB$_{\mathrm{Amp}}$-agar plates (Corning). Serial dilutions were plated to determine the transformation efficiency and complexity of the library. In total, >10^8 independent clones were obtained from 24 electroporations. The colonies were gently scraped off the plates in liquid LB medium and plasmid DNA was isolated using the NucleoBond PC 10000 Giga-prep kit (Macherey-Nagel). The total yield of plasmid DNA from 24 plates was 16 mg. In total, 20 clones were picked from a high-dilution plate and sequenced to verify the correct length and random composition of the 200 bp linker sequences.

For preparing the different biotinylated dinucleosome DNAs the pUC19 601 dinucleosome plasmid constructs were first digested with EcoRV, ethanol-precipitated and then further digested with EcoRI (New England Biolabs) to liberate the dinucleosome DNAs. After another ethanol precipitation, the EcoRI overhangs were filled in with dATP and biotin-11-dUTP (Yorkshire Bioscience) using Klenow (3′→5′ exo$^-$) polymerase (New England Biolabs). The biotinylated dinucleosome DNAs were again concentrated by ethanol precipitation, separated from the pUC19 vector DNA by preparative agarose gel electrophoresis and then purified from the excised gel slices using the NucleoSpin gel extraction Maxi kit (Macherey-Nagel). Biotinylation and the purity of the dinucleosome DNAs were verified by depletion with streptavidin Sepharose High Performance beads (Cytiva) and agarose gel electrophoresis of the inputs and supernatants (Supplementary Information). Dinucleosomes were then assembled in the presence of MMTVA competitor DNA as described above.

Eukaryotic tissue culture. HeLa S3 cells (ATCC, CCL-2.2) cells were obtained from the Cancer Research UK Clare Hall Laboratories Cell Services Facility and maintained in suspension culture at 37 °C under 5% CO$_2$ in RPMI 1640 medium. HeLa S3 cells were authenticated by morphology on the basis of their ability to grow both in suspension culture and as round spherical cells in adhesion culture. A HeLa Kyoto BAC cell line expressing the C-terminal localization and affinity purification (LAP)-tagged INO80 subunit ACTR5[48] was a gift from M. Mann (Max Planck Institute of Biochemistry). Cells were cultured at 37 °C under 5% CO$_2$ in Dulbecco's modified Eagle's medium (DMEM) containing 4.5 mg ml^{-1} glucose, 10% fetal calf serum, 1% penicillin–streptomycin and 1% L-glutamine and validated by immunoprecipitation and immunoblotting against the tagged ACTR5. MCF-7 cells (ATCC, HTB-22) were obtained from the Cell Services Facility of the IGBMC. Cells were cultured at 37 °C under 5% CO$_2$ in DMEM containing 4.5 mg ml^{-1} glucose, 10% fetal calf serum, 1 mM sodium pyruvate, 1% penicillin–streptomycin and 1% L-glutamine and authenticated by morphology and by regularly testing the induction of oestrogen-responsive genes by quantitative PCR with gene-specific primers or global RNA-sequencing after 17β-estradiol treatment. IMR90 human fibroblasts were purchased directly from ATCC (CCL-186) and cultured at 37 °C under 5% CO$_2$ in DMEM containing 4.5 mg ml^{-1} glucose, 10% fetal calf serum, 1 mM sodium pyruvate, 1% penicillin–streptomycin and 1% L-glutamine. Cells were authenticated by morphology and only maintained for a limited number of passages. All of the cell lines were tested and were mycoplasma free.

SNAP. SILAC-labelled nuclear extracts were prepared from HeLa S3 cells as previously described[25]. The isotopically light (R$_0$K$_0$) or heavy (R$_{10}$K$_8$) nuclear extracts were mixes of three independently prepared nuclear extracts. For each pull-down, nucleosomes corresponding to 12.5 µg of octamer were immobilized on 10 µl streptavidin Sepharose High Performance beads (Cytiva) in the final reconstitution buffer (10 mM Tris (pH 7.5), 250 mM KCl, 1 mM EDTA and 1 mM DTT; supplemented with 0.1% NP-40) and then rotated with 0.5 mg HeLa S3 SILAC-labelled nuclear extract in 1 ml of SNAP buffer (20 mM HEPES (pH 7.9), 150 mM NaCl, 0.2 mM EDTA, 10% glycerol) supplemented with 0.1% NP-40, 1 mM DTT and protease inhibitor cocktail (Roche) for 4 h at 4 °C. Nucleosome pull-downs with acetylated histones and the corresponding unmodified control pull-downs were supplemented with HDAC inhibitors (5 mM sodium butyrate (Sigma-Aldrich, B5887) and 250 nM TSA (Sigma-Aldrich, T1952)) to prevent removal of the acetyl modifications. After two washes with 1 ml SNAP buffer + 0.1% NP-40 and then two washes with 1 ml SNAP buffer without NP-40, the beads from both SILAC pull-downs (modified and unmodified control nucleosome) were pooled. The supernatant was completely removed, and bound proteins were eluted by on-bead digestion (see below).

Label-free nucleosome affinity purifications. Nuclear extracts were prepared from HeLa S3 cells as previously described[25] except that cells were cultured with 10% regular fetal calf serum and no isotopically labelled amino acids were used. Unlabelled nuclear extracts were a mix of three independently prepared nuclear extracts. Nucleosome pull-downs were performed in the same manner as described above for SNAP, except for the bead washing and protein elution steps, which were performed as follows: after incubation with nuclear extracts, beads with immobilized nucleosomes were washed three times with 1 ml SNAP buffer + 0.1% NP-40, the supernatant was completely removed and bound proteins were eluted by boiling the beads in 50 µl Laemmli sample buffer containing 1% SDS at 95 °C for 5 min. A 20 µl protein aliquot was then digested with trypsin using a filter-aided sample preparation (FASP) protocol and analysed using liquid chromatography–mass spectrometry (LC–MS) as described below.

Cross-linking ChIP for MS analysis. IMR90 human fibroblasts were cultured as described above. Cells were washed three times with PBS and cross-linked on the plate with 1.25 µM ethylene glycol bis(succinimidyl succinate) (EGS) and 0.75 µM disuccinimidyl glutarate in PBS for 30 min at room temperature. After the first cross-linking reaction, cells were washed twice with PBS and cross-linked with 1% formaldehyde in PBS at room temperature for 10 min. Cross-linking reactions were quenched by the addition of glycine solution in PBS to a final concentration of 125 mM and incubation at room temperature for 5 min. Cells were then washed three times with ice-cold PBS, collected by scraping and pelleted by centrifugation (1,000*g*, 5 min, 4 °C). Cells were lysed in a hypotonic buffer (10 mM Tris (pH 7.6), 5 mM NaCl, 1.5 mM $MgCl_2$) supplemented with 0.1% NP-40, protease inhibitor cocktail (Roche), 10 mM sodium butyrate and 1 mM DTT using a Dounce homogenizer as described previously[25]. Nuclei were pelleted by centrifugation (3,000*g*, 5 min, 4 °C), washed in hypotonic buffer supplemented with 300 mM NaCl and pelleted again (3,000*g*, 5 min, 4 °C). Nuclei were resuspended in nuclear lysis buffer (15 mM Tris (pH 7.6), 10% glycerol, 1% SDS) and incubated for 5 min on ice. Chromatin was pelleted by centrifugation (5,000*g*, 5 min, 4 °C), washed in chromatin wash buffer (15 mM Tris (pH 7.6), 300 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40, 0.5% Triton X-100), pelleted again (5,000*g*, 5 min, 4 °C) and resuspended in ChIP buffer (20 mM Tris (pH 7.6), 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.01% SDS) supplemented with protease inhibitor cocktail (Roche) and 10 mM sodium butyrate. DNA was fragmented to an average size of 150–300 bp by sonication (Qsonica, Q800R2, 70% amp, 10 s off, 10 s on, 40 min active sonication time, 4 °C). Chromatin debris was pelleted by centrifugation (16,000*g*, 10 min, 4 °C). Then, 25 µl of supernatant was used for DNA purification to check the average DNA fragment size and another 25 µl supernatant aliquot was transferred to a fresh tube, de-cross-linked as described below, and stored at 4 °C until it was later used as the input sample for histone PTM analysis to define the average levels of core histone PTMs in bulk chromatin. For DNA purification, the sample was mixed 1:1 with 2× de-cross-linking buffer (20 mM Tris (pH 7.6), 600 mM NaCl, 2% SDS) and incubated at 65 °C overnight. The next day, proteinase K was added and the mixture was incubated at 37 °C for 2 h. DNA was purified using the QIAquick PCR purification kit (Qiagen) and eluted in RNase/DNase-free water. RNase A was added and the mixture was incubated at 37 °C for 1 h. DNA was resolved on an agarose gel and visualized with ethidium bromide. Approximately 0.2 mg chromatin (as measured by DNA content) was used for each ChIP reaction with the following antibodies: anti-H3K4me1 (Abcam, ab8895), anti-H3K4me3 (Millipore, 17-614), anti-H3 (Active motif, 39163), anti-H4 (Abcam, ab31830). For H3K4me3 ChIP reactions, 0.6 mg chromatin was used. To boost the identification of H3K4 methylation-state-specific protein interactors, H3 and H4 ChIPs were performed using chromatin inputs partially depleted in H3K4me1- and H3K4me3-modified nucleosomes and co-bound protein factors. Specifically, H3K4me1 and H3K4me3 ChIPs were performed first, then the chromatin inputs used for the H3K4me1 and H3K4me3 ChIPs were combined and subsequently used for H3 and H4 ChIPs. This aimed to shift the composition of the bulk chromatin-associated proteome measured in H3 and H4 control ChIPs towards regions devoid of H3K4me1 and H3K4me3. The antibody–chromatin mixture was incubated overnight on a rotation wheel (25 rpm) at 4 °C. Antibodies were captured using a 1:1 mixture of protein A and protein G Dynabeads (Thermo Fisher Scientific) for 2 h at 4 °C while rotating on a rotation wheel (25 rpm); 40 µl of bead mixture was used per ChIP sample. Beads were washed three times with ice-cold ChIP buffer and twice with ice-cold ChIP buffer supplemented with NaCl to a final concentration of 500 mM. Antibodies and co-bound chromatin were eluted by boiling the beads in 30 µl of Laemmli sample buffer containing 1% SDS and supplemented with 300 mM NaCl for 10 min at 95 °C. The eluate was transferred to a fresh tube and incubated in a thermomixer at 65 °C and 500 rpm for 12 h. For the histone PTM proteomic analysis, eluted proteins as well as the input samples (see above) were resolved on a 4–20% polyacrylamide gel (Novex WedgeWell Tris-Glycin-Minigel, Invitrogen), histone bands were excised, in-gel derivatized, digested with trypsin and processed for LC–MS analysis as described below. For the identification and quantification of co-purified chromatin proteins, a 10 µl aliquot of the eluted proteins in Laemmli sample buffer was processed for trypsin digestion using a FASP protocol and analysed using LC–MS as described below.

Native chromatin immunoprecipitations for MS analysis. The HeLa Kyoto BAC cell line expressing the C-terminal LAP-tagged INO80 subunit ACTR5[48] was cultured as described above. Cells were collected by trypsinization and were washed three times with ice-cold PBS. Nuclei were isolated using a Dounce homogenizer under hypotonic conditions in the presence of 0.1% NP-40 as described previously[25]. Nuclei were resuspended in ice-cold MNase digestion buffer (10 mM Tris (pH 7.6), 15 mM NaCl, 60 mM KCl, 0.1% NP-40) supplemented with protease inhibitor cocktail (Roche) and 10 mM sodium butyrate, and MNase was added at a proportion of 150 U per approximately 20 × 10^6 nuclei. The nucleus suspension was transferred to a thermomixer and, after 2 min incubation at 37 °C and 400 rpm, $CaCl_2$ was added to a final concentration of 1.5 mM and the mixture was incubated at 37 °C for another 6 min. The MNase digestion was stopped by the addition of EDTA to a final concentration of 10 mM. The mixture was then diluted 1:1 with ice-cold 2× SNAP buffer (30 mM HEPES (pH 7.8), 300 mM NaCl, 0.1% NP-40, 20% glycerol, 0.4 mM EDTA) supplemented with protease inhibitor cocktail (Roche) and 10 mM sodium butyrate. The samples were rotated on a rotation wheel for 45 min at 4 °C and further incubated in a thermomixer at 4 °C and 1,000 rpm for another 15 min. Nuclear debris was pelleted by centrifugation (16,000*g*, 10 min, 4 °C). The resulting supernatants were transferred to fresh 1.5 ml low-protein-binding Eppendorf tubes and used for the purification of nucleosomes bound to the INO80 complex as described below. To determine the efficiency of the MNase digestion, the pellets containing the insoluble chromatin fraction were resuspended in 1× supernatant volume of SNAP buffer, supplemented with proteinase K, and incubated at 37 °C overnight. In parallel, 25 µl aliquots of the supernatants were transferred to fresh tubes, supplemented with proteinase K and incubated at 37 °C overnight. After proteins were digested with proteinase K, DNA was extracted using the QIAquick PCR purification kit (Qiagen) and eluted in RNase/DNase-free water. RNase A was added, and the mixtures were incubated at 37 °C for 1 h. The DNA was then resolved on an agarose gel and visualized with ethidium bromide. For each sample, another 25 µl aliquot of the supernatant was transferred to a fresh tube and subsequently used as the input sample to define average histone modification levels on bulk chromatin. For the purification of nucleosomes bound to the INO80 complex, 25 µl of GFP-Trap Agarose beads (ChromoTek) were added to MNase-digested supernatants and the mixture was incubated on a rotation wheel (25 rpm) overnight at 4 °C. The beads were pelleted by centrifugation (250*g*, 3 min, 4 °C), followed by two washes with ice-cold SNAP buffer and one wash with SNAP buffer supplemented with NaCl to the final concentration of 200 mM. The supernatant was completely discarded and the beads were resuspended in 40 µl of SNAP buffer supplemented with 1 µg of 3C protease (Sigma-Aldrich). The mixture was then incubated for 8 h at 4 °C. The beads were pelleted by centrifugation, and the supernatant was transferred to a fresh tube, mixed with Laemmli sample buffer and boiled at 95 °C for 5 min. To identify histone PTMs of INO80-bound nucleosomes the immunopurified proteins and input samples were resolved on a 4–20% polyacrylamide gel (Novex WedgeWell Tris-Glycin-Minigel, Invitrogen), histone bands were excised, in-gel derivatized, digested with trypsin and analysed using LC–MS as described below.

CRISPR–Cas9-mediated endogenous protein tagging. The core INO80 complex subunit INO80B was endogenously tagged at its C-terminus with a V5 epitope in the MCF-7 cell line using the tagging strategy described previously[54]. Specifically, 1 day before transfection, MCF-7 cells were seeded onto 24-well plates at approximately 1.0 × 10^5 cells per well in 500 µl of low-glucose DMEM medium supplemented with 10% FBS, 1 mM glutamine and 100 µg ml^{-1} penicillin–streptomycin. On the day of transfection, 25 µl of Opti-MEM medium was added to a 1.5 ml sterile Eppendorf tube, followed by the addition of 1,250 ng of TrueCut Cas9 Protein v2 nuclease (Invitrogen) and 240 ng of two-piece gRNA (crRNA:tracrRNA duplex) generated by annealing crRNA (IDT) and tracrRNA (IDT) according to the manufacturer's instructions. After mixing briefly by vortexing, 1 µl Cas9 Plus reagent was added to the solution containing Cas9 protein and gRNA. The mixture was incubated at 25 °C for 5 min to allow the formation of Cas9 ribonucleoprotein particles (RNPs). For co-delivery of homology donor DNA, 800 ng of single-stranded DNA oligonucleotide (IDT) was added to the Cas9 RNPs at this point. Meanwhile, 25 µl Opti-MEM medium was added to a separate sterile Eppendorf tube, followed by the addition of 1.5 µl of Lipofectamine CRISPRMAX. After briefly vortexing, the Lipofectamine CRISPRMAX solution was incubated at 25 °C for approximately 5 min. After incubation, the Cas9 RNPs were then added to the Lipofectamine CRISPRMAX solution. The mixture was incubated at 25 °C for 10–15 min to form Cas9 RNPs and Lipofectamine CRISPRMAX complexes and then added to the cells. At 48 h after transfection, the cells were collected by trypsination and seeded in 96-well plates at 1 cell per well. After reaching 60–80% confluency, the cells were trypsinized and split 1:1 into two 96-well plates where the first plate was used for immunofluorescence screening with monoclonal mouse anti-V5 primary antibodies (eBioscience, TCM5 14-6796-82, 1:250) and Alexa-Fluor-488-coupled anti-mouse IgGs as secondary antibodies (Jackson ImmunoResearch Laboratories, 715-545-150, 1:333), and the second plate was used for the subsequent expansion and further testing of V5-positive clones. The immunofluorescence screen for V5-positive clones was performed as previously described[54].

Co-IP. Approximately 1.0 × 10^7 MCF-7 WT or INO80B-V5 cells were used for nuclear extract preparations as described previously[25]. The nuclear extract was diluted with IP buffer (20 mM HEPES (pH 7.9), 50 mM NaCl, 0.2 mM EDTA, 5% glycerol, 0.1% NP-40, 1 mM DTT and protease inhibitor cocktail (Roche)) to a final protein concentration of around 1 µg μl^{-1} and a NaCl concentration of 160 mM and subsequently cleared by centrifugation at 20,000*g* for 10 min at 4 °C. Then, 1 ml of cleared nuclear extract was mixed with 5 µl of anti-V5 antibodies (Abcam, ab15828) and incubated on a rotating wheel over night at 4 °C. The next day, 20 µl of a 1:1 mixture of protein A and protein G Dynabeads (Invitrogen) were added to the sample followed by 1 h incubation on a rotation wheel at 4 °C. Magnetic beads were washed three times with the IP buffer containing 150 mM NaCl. Co-immunoprecipitated proteins were eluted from the beads by boiling in 20 µl of Laemmli sample buffer for 5 min at 95 °C. Eluted proteins were subsequently used for immunoblotting and LC–MS experiments (IP–MS). For LC–MS analysis, proteins were digested with trypsin using a FASP protocol as described below.

Protein detection by immunoblotting. Proteins were separated by SDS–PAGE and blotted onto nitrocellulose membranes (0.45 µm, Thermo Fisher Scientific) using a Bio-Rad PROTEAN mini-gel and blotting system. Antibodies were diluted in TBST + 5% milk (25 mM Tris (pH 7.5), 137 mM NaCl, 2.7 mM KCl, 0.2% Tween-20, 5% non-fat dry milk). The following primary antibodies were used for immunoblots: anti-V5 tag (eBioscience, TCM5 14-6796-82, 1:1,000), anti-INO80 (Abcam, ab118787, 1:2,000), anti-INO80B (Santa Cruz (E-3), sc-390009, 1:1,000), anti-ACTR5 (GeneTex, GTX80453, 1:1,000), anti-TBRG1 (Santa Cruz (D-9), sc-515620, 1:1,000), anti-H3K4me3 (Millipore, 17-614, 1:2,000), anti-H4 (Abcam, ab31830, 1 µg ml^{-1}), anti-H4ac (pan-acetyl) (Active Motif, 39967, 1:1,000), anti-CBX4 (Cell Signaling Technology, E6L7X 30559, 1:1,000), anti-CBX8 (Santa Cruz (C-3), sc-374332, 1:1,000), anti-H2B (Abcam, ab1790, 1:1,000), anti-H2A.Z (Abcam, ab4174, 1:1,000). Immunoblot images were acquired by CCD camera using the Bio-Rad ChemiDoc Touch Imaging System running Image Lab Touch Software (v.2.3.0.07).

MS methods

Sample preparation for MS. On-bead digestion and peptide purification for SNAP samples. The beads were resuspended in 50 µl of elution buffer (2 M urea, 100 mM Tris (pH 7.5), 10 mM DTT) and incubated on a shaker (1,000 rpm) at 25 °C for 20 min. Iodoacetamide (Sigma-Aldrich, I1149) was added to a final concentration of 50 mM and the sample was incubated on a shaker (1,000 rpm) at 25 °C in the dark for 10 min. After digestion with 0.3 µg trypsin (Promega V5113) for 2 h on a thermo shaker (1,000 rpm) at 25 °C, the supernatant was transferred to a new tube and was further digested with 0.1 µg trypsin overnight at 25 °C. The digestion was stopped by adding 5.5 µl of 10% trifluoroacetic acid. Eluted peptides were purified on C18 stage-tips (Glygen 10-200 µl TopTips) according to the manufacturer's instructions and dried using a SpeedVac.

FASP of label-free proteomics samples. Filter-aided sample preparation was performed as described previously[52]. In brief, 10–20 µl aliquots of protein mixtures in 1% SDS Laemmli sample buffer were diluted with 200 µl of 100 mM triethylammonium bicarbonate buffer (TEAB; pH 8.5). For protein reduction, 1 µl of 1 M DTT was added to each sample and the samples were incubated at 60 °C for 30 min. After cooling the samples to room temperature, 300 µl of freshly prepared UA buffer (8 M urea in 100 mM TEAB (pH 8.5)) was added to each sample. Proteins were alkylated by the addition of 10 µl of 300 mM iodacetamide solution and subsequent incubation for 30 min at room temperature in the dark. The samples were then concentrated to dryness in a 30 kDa cut-off centrifugal spin filter unit (Millipore), and washed three times with 200 µl UA buffer and twice with 200 µl of 50 mM TEAB (pH 8.5). Then, 40 µl of a 50 ng μl^{-1} trypsin solution in 50 mM TEAB (pH 8.5) was added to each sample and protein digestion was performed overnight at 37 °C. Peptides were centrifuged through the filter, and the collected flow through was acidified by the addition of trifluoroacetic acid to a final concentration of 0.5% (v/v). About 300 ng of the tryptic peptide mixtures was then used for LC–MS analysis as described below.

Histone sample preparation for proteomics analysis. Histone proteins were prepared for LC–MS analysis using a hybrid chemical derivatization protocol adopted for in-gel sample preparation. In brief, proteins were resolved on 4–20% polyacrylamide gels (Novex WedgeWell Tris-Glycin-Minigel, Invitrogen) followed by Coomassie staining. Histone protein bands were excised from the gel and destained in a destaining buffer (100 mM triethylammonium bicarbonate in 50% acetonitrile). After destaining, the gel pieces were dehydrated with 200 µl of 100% acetonitrile for 10 min at room temperature after which acetonitrile was discarded. Propionylation solution was prepared by mixing 50 mM TEAB (pH 8.5) and freshly prepared 1% (v/v) propionic anhydride solution in water at a 100:1 ratio. Immediately after preparation, 100 µl of propionylation solution was added to the dehydrated gel pieces followed by 10 min incubation at room temperature. The propionylation reaction was quenched by the addition of 10 µl of 80 mM hydroxylamine and subsequent incubation for 20 min at room temperature. The propionylation solution was discarded and gel pieces were dehydrated with 200 µl of 100% acetonitrile for 10 min at room temperature. After this, the acetonitrile solution was discarded and 20 µl of 50 ng μl^{-1} trypsin solution in 100 mM TEAB (pH 8.5) was added. Trypsin digestion was performed overnight at 37 °C. The next day, 50 µl of 100 mM TEAB (pH 8.5) solution was added to each sample followed by 30 min incubation in a thermo shaker (37 °C, 1,500 rpm). A 1% (v/v) solution of phenyl isocyanate in acetonitrile was freshly prepared and 15 µl added to each sample and incubated for 60 min at 37 °C. The samples were acidified by the addition of 24 µl 1% trifluoroacetic acid. Peptides were desalted with C18 spin columns (Thermo Fisher Scientific) according to the manufacturer's instructions, dried in a speed-vac, resuspended in 50 µl 0.1% trifluoroacetic acid and subsequently used for LC–MS analysis.

LC–MS-based proteomics measurements. MS analysis of SNAP samples. SNAP samples were processed and analysed by LC–MS on a Q-Exactive mass spectrometer (Thermo Fisher Scientific) as described previously[55]. In brief, the samples were loaded at 8 µl min^{-1} onto a trap column (Thermo Fisher Scientific, Acclaim PepMap 100; 100 µm internal diameter, 2 cm length, C18 reversed-phase material, 5 µm diameter beads and 100 Å pore size) in 2% acetonitrile and 0.1% trifluoroacetic acid. Each of the samples was loaded twice, providing two technical replicates. Peptides were eluted on line to an analytical column (Thermo Fisher Scientific, Acclaim PepMap RSLC; 75 µm internal diameter, 25 cm length, C18 reversed-phase material, 2 µm diameter beads and 100 Å pore size) and separated using a flow rate of 250 nl min^{-1} and the following gradient conditions: initial 5 min with 4% buffer B; a 90 min gradient of 4–25% B; a 30 min gradient of 25–45% B; a 1 min gradient 45–90% B; and finally 15 min isocratic at 100% B before returning to the starting conditions for a 15 min equilibration (buffer A: 2% acetonitrile and 0.1% formic acid in water; B: 80% acetonitrile and 0.1% formic acid). The Q-Exactive instrument acquired full-scan survey spectra (*m/z* 300–1,650) at 70,000 resolution. An automatic gain control target value of 3×10^6 and a maximum injection time of 20 ms were used. The top 10 most abundant multiply charged ions were selected in a data-dependent manner, fragmented by higher-energy collision-induced dissociation, and data were collected over the range 200–2,000 *m/z* at 17,500 resolution. An automatic gain control target value of 1×10^5 with a maximum injection time of 120 ms was used. A dynamic exclusion time of 30 s was enabled.

MS analysis of label-free proteomics samples. LC–MS/MS analysis of label-free nucleosome pull-downs and ChIP–MS proteomics samples was performed on the Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled in-line to a nanoEasy LC (Thermo Fisher Scientific). The samples were loaded in solvent A (0.1% formic acid) on a two-column set-up consisting of a 3.5 cm, 100 µm inner diameter pre-column packed with Reprosil-Pur 120 C18-AQ (5 µm; Dr. Maisch) and an 18 cm, 75 µm inner diameter analytical column packed with Reprosil-Pur 120 C18-AQ (3 µm; Dr. Maisch). A gradient of solvent B (95% acetonitrile, 0.1% formic acid) was applied at a flow rate of 250 nl min^{-1} as follows: 3% to 25% B in 90 min; 25% to 45% B in 30 min; 45% to 100% B in 3 min; and 100% B in 8 min. MS was obtained at a resolution of 120,000 and MS/MS as top 15 at a resolution of 15,000 and with a dynamic exclusion of 30 s. The maximum injection time was set to 100 ms for both MS and MS/MS and only peptides of charge state 2, 3 and 4 were selected for MS/MS.

LC–MS/MS analysis of INO80-V5 IP–MS samples was performed on the Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled to a nano-RSLC (Ultimate 3000, Dionex). In brief, the samples were automatically loaded onto a nano trap column (300 µm inner diameter × 5 mm, packed with Acclaim PepMap100 C18, 5 µm, 100 Å; LC Packings) before separation by reversed-phase chromatography (HSS-T3 M-class column, 25 cm, Waters) in a 95 min nonlinear gradient from 3 to 40% acetonitrile in 0.1% formic acid at a flow rate of 250 nl min^{-1}. Eluted peptides were analysed using the Q-Exactive HF mass spectrometer equipped with a nano-flex ionization source. Full scan MS spectra (*m/z* 300–1,500) and MS/MS fragment spectra were acquired in the Orbitrap with a resolution of 60,000 or 15,000, respectively, with maximum injection times of 50 ms each. Up to ten most intense ions were selected for higher-energy collisional dissociation fragmentation depending on signal intensity. Dynamic exclusion was set for 30 s.

MS analysis of histone samples. For LC–MS analysis of modified histone proteins, the acidified histone peptide digests were analysed on the Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled in-line to a nanoEasy LC (Thermo Fisher Scientific). In brief, the samples were automatically loaded onto an in-house packed 2 cm 100 µm inner diameter C18 pre-column with buffer A (0.1% formic acid) and then eluted and separated on an in-house packed Reprosil-Pur 120 C18-AQ (3 µm; Dr. Maisch) analytical column (20 cm × 75 µm inner diameter) using a 35 min linear gradient from 0% to 40% buffer B (90% acetonitrile, 0.1% formic acid). Full scan MS spectra (*m/z* 300–1,000) and MS/MS fragment spectra were acquired in the Orbitrap with a resolution of 120,000 or 15,000, respectively, with maximum injection times of 50 ms each. Up to the 20 most intense ions were selected for higher-energy collisional dissociation fragmentation depending on signal intensity. Dynamic exclusion was disabled.

MS RAW data search and quantification. Analysis of SNAP MS data. Protein abundances were quantified from the Q-Exactive raw data files using MaxQuant (v.1.5.2.8)[56] against the UniProt UP000005640 canonical proteome (downloaded in September 2016) using 2-plex labelling (Arg0/Lys0 and Arg10/Lys8). The search was performed allowing for fixed carbamidomethyl modification of cysteine residues and variable oxidation of methionine residues and acetylation of amino termini. The minimum peptide length was set to 7. All raw files resulting from the forward and reverse pull-downs, including technical replicates for each nucleosome tested, were processed together using the 'match between runs' feature. H/L ratios were computed in advanced ratio computation mode, with the minimal ratio and peptide count set to 1. The corresponding mqpar.xml file is deposited along with the proteomics data. Initial trial experiments with mono-, di- and tetra-nucleosomes (Supplementary Information) were quantified separately by MaxQuant v.1.5.1.0 against the December 2015 version of UniProt proteome with more stringent settings requiring at least two peptides for ratio estimation.

Analysis of label-free MS data. Protein identification and quantification was performed using Proteome Discoverer v.2.5 (Thermo Fisher Scientific). Data were searched against the human Swiss-Prot database using Mascot[57] as the search engine, with a precursor mass tolerance of 5 ppm and a fragment mass tolerance of 0.05 Da. Two missed cleavages were allowed for trypsin and carbamidomethylation of cysteine was set as a static modification, while oxidation of methionine was set as dynamic. Label-free quantification was achieved as match between runs by using the Minora Feature Detector, the Feature Mapper and the Precursor Ions Quantifier. The maximum retention time shift for chromatographic alignment was set to 2 min and the retention time tolerance for mapping features was set to 1 min. Peptide quantification was performed as the peak area normalized to the total peptide amount and protein quantification as the average of the top three unique peptides.

Analysis of histone MS data. For the identification and quantification of histone PTMs in ChIP–MS samples and the quality control of recombinantly produced modified histone proteins, MS raw data files were manually analysed using Skyline (v.20.1.0.31)[58]. In brief, a list of unmodified as well as differentially modified histone H3 and H4 peptides was manually compiled and used to evaluate the modification status of histones in each sample. All lysine residues not bearing acetylation or methylation were considered to be propionylated and all peptide N termini were considered to be modified with phenyl isocyanate. MS1 filtering was set to include 3 isotope peaks and the MS1 resolving power was set at 120,000. MS2 resolving power was set at 15,000. For each modified histone peptide, the relative abundance was estimated by dividing its peak area by the sum of the areas corresponding to all of the observed forms of that peptide (that is, all peptides sharing the same amino acid sequence). The relative abundance of histone variant H2A.Z was estimated by dividing the sum of peak areas of unique H2A.Z peptides (that is, only present in H2A.Z but not in any other H2A variants) by the sum of peak areas of all unique peptides corresponding to histones H2A, H2B and H2A.Z.

Data postprocessing and bioinformatic analyses

Data postprocessing. Postprocessing of SNAP MS data. MaxQuant proteinGroups entries marked as 'potential contaminant', 'reverse' or 'only identified by site' were removed from the datasets analysed. The SILAC H/L ratios for each of the remaining entries were converted to a $\log_2$ scale. In initial trial experiments (Supplementary Information), the median and first and third quartiles $\log_2$[H/L ratio] values were estimated in all experiments individually, treating forward and reverse experiments separately. Proteins were assumed to be significantly enriched if they fell 1.5× the interquartile range away from first and third quartiles for both forward and reverse experiments, matching the box plots. The data for the main set of experiments were additionally annotated with up to date (as of 30 July 2019) metadata that were downloaded from the mygene.info API service[59] based on the IDs in the 'Majority Protein ID' column. Protein identifiers were assigned readable counterparts on the basis of the associated gene names. Duplicate entries were enumerated in parentheses (for example, *SMARCA* (1) and *SMARCA* (2)), assigning lower numbers to entries with a higher MaxQuant score. Common prefixes of the gene names were collapsed (for example, *SMAD[2,3,9]*) for brevity. The principal direction of the data spread (that is, the direction of enrichment) in each of the pull-downs was estimated by determining the first principal component of the data in the top-left and bottom-right quadrants of the forward and reverse $\log_2$[H/L ratio] plot. The estimate was adjusted by re-evaluating the principal direction after removing outlier points ±2 s.d. away from the median in the second principal direction. Protein-specific variation in the second principal direction across pull-down experiments was adjusted to zero to correct systemic heavy and light cell population batch effects resulting from different abundances of proteins in the nuclear extracts from the H/L cell populations or different labelling efficiencies of proteins with the heavy-labelled amino acids. In cases in which either the forward or the reverse H/L ratio was measured for the protein (9.13% of ratio pairs), but not both, the missing ratio was imputed by projecting the measured ratio to the estimated principal enrichment line. In six cases (0.01%) in which the estimated H/L ratio was infinite as protein intensity could have been measured in the modified nucleosome, but not in the unmodified nucleosome, the ratio was imputed to the maximum ratio identified in the particular SNAP experiment. All other missing H/L ratios were imputed to zero (24.27%). Five proteins of which the forward and reverse H/L ratios were equal to zero in all of the experiments were removed. The resulting data for each of the pull-down experiments were then further rotated so the estimated principal direction of variation lays exactly on the ideal 45° diagonal, so the reverse ratio on average equals the negative of the forward one. For visualizations and computational analyses, the sign of the reverse experiment was flipped to be on the same scale as the forward one.

Postprocessing of cross-linked H3K4me1 and H3K4me3 ChIP–MS data. Protein abundances obtained from H3K4me1 and H3K4me3 cross-linking-ChIP–MS experiments were converted to $\log_2$ scale, treating zero abundances as missing data. The data were normalized to ten histone proteins observed in the data: *H2AC20*, *H2AC21*, *H2AW*, *H2AZ2*, *H2BC4*, *H2BU1*, *H3-2*, *H4C1*, *MACROH2A1* and *MACROH2A2*. Specifically, we calculated the average $\log_2$-transformed abundance for the histone proteins in each of the experiments, and calculated the residuals (that is, $\log_2$-transformed abundances minus the average (*M* value)) for the histone proteins. The data were normalized by subtracting the median of these residuals for each of the samples, so that the median *M* value of the normalized data for the histone proteins remains approximately zero across experiments. The normalized data were then further filtered to include only proteins that were detected in at least two replicates of at least one experiment.

We used limma[60] to estimate the $\log_2$[FC] values between H3K4me3 and controls (H3 and H4), H3K4me1 and controls, and H3K4me3 and H3K4me1. Specifically, we used a zero-intercept means model encoding one parameter for each experiment (H3, H4, H3K4me1, H3K4me3), and analysed the contrasts between protein abundance in H3K4me1/3 experiments and the average abundance of H3 and H4 (for example, (H3 + H4)/2), as well as a contrast between H3K4me3 and H3K4me1. The analysis was run using the default parameters of limma (v.3.50.1), with the addition of 'robust=True' in the 'eBayes' step, hypothesis testing was performed using the default settings, assuming zero $\log_2$[FC] under the null hypothesis. *P* values were corrected using the Benjamini–Hochberg procedure, and significance was assumed at an FDR of 0.05.

In some cases, the contrasts could not be estimated due to missing data. This frequently happened when proteins were detected in one of the experiments, but not in controls (or vice versa). In these cases, we imputed such $\log_2$[FC] estimates with infinities (positive and negative). Moreover, whenever it was possible to estimate the H3 or H4 controls, but not both, we imputed the $\log_2$[FC] estimates using one of such controls only. The imputed estimates are clearly flagged in the data and figures. Estimates based only on single data points (that is, an observed abundance in one of the three replicates only) are flagged as well.

To be able to link the ChIP–MS data with MARCS feature effect estimates, we mapped the ChIP–MS proteins to their MARCS counterparts through their accession numbers and gene names. The cases in which one ChIP–MS protein mapped to multiple proteins in the MARCS dataset were resolved by assigning the feature effect estimate with the lowest *P*-value estimate across all of the matched identifiers.

To obtain association statistics, we performed a Mann–Whitney *U*-test, comparing the imputed ChIP–MS $\log_2$[FC] estimates of proteins strongly recruited to or excluded by a MARCS feature to the imputed $\log_2$[FC] estimates of other proteins detected in both MARCS and ChIP–MS data. Only the groups with at least five proteins were tested. For visualization purposes, we computed the mean $\log_2$[FC] estimates in each of the groups, and their respective differences. For this purpose, we assumed the infinities to be equal to the maximum finite $\log_2$[FC] plus a small number.

Postprocessing of variable-linker nucleosome pull-down data. Label-free MS quantification datasets for the short linker nucleosome, long linker SV40 promoter nucleosome and long linker SV40 enhancer nucleosome affinity-purification experiments were analysed independently. The protein abundances were converted to a $\log_2$ scale, treating zero intensities as missing values. The data were normalized using the abundances of HIST1H4A and HIST2H2BF histones (short linkers) or H4C1 and H2BC12 histones (long linkers) as described in the H3K4me1/3 cross-linking-ChIP–MS methods.

For each set of experiments, we used a zero-intercept means model in limma and hand-crafted contrasts to measure two types of effects on protein binding to dinucleosomes: (1) modification-specific effects, that is, the $\log_2$-transformed FC in protein abundance between modified nucleosome and unmodified nucleosome, given a specific linker of certain length, for example, $\log_2$[H3K27me3 with 50 bp linker] versus $\log_2$[unmodified with 50 bp linker], as well as (2) linker-specific effects, that is, the $\log_2$-transformed FC in protein abundance between two different linkers, given a certain nucleosome modification, for example, $\log_2$[H3K27me3 with 55 bp linker] versus $\log_2$[H3K27me3 with 50 bp linker]. Owing to the large number of missing values, the second replicate of the H3K27me3 experiment with 35 bp linker was excluded from the analysis. Only proteins that had at least two values in at least one condition were analysed.

The analysis was run using the default parameters of limma (v.3.50.1), using the 'robust=True' parameter in the 'eBayes' step. *P* values were corrected using the Benjamini–Hochberg procedure, assuming significance at an FDR of 0.05. In addition to this, significant estimates were considered to be 'strong' if the absolute $\log_2$[FC] was greater than 1.

As in the H3K4me1/3 cross-linking-ChIP–MS experiment, we imputed contrasts that could not be estimated from the data using the following heuristics: proteins detected in one of the conditions, but not the other, received either infinite enrichments or infinite depletions. Such imputed estimates were flagged in the data, together with estimates based on single data points.

To aid the data visualization, we divided the proteins into three groups on the basis of the effects of the modifications and linkers on dinucleosome binding in the different analyses: (1) modification-responsive proteins, that is, proteins that have a significant and strong response to a modification signature in at least one of the linkers visualized; (2) linker-responsive proteins, that is, proteins with a significant and strong response to the linker in either modified or unmodified nucleosomes; and (3) proteins that respond to both, that is, satisfy conditions (1) and (2) simultaneously.

Postprocessing of endogenous INO80B-V5 IP–MS data. For analysis of INO80B-V5 IP-MS data, only proteins identified based on three or more unique peptides were considered. The quantified MS1 protein abundances were normalized to the IGHG1 abundance. Differential enrichment analysis was performed using a two-tailed *t*-test. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg method. The protein stoichiometry was determined using MS1-based label-free quantification[61]. Specifically, protein abundances were calculated as the mean of MS1 intensities of all unique peptides identified for the protein. To assess the stoichiometry of INO80 complex subunits, the abundance of each subunit (mean of unique MS1 peptide intensities) was divided by the abundance of INO80B (mean of unique INO80B MS1 peptide intensities) used as a bait in co-IP complex purification experiments.

Decoupling of the effects of individual modification features (SNAP dataset). Pairs of nucleosomes differing by a single modification only were identified by arranging the nucleosomes into a directed graph of which the edges track the difference by one modification, including self-informative nucleosomes that contain only one chromatin feature (for example, H3K4me3). H3K9acK14ac, full acetylation on histone H3 (H3K9acK14acK18acK23acK27ac), H4K5acK12ac and fully acetylated H4 (H4K5acK8acK12acK16ac) were treated as single modification. Only chromatin features that have two or more informative nucleosome pairs, and therefore an independent experimental replicate, were analysed. As each pull-down consists of a forward and reverse experiment, this results in at least four experimental measurements, enabling a robust statistical analysis. Moreover, a feature effect estimate was computed only for proteins that have at least one nucleosome pair with no imputed data.

The relationship between nucleosomes was modelled in limma using the following formula: '~ 0 + edge + ptm'. Here the 'edge' parameter tracks edges in the directional graph and ptm captures the direction of the edge and is set to one at the endpoint that contains the target feature and zero at the other. This expression allows the baseline effect of a nucleosome pair to be captured by the 'edge' parameter allowing the 'ptm' parameter to measure the change of the effect caused by the modification feature (that is, a PTM, histone H2A.Z or DNA methylation). Self-evident purifications were assigned no edge coefficient. Limma was run with robust empirical Bayes, with weights set to number of unique peptides detected plus one. Significance was assumed at Benjamini–Hochberg-adjusted FDR of 0.01.

Significant responses were additionally labelled as strong if their parameter estimates were greater than or equal to 1. For the proteins that respond strongly to at least one feature, the collective modification response profiles across all features were clustered. The clustering was performed using protoclust[62] (v.1.6.3) under cosine distance. The dendrogram corresponds to Minimax Hierarchical Linkage. In cases in which no estimate for the effect could be made, for clustering purposes the values were imputed using three nearest neighbours (bnstruct package[63]). The resulting dendrogram was divided into 40 flat clusters that were annotated with their respective prototype protein in Fig. 2e and Supplementary Table 5.

The joint response of protein complexes to chromatin features was analysed using CAMERA[64]. Only complexes with 3–40 members (inclusive) were analysed. Significance was assumed at a Benjamini–Hochberg-adjusted FDR of 0.01. Whenever possible, the enrichment of both the whole protein complex, and the enrichment of only the exclusive subunits of the complex, not including subunits shared with other complexes, was tested. The median effect of chromatin features on protein complexes was estimated from 100,000 random samples from the effect distributions of individual subunits. The median, as well as the empirical 95% confidence interval (CI) is reported.

Network inference (SNAP dataset). We used the network inference algorithms ARACNE, MRNET and CLR implemented in the minet package[36] to infer the protein–protein interaction networks in an unsupervised manner, using only the 1,915 × 110 matrix of processed $\log_2$-transformed heavy/light ratios of identified proteins as the input. The algorithms were configured to use Miller-Madow (mi.mm) estimator for MI and the equal width discretization strategy with the bin number set to 10. In addition to the algorithms above, the performance of the MI metric on its own (without subjecting it to network algorithms) was also evaluated (network RAW-MI).

In addition to the MI-based methods above, we have benchmarked the networks defined by the interprotein correlation matrix computed both naively (CORR) or using Ledoit–Wolf shrinkage (CORR-LW)[65]. These networks were built by assuming the adjacency between the nodes to be equal to the corresponding entry in the correlation matrix. Negative values in the correlation matrix were avoided by adding one to each of the entries and dividing the result by two.

The inferred networks were evaluated against the BioGRID database[34] (release 3.5.174) after training. BioGRID entries were linked with our identifiers through Entrez identifiers downloaded previously through the mygene.info API service[59]. Networks were evaluated by computing their precision (fraction of predicted edges in the network that were also in the BioGRID database) and recall (fraction of edges in BioGRID database that were predicted by the network) at multiple stringency levels. We used the scaled truncated area under precision and recall curve (auPRC) statistic[66], which combines the multiple precision/recall estimates into a single score as our primary metric. As we did not anticipate a full recovery of BioGRID interactions by our networks and therefore wanted to trade higher precision for lower recall, we did not consider any threshold settings with a recall of greater than 0.2 for the evaluation of the algorithms. Interactions with histone proteins, as well as self-interactions (either homodimers in BioGRID or interactions between two proteins with the same gene name) were excluded from the evaluation.

To produce the inferred networks described in the paper, we noted that the scores of the CLR algorithm can be converted to *P* values by noting that for the CLR scoring function $s(i,j) = \sqrt{\max(0, z_i)^2 + \max(0, z_j)^2}$ where z_i and z_j are assumed to follow standard normal distribution under the null hypothesis[35,36], the *P* values under null can be expressed as $P(s(i,j) \geq x > 0) = \frac{1}{4}(2 \times \mathrm{erfc}(\frac{x}{\sqrt{2}}) + e^{-x^2/2})$. Where erfc is the complementary error function. Adjusting those computed *P* values for multiple hypothesis testing using the Benjamini–Hochberg procedure (that is, converting them to a *q* value) enabled us to pick a set of intuitive thresholds to produce the networks presented in the paper.

Networks at different adjusted *q*-value thresholds were drawn using the Force Atlas 2 algorithm in gephi[67] and adjusted manually. Only proteins with at least five non-zero values were drawn. Isolated nodes (connected components with size of 1) were not drawn. Network nodes were either coloured by communities (Louvain algorithm[68] implemented in the Python-Louvain package) or overlaid by the colour-coded chromatin response estimates (see the 'Decoupling of the effects of individual modification features (SNAP dataset)' section above). In the network projection plots, the names of protein complexes were annotated manually on the basis of protein complexes that were significantly regulated by the chromatin modification (as reported by the CAMERA procedure), and had empirically estimated median effects of at least 0.3. Expert judgement was used to disambiguate complexes with a high number of shared subunits, as well as to determine which labels to exclude to reduce crowding. Protein complexes that did not form tight clusters in the network were not annotated.

An additional high confidence network was generated for protein interaction predictions by selecting a network threshold at which 70% precision was achieved. BioGRID interactions that were not predicted by the algorithm (false negatives) were added to the network plot. The network was visualized using cytoscape[69]. Network node labels and annotations were added to the network manually. Both high-confidence and standard network interaction predictions are provided in Supplementary Table 7.

Curation of protein complex list (SNAP dataset). A curated protein complex list was seeded with complexes downloaded from the EBI complex portal version 19 July 2019 (ref. 38) and the EpiFactors database (obtained on 29 July 2019)[37]. Protein members of the complexes that were not detected in our experiments were filtered out. Only complexes with at least two protein subunits left after filtering were retained, merging protein complexes that became indistinguishable (that is, had the same subunits) after filtering. Protein complex annotations from the databases that were substantially similar (for example, variants of protein complexes defined by redundant adapter proteins) were merged together based on manual review. Missing annotations from the databases were added manually based on the review of the inferred protein network and corresponding literature. In some cases, the entries were also augmented with data from CORUM[70] and UniProt[71]. Where possible, protein complexes were renamed manually to match the canonical designations. All sources of annotations were recorded and are available in Supplementary Table 8.

Integration of MARCS with ChIP–seq data. For joint MARCS and ChIP–seq analysis, the relevant ENCODE[30] ChIP–seq, DNase-seq and ATAC–seq datasets for the K562 cell line were downloaded together with the chromatin state predictions from ROADMAP[1]. We next divided the hg38 reference genome, excluding blacklisted regions[72] and chromosome Y, into a set of non-overlapping 1,000-bp-wide bins and marked the bins containing peaks from each of the NGS datasets. We have assumed each of the genomic bins to be independent and identically distributed and therefore modelled the presence or absence of a given peak as a Bernoulli event. For a given pair of NGS datasets, we therefore computed their joint distribution by counting the bins for which both datasets are co-present, co-absent and mutually exclusive (both ways). A pseudocount of 100 was added to avoid zeroes and to smooth the probability estimates. This joint distribution enabled us to compute the MI between two NGS datasets, which is equivalent to the Kullback–Leibler divergence from the joint distribution under independence. To obtain an interpretable statistic that measures the fraction of information about A that can be predicted by knowing B, the MI was divided by the Shannon entropy (H) of one of the two datasets: $U(A,B) = \mathrm{MI}(A,B)/H(A)$. We frequently refer to this ratio as fraction of entropy of A explained by B or, simply, the normalized MI. As a convention, we use this to measure the fractional entropy of a protein (for example, PHF8) NGS experiment that the knowledge of a chromatin feature (such as H3K4me3) NGS experiment provides, for example, $U(\mathrm{PHF8, H3K4me3}) = \mathrm{MI}(\mathrm{PHF8, H3K4me3})/H(\mathrm{PHF8})$ (Extended Data Fig. 4a).

We next compared these normalized MI estimates for each of the MARCS-identified proteins for which ENCODE ChIP–seq data were available in K562 cells. For each of the MARCS chromatin features, and for each of the ChIP–seq chromatin features, we measured whether the proteins predicted to be strongly recruited or strongly excluded by MARCS feature had significantly higher or lower uncertainty coefficients, when compared to proteins neither strongly recruited nor strongly excluded, or proteins identified in MARCS for which we had no MARCS feature effect estimates at all. For these comparisons, we used a Mann–Whitney U-test (two-sided) and Benjamini–Hochberg correction. For the benefit of visualization we also computed the differences between mean $\log_2$-normalized MI estimates for MARCS-feature-associated proteins and others. In cases in which proteins had multiple ChIP–seq replicates, we used the harmonic average of their normalized MI coefficients for the analysis. We treated replicates of chromatin feature ChIP–seq analyses independently. In cases in which one ChIP–seq protein mapped to multiple MARCS proteins, we used the chromatin feature effect estimates from the proteins with the lowest P value.

As an additional similarity metric to the normalized MI statistic described above, we computed the Kendall correlation between the peak heights (as defined by the column 7 signalValue in the 'narrowPeak' and 'broadPeak' file formats) for genomic bins for which the peaks were co-present. This metric is used in Extended Data Fig. 4e–j.

For verification of the network analysis results in Extended Data Fig. 7f, we divided each pair of proteins for which ChIP–seq data were available into groups based on the confidence of inferred interactions from the MARCS-based network analysis. In the case of multiple mappings to MARCS, the highest-confidence outcome was chosen. For each such pairs, we computed the symmetric variant of normalized MI statistic: $U_{\mathrm{sym}}(A,B) = (2\mathrm{MI}(A,B))/(H(A) + H(B))$, based on their ChIP–seq datasets. The statistics of replicate experiments were averaged harmonically. A one-sided Mann–Whitney U-test was used to test whether the distribution of symmetric normalized MI coefficients is statistically different across the MARCS confidence levels (Bonferroni correction).

Statistics

The details of quantification and statistical analyses are described in detail in the Methods. Where appropriate, the necessary information is also described in the figure legend.

Reporting summary

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

Data availability

Gel raw data for the immunoblots shown in Fig. 5e and Extended Data Figs. 2b and 5g,h,j and a graph source data table providing the number of feature effect estimate measurements for the H3ac and H4ac features for each of the protein complexes displayed in the bar graph in Fig. 3d are provided in Supplementary Fig. 1. The MS data have been deposited at the ProteomeXchange Consortium via the PRIDE[73] partner repository (https://www.ebi.ac.uk/pride/) under the following identifiers: SILAC dinucleosome-purification experiments (PXD018966; the H4K20me2 samples from this experiment were previously deposited with identifier PXD009281 as part of ref. 55); H3K4me1 and H3K4me3 ChIP–MS (analysis of histone PTMs; PXD042224); H3K4me1 and H3K4me3 ChIP–MS (analysis of co-purified proteins; PXD042826); label-free dinucleosome-purification experiments with 200 bp SV40 promoter linker (PXD041835); label-free dinucleosome-purification experiments with 200 bp SV40 enhancer linker (PXD041443); label-free dinucleosome-purification experiments with short linkers and heterochromatic PTMs (PXD042368); IP–MS analysis of the human INO80 complex composition and interactome (PXD020712); ChIP–MS analysis of histone PTMs co-purified with the human INO80 complex (PXD042210); analysis of the effect of native chemical ligation on protein binding (PXD042390); MS analysis of ligated and recombinant human histones H3 and H4 (PXD020773); analysis of the stability of nucleosomal modifications during affinity purification in nuclear extract (PXD042823). Moreover, the SILAC nucleosome affinity purification data presented in this publication are available in an interactive format online (https://marcs.helmholtz-munich.de). The following public databases were used for data analyses in this study: BioGRID[34] (https://thebiogrid.org/); CORUM[70] (https://mips.helmholtz-muenchen.de/corum/); Complex portal[38] (https://www.ebi.ac.uk/complexportal/home); ENCODE[30] (https://www.encodeproject.org/); EpiFactors[37] (http://epifactors.autosome.ru/); Mygene.info[59] (https://mygene.info/); UniProt/Swiss-Prot[71] (https://www.uniprot.org). A detailed list of ENCODE datasets used for the integration of MARCS with ChIP–seq data, including ENCODE accession numbers, is provided in Supplementary Table 4. A list of key resources and reagents used in this study is provided in Supplementary Table 10 and the Supplementary Information.

Code availability

The source code developed for this study for data processing and analyses (https://github.com/lukauskas/publications-lukauskas-2024-marcs) and for the interactive web interface (https://github.com/lukauskas/marcs) are available at GitHub. Detailed information about software used in this manuscript is provided in the ‘key resources table’ in Supplementary Table 10 and the Supplementary Information.

51. Dyer, P. N. et al. Reconstitution of Nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* **375**, 23–44 (2003).
52. Tvardovskiy, A., Nguyen, N. & Bartke, T. Identifying specific protein interactors of nucleosomes carrying methylated histones using quantitative mass spectrometry. *Methods Mol. Biol.* **2529**, 327–403 (2022).
53. Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.* **327**, 85–96 (2003).
54. Dewari, P. S. et al. An efficient and scalable pipeline for epitope tagging in mammalian stem cells using Cas9 ribonucleoprotein. *eLife* **7**, e35069 (2018).
55. Nakamura, K. et al. H4K20me0 recognition by BRCA1–BARD1 directs homologous recombination to sister chromatids. *Nat. Cell Biol.* **21**, 311–318 (2019).
56. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
57. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
58. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
59. Xin, J. et al. High-performance web services for querying gene and variant annotation. *Genome Biol.* **17**, 91 (2016).
60. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
61. Fabre, B. et al. Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteom.* **4**, 82–86 (2014).
62. Bien, J. & Tibshirani, R. Hierarchical clustering with prototypes via minimax linkage. *J. Am. Stat. Assoc.* **106**, 1075–1084 (2012).
63. Franzin, A., Sambo, F. & Camillo, B. D. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btw807 (2016).
64. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
65. Ledoit, O. & Wolf, M. Honey, I shrunk the sample covariance matrix. *J. Portf. Manage.* **30**, 110–119 (2004).
66. Saito, T. & Rehmsmeier, M. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics* **33**, 145–147 (2016).
67. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. In *Proc. Third International AAAI Conference on Weblogs and Social Media* 361–362 (AAAI, 2009).
68. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
69. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
70. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2018).
71. Bateman, A. et al. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
72. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
73. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2018).
74. Schuettengruber, B., Bourbon, H.-M., Croce, L. D. & Cavalli, G. Genome regulation by polycomb and trithorax: 70 years and counting. *Cell* **171**, 34–57 (2017).
75. Filippakopoulos, P. et al. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **149**, 214–231 (2012).
76. Kuo, A. J. et al. The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier–Gorlin syndrome. *Nature* **484**, 115–119 (2012).
77. Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B. & Borggrefe, T. The histone variant H2A.Z in gene regulation. *Epigenet. Chromatin* **12**, 37 (2019).

Acknowledgements We thank the members of the Bartke group for input and discussions; T. Schalch for providing plasmids; M. Mann for the ACTR5–GFP HeLa cell line; and C. Wooding and P. Jansen for technical support. S.L. was supported by a BBSRC DTP PhD studentship award (1377205). M.S. was funded by the Helmholtz Gesellschaft through a Munich School for Data Science (MUDS) PhD Studentship. P.F., H.F. and A.P.S. were supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001999), the UK Medical Research Council (FC001999) and the Wellcome Trust (FC001999). Proteomics and MS research in the O.N.J. laboratory is supported by grants to the VILLUM Center for Bioanalytical Sciences (VILLUM Foundation grant no. 7292), PRO-MS: Danish National Mass Spectrometry Platform for Functional Proteomics (grant no. 5072-00007B) and INTEGRA (Novo Nordisk Foundation, grant no. NNF20OC0061575 to O.N.J.). The M.V. laboratory is part of the Oncode Institute, which is partly funded by the Dutch Cancer Society. R.S. was supported by the Deutsche Forschungsgemeinschaft (DFG) through SFB 1064 (project ID, 213249687) and SFB 1309 (project ID, 325871075), as well as the AmPro programme and the Helmholtz Gesellschaft. Work in the T.B. laboratory was funded by the UK Medical Research Council (grant no. MC_UP_1102/2), the European Research Council (ERC StG no. 309952), the Deutsche Forschungsgemeinschaft (DFG project IDs, 213249687/SFB 1064, 431163844 and 450084515) and the Helmholtz Gesellschaft. The MARCS website (https://marcs.helmholtz-munich.de) is hosted on the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Author contributions S.L., A.T., N.V.N. and T.B. designed the study and planned experiments and analyses. A.T. and N.V.N. performed experiments. S.L. performed computational analyses and designed the MARCS online resource. M.S. performed computational analyses. P.F., T.K.B. and T.R. performed MS measurements. B.Ö.A. and S.D. prepared reagents and performed quality control. K.B. helped with tissue culture experiments. S.L., A.T., P.F., H.F., R.G.H.L., T.K.B., T.R., S.M.H., O.N.J., M.V. and A.P.S. analysed MS data. S.L., N.V.N., A.T., R.S., C.L.M., P.A.D. and T.B. analysed and interpreted data. P.A.D. and T.B. supervised experiments and data analysis. T.B. coordinated the study. S.L., N.V.N., A.T. and T.B. wrote the manuscript with input from all of the authors.

Funding Open access funding provided by Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH).

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07141-5.
Correspondence and requests for materials should be addressed to Till Bartke.
Peer review information *Nature* thanks Michael Washburn, Shoshana Wodak and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 30.4.2024

Mara Stadler