Biomedical Machine Learning
Beyond the Training Distribution

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Giovanni Visonà

aus Valdagno / Italien

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	29.10.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Bernhard Schölkopf
2. Berichterstatter/-in:	Prof. Oliver Kohlbacher

Abstract

Machine learning (ML) holds the potential to impact many aspects of our lives, particularly in high-stakes areas like law, autonomous systems, and healthcare. The prospects of leveraging large quantities of data to mine patterns, improve decision-making, and navigate the complexity of biological systems are especially appealing and can have far-ranging consequences; however, ensuring the robustness and reliability of machine learning models has proven a remarkably difficult challenge, leading to considerable efforts by the research community.

In particular, understanding how ML models generalize to new observations is a necessary condition for the fruitful translation of these advancements in machine learning to clinical practice or to expand biological domain knowledge. When the training and test settings correspond, and the individual observations do not affect each other—the so-called independent, identically distributed (IID) setting—machine learning and deep learning have displayed remarkable capabilities. But when the data-generating distribution shifts, or when we want to solve related but slightly different tasks, then the quality of the predictions of a model can rapidly deteriorate.

In this thesis, I will examine the challenges that arise when generalizing beyond the training distribution in biomedical machine learning and the approaches developed to tackle such challenges. The first part of the thesis will provide a broad overview of the topic of generalization in machine learning, starting from a conceptual formulation of the generalization problem and the progress made in laying theoretical foundations for generalization in ML. Delving into the topic, I will provide an examination of the most common paradigms developed to improve predictive performance when generalizing outside the training distribution, and I will discuss the role of causal reasoning within this picture.

Afterwards, I will review the state of biomedical applications of machine learning, highlighting some of the most well-studied areas of research, as well as fields where

the use of ML has yet to deliver on its promise. Of particular interest is the topic of biases in biomedical data: given the staggering complexity of biological phenomena, and the considerable experimental constraints on gathering relevant data, it is crucial that we understand how to separate noise and natural variability from meaningful signal. Related to this idea, I will also discuss the ever-present challenge of validating the results of biomedical ML models.

Following these broad overviews of generalization and biomedical machine learning, I will present two works revolving around the application of deep learning to biological and clinical data. In each of them, the generalization challenges and paradigms presented in the earlier chapters play a crucial role, enabling novel prediction tasks or revealing insights into the properties of the models. The first work, that focuses on the task of imputing epigenomic signals, showcases how the use of transfer learning enables the out-of-distribution imputation of individual-specific epigenomic patterns, a case study in personalized epigenomics that is, to the best of my knowledge, the first of its kind. Afterwards, I will present a research project that tackles the task of predicting antimicrobial resistance from clinical proteomics data; when delving into the workings of the models proposed, the analysis of zero-shot prediction tasks offers a window into their robustness, which can guide future developments and offer insights for the data collection efforts required to progress further.

Zusammenfassung

Maschinelles Lernen (ML) hat das Potenzial, viele Aspekte unseres Lebens zu beeinflussen — insbesondere in kritischen Bereichen wie Recht, autonome Systeme und dem Gesundheitswesen. Die Aussicht, große Datenmengen nutzen zu können, um Muster zu erkennen, Entscheidungsprozesse zu verbessern und sich in der Komplexität biologischer Systeme zurechtzufinden, ist besonders reizvoll und kann weitreichende Folgen haben. Es hat sich allerdings auch als eine bemerkenswert schwierige Herausforderung erwiesen, die Robustheit und Zuverlässigkeit von ML-Modellen zu gewährleisten, was zu erheblichen Anstrengungen in der Forschung geführt hat.

Insbesondere das Verständnis, wie gut oder schlecht Modelle auf neue Beobachtungen verallgemeinern, ist eine notwendige Voraussetzung dafür, um die Fortschritte aus der ML-Forschung in die klinische Praxis zu übertragen oder um das biologische Fachwissens zu erweitern. Wenn die Trainings- und Testszenarien übereinstimmen und die einzelnen Beobachtungen sich nicht gegenseitig beeinflussen (das sogenannte independent, identically distributed (IID) Setting), haben maschinelles Lernen und Deep Learning bemerkenswerte Fähigkeiten gezeigt. Wenn sich jedoch die datenerzeugende Verteilung ändert, oder wenn wir verwandte, aber leicht unterschiedliche Aufgaben lösen wollen, kann die Qualität der Vorhersagen eines Modells schnell abnehmen.

In dieser Arbeit werde ich die Herausforderungen untersuchen, die beim Generalisieren über die Trainingsverteilung hinaus im biomedizinischen maschinellen Lernen auftreten, sowie die Ansätze besprechen, die entwickelt wurden, um diese Herausforderungen anzugehen. Der erste Teil der Arbeit wird einen breiten Überblick über das Thema Generalisierung im maschinellen Lernen geben, angefangen mit einer konzeptionellen Formulierung von “Generalisierung” und den Fortschritten bei der Schaffung theoretischer Grundlagen hierfür. Weiterhin werde ich die gängigsten Paradigmen vorstellen, die entwickelt wurden, um die Vorhersageleistung bei der Generalisierung über die Trainingsverteilung hinaus zu verbessern, und die Rolle des kausalen Denkens in diesem Zusammenhang diskutieren.

Anschließend werde ich einen Überblick geben über den aktuellen Stand der Anwendung von maschinellem Lernen in der Biomedizin und dabei einige der am intensivsten beforschten Bereiche hervorheben sowie Gebiete aufzeigen, in denen der Einsatz von ML sein versprochenes Potenzial noch nicht gänzlich eingelöst hat. Besonders interessant ist dabei das Thema Bias (Verzerrungseffekte) in biomedizinischen Daten: Angesichts der überwältigenden Komplexität biologischer Phänomene und der erheblichen experimentellen Einschränkungen beim Sammeln von relevanten Daten ist es entscheidend, dass wir verstehen, wie man Rauschen und die natürliche Variabilität von echten Signalen voneinander trennt. Damit verbunden werde ich auch die allgegenwärtigen Herausforderungen bei der Validierung der Ergebnisse biomedizinischer ML-Modelle diskutieren.

Nach diesem breitgefasstem Überblick über die Generalisierung und das biomedizinische maschinelle Lernen werde ich zwei Arbeiten vorstellen, die sich um die Anwendung von Deep Learning auf biologische und klinische Daten drehen. In beiden diesen Arbeiten spielen die in den früheren Kapiteln dargestellten Herausforderungen und Paradigmen in Bezug auf Generalisierung eine entscheidende Rolle, indem sie neue Vorhersagen ermöglichen oder tiefere Einblicke in die Eigenschaften der Modelle geben. Die erste Arbeit, die sich auf die Imputation (das heißt, das Ergänzen fehlender Daten) epigenomischer Signale konzentriert, zeigt, wie der Einsatz von Transfer Learning die Out-Of-Distribution-Imputation von individuellen epigenomischen Mustern ermöglicht — eine Fallstudie in der personalisierten Epigenomik, die meines Wissens die erste ihrer Art ist. Anschließend werde ich ein Forschungsprojekt vorstellen, das sich damit beschäftigt, antimikrobielle Resistenzen anhand klinischer Proteomikdaten vorherzusagen. Bei der Untersuchung der Funktionsweise der vorgeschlagenen Modelle bietet die Analyse von Zero-Shot-Vorhersagen einen Einblick in deren Robustheit. Dies kann zukünftige Entwicklungen leiten und Erkenntnisse liefern für die Erfassung zusätzlicher Daten, die für den weiteren Fortschritt erforderlich sind.

Acknowledgements

There are certain moments in your life when you reach a crossroad, a transition to new paths that can shape your future in a myriad of possible directions. These points of change can be scary, I know they are for me, but at the same time they are just perfect to pause for a minute and reflect on your journey up to now.

As I reach the end of my doctoral studies, I look back and feel overwhelmed by gratitude and appreciation for what I experienced, in a way that I rarely felt before. The past four and a half years did not lack in challenging moments, but they were also full of wonderful experiences and, most important of all, people that enriched my life and made all of this possible.

I want to first thank Bernhard for welcoming me to his lab, and allowing me to pursue my Ph.D. in a wonderful department that encouraged my growth both as a researcher and as a person. Here I found unwavering support and freedom in pursuing many research directions, as well as unbelievably talented colleagues and an environment that encouraged discussion and professional development. I could not have asked for a better place to be than your group.

A special thank you must go to Gabriele, who has worked tirelessly to support me every step along the way. Thanks to you, I took my first stumbling steps as a researcher four years ago; slowly but surely, I have found my stride, and I have learned so much since then. Your passion for all aspects revolving around machine learning and science, including ethics and a drive to really understand molecular biology, inspires me as a researcher.

I would also like to thank the members of my thesis committee, Oliver Kohlbacher and Nico Pfeifer, for their kind availability in accompanying me at the finish line of this project. I am elated that you can join me for the completion of my journey.

The MLFPM network was perhaps one of the most influential experiences that shaped my doctorate. In this project, I found talented and driven peers that motivated me to grow, mentors that provided insights and supported me along the way, and collaboration opportunities that widened my horizons. So thank you to my fellow ESRs Lucas, Emese, Giulia, Diane, Maguette, Pelin, Christopher, Bowen, Pradeep, Vesna, Kadri, Anastassia, and Rime. A heartfelt thank you to Katharina Heinrich and the organizers, as well as the PIs and partners involved.

The network also gave me the chance to collaborate directly with amazing scientists and researchers, so I want to express gratitude to Florence Demenais, Emmanuelle Bouzigon, Carl-Johan Ivarsson, Magnus Fontes, Michal Rosen-Zvi, and Karsten Borgwardt. Working with you has been a fantastic, formative experience that will continue to shape the researcher I aim to become.

Together with them, I had the absolute pleasure of collaborating with amazing researchers and professionals outside the institute; my work with Alex Hawkins-Hooker, Katharina Wenger-Alakmeh, and Carlos Oliver exposed me to novel ideas and challenged my views.

I could not have weathered my doctoral studies without close colleagues and friends who wholeheartedly listened to me and alongside whom I could grow. Tanmayee and Crhistian have made me feel part of a wonderful group, and my experience would not have been the same without you.

At the same time, I had the fortune of meeting so many talented and motivated researchers at MPI that were always ready to discuss ideas and share their work, and I hope the friendships that were born here will continue even when we all move to new endeavours. So I want to thank each of you that pushed me to be a better scientist, starting with my office mates Heiner, Lennart, Yassine, Hamza, Alex, and now Frederick, Jonas W., and Nasim. I will always treasure the experiences I made and the discussions I shared with all other researchers in our department, including (in no particular order) Simon, Erick, Maximilian D., Sergio, Timothy, Siyuan, Yucen, Zhijing, Armin, Annalena, Felix, Wendong, Maximilian M., Frederike, Hsiao-Ru, Junhyung, Amartya, Jonas K., Aaron, Julius and Partha. Additionally, I could always count on the support of the people who made the Empirical Inference department work, and I want to express my heartfelt gratitude to Sabrina, Ann-Sophie, Lidia, Annika, Vincent, and Sebastian.

Outside Tübingen I had the great fortune of holding friendships that kept up through my doctoral experience in another country. Many people would deserve a shoutout, but I have to thank in particular Francesco G. and Francesco F., who were always ready to support me when I needed it.

Finally, none of this would have been possible without my family, who always cared and cheered for me while I walked my path. My mum and dad, my sisters Cecilia and Margherita, and my brother Matteo, who always made sure to tell me how proud they are of me, and encouraged me at every step along the way.

Last but certainly not least, my greatest supporter, my partner, and confidant. Cecilia, without you, I would not be the person I am today. With you walking along my side, I have grown and learned, I have challenged myself and became better for it. I hope I will always continue to make you proud.

Declaration

The main research results of this thesis are composed of two self-contained chapters based on the following publications:

Chapter 4
Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning.
Hawkins-Hooker A.*, **Visonà G.***, Narendra T., Rojas-Carulla M., Schölkopf B., Schweikert G.
Nature Communications (2023)
*equal contribution

Chapter 5
Multimodal learning in clinical proteomics: enhancing antimicrobial resistance prediction models with chemical information.
Visonà G.*, Duroux D.*, Miranda L., Sükei E., Li Y., Borgwardt K., Oliver C.
Bioinformatics (2023)
*equal contribution

Every chapter provides the required background material and relevant literature. Supplementary materials can be found in the corresponding appendices.

In the course of my Ph.D. studies, I have contributed to the following publications, the material for which is not directly included in this thesis:

Network propagation for GWAS analysis: a practical guide to leveraging molecular networks for disease gene discovery.
Visonà G., Bouzigon E., Demenais F.*, Schweikert G.*
Briefings in Bioinformatics (2024)
*shared supervision

A historical perspective of biomedical explainable AI research.
Malinverno L., Barros V., Ghisoni F., **Visonà G.**, Kern R., Nickel P. J., and Others
Patterns (2023)

Machine-learning-aided prediction of brain metastases development in non-small-cell lung cancers.
Visonà G., Spiller L. M., Hahn S., Hattingen E., Vogl T. J., Schweikert G., and Others
Clinical Lung Cancer (2023)

Giovanni Visonà
2024

Table of contents

Chapter 1

Introduction

> The essence of knowledge is generalisation. That rubbing wood in a certain way can produce fire is a knowledge derived by generalisation from individual experiences; the statement means that rubbing wood in this way will always produce fire. The art of discovery is therefore the art of correct generalisation. [...] The separation of relevant from irrelevant factors is the beginning of knowledge.
>
> *The Rise of Scientific Philosophy*
> Hans Reichenbach

What does it mean to generalize? The concept of generalization is so integral to our understanding of intelligence and cognition that it is challenging to provide a single answer encompassing all the exquisite complexity hidden behind this question. We commonly use the term to describe the setting in which previously acquired knowledge applies to novel situations and inputs. We human beings are adept at generalizing knowledge in many ways, and we excel at learning abstractions and generalizations from examples: a child will learn what a tree is not by receiving a complex and nuanced description, but by repeatedly pointing at trees and asking "what is that?" or "is that also a tree?"

Generalization is an exceedingly broad task that has shaped our inquiries into the nature of intelligence for as long as humans have wondered about the world, and we learned to recognize its role in different settings. From an evolutionary perspective, for example, the capacity to learn information and behaviours that can be used to face a

multitude of situations is such a powerful survival tool that it has become inextricably linked with our definition of "higher intelligence." We consider an organism "intelligent" when it can use its limited experience to adapt to novel situations through the use of strategies, tools, interactions, and knowledge; the more abstraction required in the process, the higher the level of intelligence we attribute to the organism. Similarly, it is no coincidence that the highest ambition of artificial intelligence (AI) research is the creation of artificial general intelligence, highlighting how intrinsic the capability of generalization is to the notion of intelligence.

For a long time, top-down deductive reasoning was considered the more rigorous approach to generalizing knowledge: starting from known principles, a robust chain of reasoning leads to deriving new conclusions which must necessarily hold true as well [1]. With the advent of the Renaissance and the rise of empiricism, a growing acceptance of the imperfect nature of our observations of the world led philosophers to emphasize inductive reasoning more [2]. This bottom-up approach attempts to derive abstract principles from finite observations of reality; if this process of abstraction has captured some "truth" of the world, then these derived principles can then be applied to novel settings, thus generalizing from an imperfect set of observations.

These two forms of reasoning are core to the formalization and diffusion of the scientific method, the crowning achievement of systematic reasoning. The conjoined nature of inductive and deductive reasoning has been the object of debate since Aristotle [3], and historical trends in the development of a system of reasoning have leaned towards one or the other without ever being able to fully separate them. Scholars such as John Stuart Mill focused firmly on inductive reasoning [4], while following conceptions formalized by Karl Popper pushed for a larger role of deductive reasoning [5]. Over time, the scientific method replaced the certainty of logical reasoning with the uncertain guarantees provided by repeated attempts to falsify the theory tested. Crucially, with this paradigm of learning, the complete certainty in the truth of a hypothesis is no more than a mirage: only its falsehood or incompleteness have a chance of being fully determined. This newfound humility is even reflected in the naming conventions developed over centuries, moving from absolute laws (Newton's law of gravitation), to tested theories (Einstein's theory of general relativity, Darwin's theory of evolution). The latter, even though supported by ample amounts of evidence, can never be labelled as complete descriptions of the world.

The tools to systematize the level of certainty in scientific theories were developed from the 18th century onward in the field of statistics. Starting from the collection of

demographic and economic data by states—hence the etymology of the term—the field evolved to offer refined mathematical tools that promised to quantify the "correctness" of a model or a theory. Unable to establish an objective truth judgment of a theory, we moved to p-values, effect sizes, confidence intervals in the so-called frequentist approaches, and to posterior probabilities, Bayes factors, and model comparisons in Bayesian formalisms.

This much more fluid conceptualization of correctness proved to be a valuable tool in the late 20th century, when the advent of modern computers enabled the modelling of observed phenomena with unprecedented richness. The paradigm of inductive learning was systematized and automated in the field of machine learning (ML), wherein computer programs and mathematical models learn from observed data. It is here that measuring how well a model generalizes to new inputs offers a proxy measure of how much it has learned the true information underlying the phenomenon studied. Conversely, understanding the factors that hinder this generalization—and how to address them—is crucial for the robust application of machine learning in practice.

The last two decades have seen a veritable explosion in the development and deployment of machine learning in real-world applications. From recommender systems to diagnostic models, from image generation to speech synthesis, this revolution has been powered by an unprecedented increase in availability of data and computational power [6]. Due to its ubiquity, machine learning is affecting the lives of many people, with a large portion of them not even realizing the spread of these technologies in our everyday life [7]. While in some cases the reliability of these computational systems is not the primary concern, machine learning is increasingly adopted to improve decisions and processes in high-stakes settings. ML models are used to guide self-driving cars [8], predict crime [9], and solve cybersecurity issues [10], to name a few.

In such situations, the failure of a model can inflict serious harm, so it becomes more important than ever to determine how we can know that the model works as intended and evaluate its safety [11]. Unfortunately, there is no silver bullet, and many factors can influence this robustness. To be able to trust and rely on the predictions of a model, we need to identify the subpopulations or samples on which it performs well, and to pinpoint and manage appropriately those on which it performs poorly [12, 13]. The way to achieve this is to design and conduct robust validation procedures for a model, or to focus specifically on the interpretability of such models [14].

Philosophically, this demand for reliability can be linked to the difference between learning and intelligence. This distinction, which in humans can be blurry [15], is especially relevant for machine learning systems as they model correlations from observational data rather than building an understanding of the world. These models learn from data, but this process does not automatically lead to what we would consider an intelligent system; this limitation makes them generally unsuitable to go beyond the observational setting and consider interventions [16], which is often what we desire in practice. Ultimately, we develop machine learning models as a mean to an end: to improve our decision-making, prioritize our limited resources, and more generally bring to fruition a desired outcome.

Among high-stakes settings, the biomedical applications of machine learning are perhaps the most crucial field in which robustness is required. With the label of "biomedical" applications I indicate both those models that examine the processes involved in molecular biology and human health, as well as those systems that aim to improve clinical practice and medical care. In biomedicine, failures in predictive capabilities of ML models can lead to expensive and time-consuming errors—such as selecting the wrong candidate drugs for clinical trials—and can culminate in outright inflicting harm by recommending unsuitable treatments or misdiagnosing a disease [17], or unfairly determining who needs extra clinical care [18, 19].

These challenges are not just technical hurdles, but also social ones. The field of statistics as a whole has been developed and coopted by systematic efforts to ensure unequal structures of power, with several founding fathers of the field like Karl Pearson and R. A. Fischer having clear and deplorable views on the role of this tool for eugenics [20, 21]. Mechanisms of oppression for underrepresented populations are just as relevant for machine learning, with the added complication of the opacity and inscrutability of many ML models; such concerns are not purely speculative, and have already been shown to disproportionately affect certain sections of the population [18]. Facing this challenge will require concerted social and political efforts that go well beyond technical improvements to the robustness of ML models; ultimately, the intersection of machine learning and healthcare is a multifaceted effort, and we will need a more in-depth understanding of all its complexities to produce meaningful progress.

Despite these risks and challenges, ML also opens up unprecedented opportunities to enhance the biomedical field; it is the core component behind the development of the field of precision medicine, an alternative paradigm of care that aims to stratify patients into small groups—or even at the level of the individual—to provide more effective

personalized medical care [22]. In addition to clinical care, ML offers compelling possibilities for mining large-scale datasets and inferring patterns and connections that can progress our understanding of biology. This latter prospect is also appealing on a conceptual level: many facets of molecular biology are controlled by pattern recognition mechanisms, and training a model to learn such patterns has the potential to help us decode the complexity of biological systems and identify potential biomarkers or targets for intervention.

Given the potential of ML applied to biomedicine, the exploration and systematization of the generalization properties exhibited by these models is a crucial, high-impact area of research. The successful translation of research results to novel insights and improved clinical practices will require enormous collaborative efforts so that we can ensure equitable outcomes and lead to transformative changes for the benefit of all.

This thesis will present a structured discussion of the technical aspects of the challenges described, where I will analyse the role of predictions beyond the training distribution in biomedical machine learning. I will first examine the concept of generalization in machine learning as a whole, the formalisms developed to understand it, and the approaches developed to improve generalization performance. Afterwards, I will focus on the biomedical field, discussing the specific biases and systemic issues that affect the robust application of machine learning. Finally, diving into the opportunities of precision medicine, I will explore two research projects that tackle generalization tasks in the biomedical setting and that well represent the challenge of producing accurate predictions when generalizing outside the training distribution.

1.1 Outline

Aside from this brief introduction, this thesis includes 5 main parts:

- Chapter 2 contains a discussion of the forms of generalization in machine learning, the paradigms developed to improve this aspect of ML models, and their relevance in the biomedical field. I will present an overview of the most common methodologies to increase robustness, selected based on their role in Chapters 4 and 5, or because of their broader role in biomedicine. I will additionally discuss the progress in laying theoretical foundations of generalization, as well as the role of causality in shaping machine learning towards robust and reliable systems.

- Chapter 3 is a review of the current state of biomedical ML and the specific challenges encountered when applying ML to the biosciences, such as the sources of bias in biomedical data or systemic issues like validating the predictions of biomedical ML models.
- Chapter 4 describes an application of deep learning for imputing the incomplete human epigenome. This work includes a case study in personalized epigenomics, demonstrating how the use of transfer learning enables the imputation of individual-specific epigenetic signals. This chapter is based on [23].
- Chapter 5 presents a research work that employs deep neural networks to combine clinical proteomics and chemical features for predicting antimicrobial resistance outcomes and recommending effective drugs. This work shows how different data-generating processes affect our predictive models, and examines several zero-shot prediction tasks to gain more in-depth insights into the workings of these predictors. This chapter is based on [24].
- Chapter 6 is a discussion of the results presented, including how the generalization paradigms described in the introductory chapters relate to the two applied projects. I will describe the limitations of the two studies presented, and present some considerations on possible future directions.

Chapter 2

Generalization in Machine Learning

2.1 Generalizing to new data

In applied machine learning, we are generally presented with a set of data points on which to train a model (i.e., fit a function), which will then be used on new sets of observations. This model is commonly supposed to serve a specific aim, and not to be the final result of our analysis; ultimately, what we are truly interested in is improving our decision-making, supporting interactions of automated systems with their environment, or detecting meaningful patterns in the data.

The main challenge keeping us from realizing the full potential of applied machine learning is that it relies on the (generally implicit) assumption that the training and test data come from the same distribution. In the so-called IID (independent, identically distributed) regime, ML models have shown high performance and robust predictive capabilities. Under the IID assumption, we can obtain an optimal model using frameworks such as empirical risk minimization (ERM), with robust guarantees that its predictions will successfully generalize to the test set [25, 26]. However, when the predictive setting diverges from the training conditions, the performance of ML models can rapidly deteriorate [27].

This issue is especially relevant in applied biomedical machine learning, as having test data from the same distribution as the training data is the exception rather than the norm [28], especially when combining datasets collected from several laboratories and institutions [29]. Factors like shifts over time, biased data collection, data processing

procedures all impact the applicability of ML models in practice (see Section 3.2 for an in-depth discussion of the issue).

It is thus clear that the capability to generalize is perhaps the most crucial property for a machine learning model, as it concerns its effectiveness in practical scenarios beyond controlled experiments and training environments. It should not come as a surprise, then, that this research field is flourishing with varied contributions, attempting to paint a more complete picture of the practical and theoretical aspects involved. Many works in the literature focus on the development of specific architectures and techniques that make models more robust to shifts, while other researchers attempt to lay the theoretical foundations to understand the generalization mechanisms in learning algorithms. The last two decades have seen notable advancements on both fronts, which will be described in this chapter.

2.2 Generalization in the IID setting

To set up the appropriate context and to properly compare the IID setting to the challenge of generalizing outside the training distribution, let us first define what generalization is and provide adequate notation to support our discussion. While describing every possible task for machine learning with a single formalism can lead to exceedingly abstract representations, we can look at generalization through the lens of a specific problem formulation that is common in many real-world scenarios, and therefore suitable for our discussion.

In many applied settings, we gather observations $x \in \mathcal{X}$, where $\mathcal{X}$ represents any space, and we use them to make predictions. For example, the space of covariates $\mathcal{X}$ could be a Euclidean vector space $\mathbb{R}^d$, or the set of all 256×256 images, or the set of texts written in English. The data-generating process of these observations can be described by the distribution $P(X)$, where X is the random variable representing the observations x. Each observation x is associated to some output $y \in \mathcal{Y}$, which can be an explicit output in the case of supervised machine learning (e.g., labels associated to images), or an implicit quantity to determine in the case of unsupervised learning (e.g., cluster assignments). These outputs are generally described by a probability distribution $P(Y)$; however, what we are truly interested in is the joint dependency of y and the covariates x, described by the joint distribution $P(X, Y)$, or alternatively the conditional distribution $P(Y|X) = P(X, Y)/P(X)$.

Machine learning is the automated process of modelling this relationship and inferring its characteristics from data. The approximation of the joint distribution $P(X, Y)$ is the subject of generative modelling, where paradigms such as variational autoencoders, normalizing flows, and diffusion models offer the flexibility to learn these complex probabilistic relationships and enable tasks such as sampling new observations [30].

Alternatively, the modelling of the conditional relationship $P(Y|X)$ is the object of fields such as supervised learning. In this setting, we have access to a set of training instances $S_{\text{train}} = \{(x_1, y_1), ..., (x_n, y_n)\}_{\text{train}}$, and we want to obtain a predictive model that can produce accurate approximations of the outputs $y_1, ..., y_n$ by using as input the covariates $x_1, ..., x_n$. A common assumption adopted in this application of machine learning models is that the stochastic dependency of the outputs on the covariates can be split into two components, a deterministic part and external noise. This decomposition can be expressed as follows:

$$P(Y|X = x) = f(x) + \epsilon \tag{2.1}$$

where ϵ can model random stochastic fluctuations (homoskedastic noise, $\epsilon \perp\!\!\!\perp x$), but also include unaccounted confounders that depend on the covariates (heteroskedastic noise, $\epsilon \not\perp\!\!\!\perp x$). In common use cases of machine learning—particularly in deep learning—we model the deterministic link f using a parametric model $f_\theta : \mathcal{X} \to \mathcal{Y}$, where the θ represents the parameters of the model.

Intuitively, we can say that a model that was trained on a set of observations S_{train} generalizes well if it produces accurate predictions even when employed for a new set of observations $S_{\text{test}} = \{x_1, ..., x_m\}_{\text{test}}$. But what does it mean to "perform well" on this test data? How can we formalize this concept? Often, the generalization capability of a model is discussed within the formulation of the generalization error. Let us consider the previously described supervised learning setting, where we want to predict the label y given the covariates x. To quantify the correctness of a prediction $\hat{y} = f_\theta(x)$ produced by the model f_θ, we define a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, which produces low values if the prediction $\hat{y}$ is similar to y, and higher values otherwise. The population risk is then the expected value of $\mathcal{L}$ under the data-generating distribution:

$$R[f_\theta] = \mathbb{E}_{x,y \sim P(X,Y)}[\mathcal{L}(y, \hat{y})] \tag{2.2}$$

Since in practice we only have access to a finite sample S, such as our previously defined training set $\{(x_1, y_1), ..., (x_n, y_n)\}_{\text{train}}$, we use the empirical risk as estimator:

$$R_S[f_\theta] = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) \tag{2.3}$$

We can define the generalization error—or sometimes generalization gap—as the difference between the population risk and the empirical risk, i.e. $R[f_\theta] - R_S[f_\theta]$; we can then claim that a model generalizes well if the generalization error is low, ideally going to zero as the number of training samples grows sufficiently large. We usually do not have access to the population risk $R[f_\theta]$, so different approaches have been developed to make use of this concept. Theoretical formulations often seek to determine bounds for the generalization error as just defined, eschewing the need for explicit estimates of the population risk.

Empirical approaches, on the other hand, commonly rely on evaluations of the model performance on a hold-out set of data. For this applied setting, we can calculate the empirical generalization error $R_{S_{\text{test}}}[f_\theta] - R_{S_{\text{train}}}[f_\theta]$, which acts as an estimator for the generalization error.

At its heart, the problem of minimizing the generalization error in the IID setting revolves around the necessity of providing a functional approximation that fits the information contained in the data (i.e., the link f), without fitting the noise that inevitably affects finite samples gathered in an experimental setting. This challenge is captured well in the so-called bias-variance tradeoff, which classically describes the necessity of hitting a "sweet spot" in the flexibility of the model to achieve this balance [31].

When training a model, its bias represents the systematic errors between the expected predictions and the true values. This bias stems from the simplifying assumptions made when modelling complex real-world phenomena (i.e., the inductive biases), which are a core component of robust applications of ML. A model with high bias is prone to underfitting, and might not possess the necessary expressive power to model complex functional relationships; however, model biases allow us to restrict the hypothesis space available to the learning algorithm, reducing the amount of data needed to obtain a robust predictor [32]. The variance of a model, instead, refers to the error introduced by the model's sensitivity to small fluctuations in the training set. High variance is typical of models with a great capacity to represent complex functions, that can learn

to fit not just the functional dependence between observables and targets, but also the noise of the measurements, in the phenomenon called overfitting.

Both high variance and high bias can lead a model to generalize poorly to new data, and we would ideally minimize them both; in practice, however, there appears to be a tradeoff between the two, so that a model with low bias will tend to have high variance, and vice versa.

Counterintuitively, the selection of biased estimators for the function we want to fit may outperform unbiased estimators. One of the most bizarre phenomena that demonstrates this possibility is given by Stein's paradox and the related Stein-James estimator; this setting shows that, in estimating the means of multiple Gaussian distributions, using a combined shrinkage estimator can yield more accurate results than estimating each mean separately, even if the distributions are independent [33, 34].

This observation has profound implications for applied machine learning as a whole. Several widespread techniques that are ubiquitous in practice have their roots in an advantageous exchange where an increase in bias is counterbalanced by a greater reduction in variance. Among these methods, we find regularization approaches such as Tikhonov regularization and lasso regularization [35], ensemble techniques like bagging and boosting [36], and (arguably) also Bayesian modelling.

2.3 Theoretical foundations of IID generalization

2.3.1 The challenge of systematizing generalization

Machine learning as a whole is a field that has found its modern-day success after embracing its deeply experimental nature [37–39]. Many of the seminal advances in the field of learning algorithms were conceived to mimic biological intelligence and were not founded on a strong theoretical basis. The function of the neuron, for example, served as inspiration for the formalism of the McCullogh-Pitts neuron [40], which led to the creation of the Perceptron [40, 41], and ultimately to the development of artificial neural networks and deep learning. Similarly, the concepts underlying Convolutional Neural Networks—such as receptive fields—are inspired by earlier work on the stimulation of the visual cortex in cats [42].

For many developments in machine learning, first comes empirical observation and the development of algorithms that improve learning and predictive performance, and only afterwards are theoretical justifications researched. To cite an example relevant to

the work presented in Chapter 4, the core ideas at the base of transfer learning (TL) can be traced back as far as the 1970s [43], and were popularized in the early 1990s [44]. However, the first theoretical foundations for TL were laid in the years following the initial implementations [45], with more considerable focus from the research community in the last decade [46–49].

Finding a complete and cohesive grounding of machine learning is an extremely challenging task. This *post hoc* derivation of theoretical justifications often leads to significantly different theoretical analyses for the same approach. Consider, for an example relevant to the ResMLP model in Chapter 5, batch normalization (BN), a technique that improves the training of neural networks by normalizing layer inputs across the elements in a mini-batch. The research paper that proposed the method justifies its effectiveness by claiming that it works by addressing internal covariate shift [50]; however, follow-up works have argued that internal covariance does not explain the success of BN [51], or that its value lays in enabling the phenomenon of "super-convergence" [52], or that its primary role is that of a regularizer that smooths out the loss landscape [53, 54]. All of these descriptions offer information on different facets of this methodology; however, a complete theoretical description is still not available.

Despite the daunting effort required, the systematization of machine learning is a crucial research field that attempts to bring order to the complex interactions between mathematical formalisms, computer science, statistics, and engineering that converge in machine learning. Naturally, a large portion of these efforts extend to the understanding of the generalization capabilities of ML models. I will present here a brief overview of the main threads that have guided the systematization of generalization in machine learning, and examine why these efforts are not always applicable to deep learning.

2.3.2 Statistical learning theory and computational learning theory

Foundational approaches to establishing a theoretical basis of generalization in ML lay in the work of Vapnik, who provided comprehensive explorations of statistical learning theory (SLT) and its applicability to real-world problems [55]. This theory focuses on the statistical properties of learning algorithms and is complemented by the field of computational learning theory (CLT) [56], that describes the computational aspects of learning algorithms. These two branches of machine learning and statistics exhibit a

large degree of overlap, and the formalisms developed in both tend to come into play synergistically to paint a richer picture of learning algorithms.

One of the core concepts established in SLT is the Vapnik-Chervonenkis (VC) dimension [26, 57], a criterion that describes representational capacity of learning algorithms based on their ability to perfectly classify all possible labellings of a given set of points. The VC dimension serves as a foundation to demonstrate the validity of empirical risk minimization as a learning rule, and can be used to analyse the generalization properties in the IID setting. In general terms, SLT and CLT show that it is desirable to select an algorithm with a sufficiently broad hypothesis space to contain the true function to be fit, while limiting the complexity as much as possible. This modern statistical form of Occam's Razor [58] is reflected in frameworks such as structural risk minimization [55], but also in statistical measures like the Akaike information criterion [59] and the Bayesian information criterion [60], which directly encapsulate this tradeoff between complexity and goodness of fit.

The concept of VC dimension is an appealing tool to describe the expressive power of a class of models; however, precise calculations of the VC dimension for complex models are far from trivial, and it is often necessary to rely on upper and lower bounds instead [61]. More crucially, the VC dimension does not properly account for the inductive biases in the learning algorithms, nor the data distribution; both of these factors are known to be important for the generalization properties of deep neural networks [62], which limits the practical usefulness of the VC dimension in deep learning. For this reason, the complexity of deep neural networks may be better described by data-dependent criteria like Rademacher complexity [63].

The VC dimension has nonetheless led to many insights into the generalization properties of learning algorithms and their sample complexity. This formalism is central to the derivation of bounds on the generalization error within probably approximately correct (PAC) learning theory [64], a framework that offers statistical guarantees on the learnability of functions from data within a probabilistic margin of error. Most crucially, the PAC formalism examines this learnability in a manner that is independent of any specific data distribution, thus deriving general properties of the learning algorithm; as a consequence, PAC learning aims to bound the generalization error uniformly for all distributions.

Some works have related these generalization bounds to the complexity of the model class [65], following the spirit of the previously presented information criteria. Other

approaches have framed the upper bound as a function of algorithmic stability, which is a description of how much a change in the training data affects the output [66, 67]. Both these formulations bear strong connection to the bias-variance tradeoff: complex models tend to have lower bias, as they can describe richer classes of functions, but are generally less stable. Alternative theoretical descriptions have attempted to overcome the limitations of traditional PAC learning theory and focused on the concept of model compression, highlighting how optimizers are likely to focus only on a small subset of the possible model classes, thus enabling learning [68–71].

PAC learning theory offers many appealing ideas for laying a robust foundation of generalization in ML. However, just like the concept of VC dimension, PAC bounds for the generalization error do not account for the data distribution, which makes it difficult to extend them to deep learning. Some works have attempted to use PAC learning to tackle auxiliary tasks such as estimating confidence sets [72], but the PAC framework in its original formulation is not suitable to provide a complete formalization of deep learning.

2.4 Generalization in deep learning

2.4.1 The surprising generalization properties of deep neural networks

When considering the generalization performance of machine learning models, the case of deep learning (i.e., deep neural networks) appears especially puzzling when compared to simpler statistical models, as neural networks display excellent generalization performance in many settings [73]. This holds true even though there are many factors that should theoretically hinder neural networks [74], such as the high dimensionality of the inputs leading to the set of problems colloquially called the curse of dimensionality [75], and the highly non-convex loss landscapes that give no guarantee of the system converging to a good local minimum [76].

Most perplexing, perhaps, is the fact that even overparameterized neural networks can learn to generalize well [77]. Based on the established theories in statistics, increasing the number of learnable parameters beyond a certain point should lead to severe overfitting [78], and analysis of the properties of neural networks in their earlier iterations seemed to support this view [79]. However, many applications have

empirically demonstrated the contrary, with certain state-of-the-art models like GPT-4 [80] depending on a truly impressive number of learnable parameters.

Counter-intuitively, raising the number of parameters far beyond the interpolation regime can lead to a considerable improvement of the generalization performance, a phenomenon dubbed double descent [81]. This occurrence seems to break the classical bias-variance tradeoff [82], leaving us puzzled regarding how to reconcile empirical results with our theoretical understanding. To delve into the workings of double descent, some works in the literature have examined this phenomenon through the properties of the Hessian at the optimum [83], of gradient descent [84], and of certain forms of regularization [85].

While the precise mechanisms of this surprising phenomenon are not yet fully understood, this property of neural networks is one of the factors that led to a new paradigm in machine learning, the use of foundation models. Foundation models are large-scale neural networks pre-trained on vast quantities of data, which can then serve as foundations for fine-tuning to specific tasks [86]. This novel approach presents considerable challenges, but it has the potential to transform certain fields of application for machine learning by supplementing insufficient quantities of data with features and relationships learned from large-scale settings. Robust applications of this methodology depend on properly understanding the influence that shifts between the larger training datasets and the small application setting have on the foundation model; this deviation from the training distribution is the subject of Section 2.5.

It is clear, then, that unravelling the mysteries of generalization in deep learning—especially outside the training distribution—is a vital research effort, crucial to guaranteeing the robust application of large-scale models in a wide variety of settings. In particular, many sectors of the biomedical sciences would massively benefit from robust methodologies to transfer knowledge and combine datasets, as the field is often hindered by cost and resource constraints, biases in the data, and challenges in validating the predictions of a model.

2.4.2 The role of machine learning architectures in enhancing generalization

Advancements in our understanding of generalization rely on the systematization of its theoretical basis and the development of techniques such as regularization. However,

one of the most interesting aspects that have driven the progress in machine learning is the creation of innovative architectures that are particularly suited for specific settings.

In certain tasks, machine learning languished until the introduction and widespread adoption of novel paradigms that encode suitable inductive biases for that specific application. A most emblematic example is found in the task of image classification: for many years, models struggled to achieve meaningful performance, and it was only with the introduction of AlexNet [87], a model based on the convolutional neural network paradigm proposed three decades earlier [88], that we saw a considerable leap forward.

Interestingly, while some of these methodologies are first developed to solve specific tasks, they often find further success in applications that differ from the original conception. This is also the case for two such advancements that are used in the works presented in Chapters 4 and 5, the Transformer and the Residual Network, respectively.

The Transformer architecture was first introduced in 2017 by Vaswani et al. [89] to implement a form of attention that would be suitable for sequence-to-sequence models. Attention mechanisms attempt to mimic the selective focus that humans place on stimuli from the external world, which allow us to prioritize cognitive resources and assign different importance to certain signals. Early approaches to implementing attention had found moderate levels of success a few years prior [90], however the introduction of the Transformer led to a qualitative jump in the performance of machine learning on natural language tasks.

This architecture, in its original conception, implements the so-called scaled dot product attention. In this formalism, vectorial representations of the inputs are linearly projected into three separate vector spaces, and named queries, keys, and values. The first two are used to describe relationships between elements of the input: a query is compared to each key using a dot product operation, to obtain an attention weight that describes their relatedness. Afterwards, these attention weights are used to produce a weighted sum of the values corresponding to each element, thus obtaining the output representation of the elements of the input.

The Transformer has been developed with the aim of improving natural language modelling, where it has proven remarkably successful [80, 91]; however, it has now shown positive results also in computer vision [92], time series analysis [93], and speech recognition [94]. The eDICE model presented in Chapter 4 makes use of a variant of the Transformer (derived from the Set Transformer [95]) to impute unperformed epigenomic assays; the flexibility of this attention mechanism enables the model to

learn relationships between epigenetic modifications and between cell types to better predict the unobserved signals.

Chapter 5 introduces the ResMLP model, a neural network with residual connections that we employ to predict resistance to antimicrobial compounds. Residual connections, or skip connections, are architectural features in neural networks that enable the bypassing of one or more layers by directly feeding the input of a previous layer to a subsequent layer. The use of skip connections, introduced in 2015 by He et al. [96], improves the properties of the loss landscape [97], allows models to learn multiscale features [98], alleviates the issue of vanishing or exploding gradients [99], and generally enables the training of very deep models [100]. The result of all these appealing properties is that skip connections lead to a more robust training procedure and improve generalization performance more broadly [101, 102]. In our work, the use of a residual neural network enhanced the training and performance of the predictive model for the antimicrobial resistance of pathogenic samples.

2.4.3 Training dynamics in deep learning

Section 2.3.2 described several approaches to understand generalization in machine learning, as well as their limitations concerning the analysis of deep learning models. Given their widespread use, a more in-depth understanding of deep neural networks is crucial for innovative applications of these models. As frameworks like PAC learning proved to be unsuitable for describing the properties of neural networks, researchers have focused considerable efforts into examining the setting in which the inductive biases and data distributions interact more directly: the training procedure of deep learning models.

Deep learning is a vast field that includes a variety of paradigms for fitting experimental data and modelling data distributions. There are several possible strategies to train a model for a predictive task; however, the most common by far is the use of gradient descent with backpropagation [103]. In this approach, the gradient of the loss function is calculated with respect to all the tunable parameters of the models, which are then slightly modified in the direction opposite the gradient. We can view the loss function as a surface in the $(p+1)$-dimensional space defined by the loss as a function of the p parameters of the model, and a specific configuration of a model as a point on this surface; the gradient descent procedure moves the model configuration downward on this landscape to reach the lowest height possible from its starting point, one step at a time.

This paradigm is pervasive in many applications of machine learning, and as a natural consequence, considerable research efforts have been dedicated to understanding the properties of this learning procedure and how they relate to generalization performance.

The most direct analysis involves the characteristics of the loss landscape itself. Simple statistical models (e.g., a linear regression model with a mean squared error loss) display a convex loss surface, which is a well-studied setting with established solutions [104]. Complex non-linear models such as deep neural networks, on the other hand, exhibit a highly non-convex landscape [76], leading to the necessity of methodologies such as gradient descent that can find local minima of the loss function. As these loss landscapes lack the theoretical guarantees of convex losses, we can investigate what influences the performance of the model. In what conditions does gradient descent lead to "good enough" local minima? Are there characteristics that distinguish good minima from bad minima?

For example, it has long been speculated that the shape and curvature of the loss landscape has strong connections to the generalization performance; particular attention has been dedicated to "flat minima," i.e. local minima that are not very sensitive to small perturbations of the parameters [105–107]. Indeed, more rigorous analyses based on local Gibbs distributions [108] or PAC Bayes bounds [109] support this hypothesis, although these works often rely on unrealistic assumptions.

Other works have looked at the learning process itself, to relate the dynamics of descending the loss landscape to the generalization performance. A rich line of works has examined the stability of the training procedure [110–112], noting how optimization procedures that produce similar results under small variations of the data generalize better.

The analysis of the loss optimization procedure can rely on diverse mathematical formalisms such as random matrix theory, which has been used to explain how the training process of deep neural networks implicitly implements a form of self-regularization [113], influencing the dynamics of learning in a way that protects against overfitting even in high-dimensional regimes [114]. The random initialization of large neural network appears to be a crucial factor in explaining the performance of overparameterized networks, allowing the derivation of generalization guarantees in certain settings [115].

Alternative approaches have sought to link the dynamics of gradient descent to statistical mechanics, enabling the analysis of more complex qualitative properties of the generalization performance of networks [116]. This approach reconciles empirical

results with theoretical predictions by incorporating factors such as effective data load and temperature interpretations.

Perhaps one of the most popular theoretical frameworks to analyse the training dynamics of neural networks and their generalization properties is the neural tangent kernel (NTK) [117]. The NTK provides insights into how neural networks learn and generalize by connecting them to kernel methods, which are well-studied in statistical learning theory.

The concept of NTK is particularly relevant for understanding the behaviour of neural networks in the regime where the number of parameters tends to infinity [118], where it can be shown that the NTK effectively governs the learning dynamics throughout training. Kernel methods enable the derivation of analytic bounds on the generalization error [119], and their connection to infinitely wide neural networks extends these results to them under the NTK regime [118].

The analysis of training dynamics under the NTK can be used to guide the design of more efficient architectures that generalize better [120]. More generally, a theoretical understanding of the influence of various factors such as initialization, network size, and training time can reveal many properties of how neural networks learn [121]. By providing a bridge between neural networks and kernel methods, the NTK offers theoretical insights and practical benefits for improving machine learning models, and constitutes a crucial concept to expand our theoretical understanding of generalization in machine learning.

2.5 Generalizing beyond the training distribution

Much of the success in generalizing the predictions of machine learning algorithms is found when generalizing to new data that comes from the same distribution as the training data. However, in many real-world scenarios we observe shifts over time, changes in the functional connection between outputs and covariates, and more generally shifts in the data-generating distribution, i.e. $P_{train}(X, Y) \neq P_{test}(X, Y)$. The generalization under distributions that differ from the training distribution is often referred to as out-of-distribution (OOD) generalization [27].

What does it mean, concretely, when we say that the test distribution differs from the training distribution? Broadly speaking, anything that causes a change in the joint distribution $P(X, Y)$ satisfies this definition. To provide more details,

we can keep in mind the relationship between joint and conditional distributions $P(X, Y) = P(Y|X)P(X)$, and consider how each part may change. A shift in the test distribution, therefore, can be caused by one or more of the following situations:

(i) the data-generating process of the population shifts, $P(X) \rightarrow P'(X)$, a phenomenon called covariate shift or data drift

(ii) the features change to a different set, $X \rightarrow X'$, generally referred to as domain shift

(iii) the conditional distribution changes, $P(Y|X) \rightarrow P'(Y|X)$ (either the noise or the functional dependency), a setting often labelled concept drift

(iv) we are predicting a different but related outcome $Y \rightarrow Y'$, i.e. generalization to new tasks

Such shifts are common in practical settings in biomedicine. Imagine a situation where we have developed a predictive algorithm for the treatment response to an infectious pathogen based on the patient's clinical history. We might have to adapt the model if the population of the area around the hospital changes over time (i), if we start gathering metabolomics data instead of using the patient's clinical history (ii), if the pathogen develops resistance to certain treatments (iii), or if we predict specific complications rather than the response to treatment (iv).

Each of these situations represents a different challenge to address, and this complexity has led to the development of various methodologies to improve generalization performance, ranging from transfer learning, to domain generalization, to meta-learning. The following sections will address these paradigms for teaching a model to generalize, and examine the current state of the theoretical support describing their workings.

2.6 Paradigms to improve OOD generalization in machine learning

Depending on the structure of the predictive problem (as described in Section 2.5), several related but distinct approaches have been developed in machine learning to improve generalization performance when deviating from the training distribution. From an evaluation perspective, testing a model on hold-out OOD sets can be a good proxy for measuring generalization performance; it has been observed that in- and

out-of-distribution performance tend to increase jointly, although there is a strong dataset-dependent influence [122].

Rather than just evaluating the OOD performance of the model, several methodologies have been developed to improve the robustness of ML models and increase their predictive capabilities in complex settings. I will present here an overview of some of the main concepts employed in this field, which were selected because of their relevance for biomedical applications generally, or more specifically for the applications presented in the following chapters. In particular, the epigenomic imputation task presented in Chapter 4 makes use of transfer learning to predict individual-specific epigenetic patterns, and the prediction of antimicrobial resistance from Chapter 5 employs variants of zero-shot learning (ZSL) to analyse the generalization challenges for predictive models.

The literature on improving generalization in ML models includes a much wider array of methodologies, including distributionally robust optimization [123], adversarial learning [124], active learning [125], and many more. However, an in-depth discussion of these methodologies goes beyond the scope of this thesis.

2.6.1 Transfer learning

The standard setting in machine learning assumes that the training and test data share the same data-generating processes, and the same feature space. In many situations, however, it may be difficult or outright infeasible to obtain sufficient training data. The method of transfer learning (TL) relies on transferring knowledge or obtaining a high-performance predictor for the target set from a related source domain.

The principles underlying TL are easy to contextualize, and we can find intuitive examples of their workings even in human behaviour. We expect, for example, that an artist skilled in drawing with a pencil would have a much easier time learning to paint than somebody who has no experience with the visual arts: the two tasks bear enough similarity that the shared information is beneficial for the learner. The same concept is then applicable to machine learning: a learning algorithm trained on related data or a similar task can be used to overcome limitations of insufficient data.

To formally define TL, let us follow the convention adopted in previous works in the literature [126, 127], wherein a domain $\mathcal{D}$ is defined as a pair

$$\mathcal{D} = \{\mathcal{X}, P(X)\} \tag{2.4}$$

where $\mathcal{X}$ is a general feature space, and $P(X)$ is the marginal probability distribution for the observations $x \in \mathcal{X}$. Given a domain, we generally want to solve a task $\mathcal{T}$, which can be represented as a pair

$$\mathcal{T} = \{\mathcal{Y}, f(\cdot)\} \tag{2.5}$$

where $\mathcal{Y}$ is an output space representing the prediction target, and $f : \mathcal{X} \to \mathcal{Y}$ is a predictive function.

Following the notation just introduced, then, transfer learning is defined as improving a predictive function f_T for a target task $\mathcal{T}_T$ on a target domain $\mathcal{D}_T$ by using information from a source domain $\mathcal{D}_S$ and corresponding task $\mathcal{T}_S$ (compared to simply training a model f_T using data from the target domain). Crucially, TL methods rely on having access to data from the target domain at training time; if we aim to overcome differences in the source and target setting, they need to share information on some level, and this information must be available to the algorithm.

The specific methodologies to implement TL are varied, but all of them require additional interventions on source models or data. We can see this, for example, in the case of shifts in the marginal distributions between domains, where simply using a function trained on a source domain does not lead to optimal performance on a target domain [128]; the model requires, therefore, some form of additional training or adaptation.

Transfer learning is a general term that covers a variety of approaches to transferring knowledge between domains and tasks. TL can address changes in both the marginal and conditional data distribution between sources and targets, but importantly it relies on access to data from the target, whether labelled or unlabelled. Because of this generality, transfer learning is strongly associated with other branches of machine learning, and some special cases of TL are treated as a research field of their own. The main example of this latter situation is domain adaptation (DA), which refers to the setup where there is a shift in the distribution of the covariates $P_{source}(X) \neq P_{target}(X)$, but the conditional distribution remains the same $P_{source}(Y|X) = P_{target}(Y|X)$ [129]. DA is a well-studied setting, for which it is possible to derive theoretical bounds [130],

and it is especially prevalent in biomedical applications, particularly for biomedical imaging.

Semi-supervised learning (SSL) [131] also bears some resemblance to TL; however, the crucial difference is in the fact that in SSL methods the labelled and unlabelled data come from the same domain, and the task is shared between the training and test settings. In a sense, then, TL is a broader and more challenging paradigm than SSL.

While certain implementations of TL require complex workflows, a very common approach in practice is the pre-training and fine-tune paradigm. Essentially, the model is first trained on the large source dataset; subsequently, the model is fine-tuned on the smaller target dataset using a much lower learning rate, so that the pre-training phase essentially acts as a good initialization of the model parameters, allowing the convergence to a good local minimum of the loss function even with a small amount of data. A huge challenge of this and similar approaches is the phenomenon of catastrophic forgetting, where subsequent training erases learned information from the network [132]. To obviate this risk, some parameters in the pre-trained model may be frozen before tuning to preserve the feature maps learned by the model.

Additionally, when the source and target domains are not sufficiently related (or similar), or the model chosen is unsuitable to transfer knowledge between the two domains, we may incur in negative transfer, meaning that the use of source knowledge actively deteriorates the target performance [133].

Within the context of healthcare and biomedicine, transfer learning has been employed for disease diagnosis from medical imaging [134], glucose level prediction for diabetic people [135], analysis of EEG data [136], MRI brain imaging [137], analysing genomic sequences [138], and imputing individual-specific epigenetic patterns [23], among other tasks. For more extensive reviews on the topic of Transfer Learning, I refer to [127, 139–141].

2.6.2 Domain generalization

The goal of domain generalization (DG) is to learn a model using one or multiple related domains, such that the model generalizes well to a target unseen domain [142, 143]. For example, a classifier model could be trained on sets of photographic images and pencil drawings, and then be used to classify the content of oil paintings.

Crucially, domain generalization differs from domain adaptation and transfer learning because DG methodologies do not have access to data from the target domain at training time [144]. The most typical DG setup involves the use of multiple training domains, which is labelled multi-source DG [145]; when training only on a single source domain, the problem essentially reduces to the analysis of OOD robustness [146].

DG shares some similarities with the field of multitask learning, an approach where a model is optimized on several tasks at the same time to learn robust representations from the covariates [147, 148]. However, multitask learning does not generally attempt to extend to new domains, but rather focuses on the specific training tasks; DG can then be considered an extension of multitask learning.

Without access to the target domain, the task of selecting the best model becomes a considerable challenge [149]. The common methodology of using a validation set provides relevant information only if the validation set is sufficiently similar to the test set; in case this condition does not hold, there is no reason to expect that a model that performs well on the validation set will also display good results at test time.

One of the most common methodologies for domain generalization is based on learning domain-invariant representations of the data by minimizing differences across training domains [150, 151]. Intuitively, if the learned features are robust to domain shift, we can expect them to offer valuable information also on a new unseen domain. This domain alignment can be achieved with a variety of methods, ranging from contrastive losses [152], to minimizing maximum mean discrepancy [150], to adversarial learning [145].

Given its potential for improving the robustness of ML models, DG is a topic of considerable interest for the biomedical field. Exploration of DG methods has shown some results in EEG classification [153], medical image segmentation and classification [154, 155]; however, its effectiveness has been put into question in certain settings [156], highlighting the need for more extensive research work.

Domain generalization remains an extremely challenging task that is nonetheless crucial for improving the robustness of machine learning models. For more extensive reviews of the field of DG, I refer to [144, 145].

2.6.3 Meta-learning

Meta-learning, often referred to as "learning to learn," is a paradigm in machine learning where the goal is to automatically derive suitable inductive biases, starting from several related tasks to generalize to unseen tasks [157]. Meta-learning is about adapting the learning process itself, utilizing past learning experiences to tackle new tasks more effectively. It contrasts with traditional machine learning approaches that typically start each task from scratch [158]. By learning optimal learning strategies from a variety of tasks, meta-learning models can generalize better to new tasks. This is because they develop an in-depth understanding of task structures and dependencies, which can be transferred across tasks [159]

The field encompasses various strategies, such as optimizing hyperparameters, algorithm selection, and architecture design. These methods help in identifying the most suitable learning approaches for specific problems based on the characteristics of the data and the tasks. We can also classify some of the most commonly used techniques for improving the learning process of a model, such as bagging [160] and boosting [161], under the label of meta-learning.

A popular approach to meta-learning focuses on the gradient descent process used to train ML models. This gradient-based meta-learning is best exemplified by model-agnostic meta-learning [162], a framework to optimize a model's initialization such that it can quickly adapt to new tasks using only a few gradient updates. Other meta-learning applications focus on memory-based strategies to learn from past experiences [163, 164]. Such approaches attempt to mimic the flexibility of human intelligence, where individual observations can lead to large shifts in behaviour. The hope is that with these procedures, and given sufficient computation resources and data, a learning system can learn not just a model for its environment, but rather a reasoning procedure [163].

Meta-learning has been effectively applied in areas like few-shot learning, where models learn from a minimal amount of data, and reinforcement learning, where it aids in faster adaptation to new environments [158]. Despite its potential, however, complex meta-learning solutions face challenges such as computational efficiency [165], the need for effective task design [166], and the risk of overfitting to the meta-tasks [167]; these issues can hinder its ability to generalize across truly novel tasks [168].

Due to the ubiquity of its most established approaches like bagging and boosting, meta-learning approaches have long been employed in the biomedical field. Some

more complex versions of meta-learning have been employed for disease prediction [169], analysis of electronic health records [170], and drug design [171]. For a review of meta-learning approaches in machine learning, I refer to [172].

2.6.4 Zero-shot and few-shot learning

Classification is one of the most studied tasks in machine learning, due to its wide applicability, the availability of considerable theoretical tools such as diverse and meaningful evaluation metrics, and the good scaling properties of classifiers.

Standard classification models can only recognize samples belonging to classes observed in training. Frequently this characteristic is hardwired in the architecture, with many models making use of softmax activations to assign a label to each sample. This methodology is convenient for mapping a representation produced by a hidden layer of a neural network to a set of predefined labels; however, it generally precludes extending predictions to new and unseen classes. Methodologies to generalize prediction to classes that are not seen during training have been developed under the label of zero-shot learning (ZSL), whereas the slightly less stringent case in which just a few samples for the new class are available is called few-shot learning.

ZSL is a challenging research field, one which has many practical implications; some important application settings include the cases where target classes are rare or change over time, and when the space of target classes is huge [173].

One of the key parts of ZSL is the inclusion of semantics describing the classes to be able to bridge from training classes to test classes [174]. Essentially, to perform zero-shot prediction, we need a method that can obtain suitable representations for the new classes, such as a language model used to embed the description of the unseen class.

While in the original setting of ZSL the test set is composed entirely of novel classes, a more relevant approach that has been gaining traction is the so-called generalized zero-shot learning [175], where the test set is a mix of novel classes and classes for the training set. This approach is a much closer approximation of many real-world scenarios where the generalization of a zero-shot model might be relevant

ZSL methods can be further distinguished into inductive and transductive learning approaches, where the differentiating factor is that transductive learning includes unlabelled data from the unseen class in the training procedure, while inductive learning

does not [176]. Transductive learning is strongly associated with semi-supervised learning, with the two approaches corresponding when the unlabelled data belongs solely to the unseen classes. As transductive ZSL has access to some unlabelled data from the unseen classes, it holds an edge compared to the inductive method [177].

Focusing on biomedical applications, ZSL methodologies have been used in disease classification [178], medical image classification [179], clinical natural language processing [180], and protein function prediction [181]. For a general review of the topic of ZSL, I refer to [173].

2.7 The role of causality in generalization

Machine learning has been exceptionally efficient in modelling correlations from observed data in certain settings. However, it has proven remarkably challenging to extend this success when generalizing to new data, and even more so to new tasks and problems. In recent years, the connection between causality and machine learning has gained increasing attention for its potential to address these shortcomings [182].

Reasoning in animals—and especially humans—makes use of complex structures and subtasks such as counterfactual outcomes, temporal shifts, and relationships between domains. By learning causal connections between observations, we are able to imagine alternative futures and plan the actions that will lead to the desired outcome. Extending this reasoning framework to computational models would offer appealing possibilities to increase robustness and transparency; contrary to neural networks, human perception turns out to be robust to many types of perturbations of the input [183, 184], which is a highly desirable feature for high-stakes applications of machine learning.

To this end, progress in causality for machine learning has required the development of new tools that allow us to discover causal structures and infer the functional relationships between variables, ranging from Pearl's Do Calculus [185] to Rubin's Potential Outcomes framework [186]. Moving from correlational observations to causal queries like interventions and counterfactuals is a daunting task, but the potential impact on machine learning is remarkable.

One of the most tantalizing possibilities offered by causality is the robust and meaningful factorization of data distributions into a form that modularizes the dependencies between variables to only include direct causal connections. The so-called independent

mechanisms assumption states that the generative model of a causal process is composed by independent, autonomous modules that do not affect each other [187, 188]. This characteristic leads to interesting consequences for generalization capabilities and practices like transfer learning: for example, a shift in certain covariates will only impact a few modules of the causal model, leaving the remaining modules unaffected [189, 190]. This limited shift, in turn, enables the adaptiation of a model with only small adjustments.

There are many formalisms that can be used to describe these modular mechanisms, but perhaps the most popular is the structural causal model, in which we have a set of random variables $X_1, ..., X_n$ associated with the nodes of a directed acyclic graph $\mathbf{G}$ [185]. Each variable is affected by the incoming edges from the parent nodes, and can be described as:

$$X_i = f(\mathrm{Pa}_i, U_i) \quad (i = 1, ..., N) \tag{2.6}$$

where Pa_i represents the set of parent nodes and U_i represents the exogenous variables not described by the model.

Access to a full causal model (i.e., knowledge of the functions f and the exogenous variables U_i) enables queries that would prove impossible with a purely correlational machine learning model. Interventional and counterfactual queries, where we answer the questions "what if...?" and "what would have happened if...?" enable the exploration of alternative outcomes in a systematic manner. This transparency is particularly desirable in high-stakes applications like biomedicine, since it enables interpretation of the decision process of a model, which can then be communicated to patients and other stakeholders, leading to a more robust and ethical use of predictive models [191]. It is no coincidence, then, that causal formalisms constitute a sizeable portion of explainable approaches in biomedical ML [192].

Discovering and quantifying the causal relationships between variables is, however, a major challenge. The first obstacle arises with the collection of data: to analyse causal relationships, we can either collect explicitly causal data (e.g., perform interventions like gene knockout experiments), or we can use observational data and make assumptions on the structure of the underlying causal graph $\mathbf{G}$ [193]. This proves especially complex in the realm of biology, where data is inherently noisy, and it is often hard or impossible to access information on the ground truth.

Due to this challenge, an alternative research direction is to use state-of-the-art machine learning models to learn representations of the data that capture the causal structure of the data generating process [194]. This causal representation learning task is extremely complex; nevertheless, several methodologies have shown promise in applications such as domain generalization and generative modelling [195, 196]. Other related approaches, such as modelling causal structure into neural networks, hold promise for improving generalization in the zero- and few-shot settings [197].

The formalization of causality is a difficult research direction that raises unique challenges for the common approaches to machine learning. For example, most machine learning classification tasks are anti-causal, as we infer causes (labels) from effects (observations) [198]. In this case, the search for a causal mechanism that leads to an observed effect is driven by anti-causal models, and yet we still need a causal model to validate the results.

Many of the approaches developed for the theory of causality are restricted to relatively simple settings, such as acyclic causal graph, and a limited number of variables. Additionally, causal methods are often tested on extremely idealized datasets, that do not resemble real-world datasets [199]. In contrast, many biomedical applications deal with high-dimensional data that is generated by processes composed of feedback loops, and that often present strong shifts over time; the direct translation of causality methods to biomedical practice, therefore, still proves challenging. However, recent works have shown promise in leveraging insights from causality, for example for the identification of biomarkers [200], the analysis of single-cell perturbations [201], and the prediction of treatment outcomes [202].

Chapter 3

Biomedical Machine Learning

3.1 The state of machine learning applied to biomedical data

One of the core aims of machine learning algorithms is learning to identify patterns in data. Applying this capability to biology and healthcare is especially appealing so that we can learn to identify mechanisms of disease, risk factors, and targets for intervention and treatment. Aside from practical considerations, this pattern recognition paradigm that enables machine learning mirrors many of the biological processes that control and regulate the biochemistry of life; after all, many facets of molecular biology are controlled by mechanisms of complementary patterns, where chemical and structural properties of complex molecules control their interactions and combine into an exquisite cascade of pathways.

The mechanisms for the detection of patterns are truly pervasive at every level of life. For individual cells, forms of biological pattern recognition range from the fundamental Watson-Crick-Franklin base pairing that controls interactions between nucleic acids [203], to more nuanced mechanisms such as motifs, short recurring patterns in sequences of nucleotides that influence the binding of proteins to DNA and RNA [204–206], and more generally to active sites on enzymes or proteins [207].

At a higher level of the hierarchy of biology, the pattern recognition paradigm is a crucial component to regulate and maintain the extraordinarily complex processes that govern multicellular systems. Perhaps no example is more emblematic in this sense than the mechanisms of immunity and the recognition of foreign dangers. Evolutionarily, the

necessity of recognizing external pathogens and dangerous substances has arisen from the earliest stages of life [208]. However, a truly advantageous system of immunity has to rely on patterns to distinguish the "self" from the "non-self" and to preserve memories of dangerous pathogens for future interactions. Such mechanisms are hundreds of millions of years old [209] and have taken many diverse forms, including the bacterial CRISPR-Cas system [210], Argonaute proteins [211], pattern recognition receptors [212], and up to the exquisitely complex systems of antigen presentation and recognition by T-cell and B-cell receptors [213–215].

Finally, zooming out further, recognition of patterns has been posed as one of the main components underpinning higher cognitive processes, regulating complex social interactions and subconscious processes [216, 217].

Given their ubiquity and importance, investigating these biological pattern recognition mechanisms is crucial for improving our understanding of biological processes and human disease; however, the extremely complex interplay that happens between organic molecules and biological systems makes deciphering their patterns a daunting task. Here is where machine learning comes into play, with the promise of leveraging vast amounts of data to decode the latent patterns of interactions and guide experimental results. This idea is appealing for numerous processes in molecular biology; an example can be found in the interplay of epigenetic modifications, which are control mechanisms that regulate gene activity, and play a crucial role in development and disease [218]. Of particular interest for the work presented in Chapter 4 are the post-translational modifications of histone proteins, the structural support around which DNA is wrapped to form nucleosomes. Histone modifications are speculated to follow a hidden "code" that relates patterns of modifications to phenotypic outcomes [219, 220], and decoding these patterns could reveal crucial information on human disease.

This is then the most tantalizing possibility of biomedical ML: the hope that, if a model has learned to accurately predict the outcome of a biological process, we can understand the patterns governing the process itself by examining the patterns learned by the algorithm. In practice, however, this decoding of patterns from data has not been the primary output of biomedical ML research, due to the complexity of the task. Currently, the most impressive advancements in biomedical ML consist in large improvements in the predictive performance of models applied to specific tasks, which does not necessarily translate to a more in-depth understanding of the biological mechanisms involved.

Some areas where large quantities of structured data are available, such as protein structures and biomedical images, have advanced considerably; other applications where that is not the case, such as phylogenetics, are still immature and yet to produce innovative outcomes [221]. Among the most well-studied applications of ML to large-scale biomedical data we find protein structure predictions [222, 223], protein function prediction [224, 225], multi-omics data integration [226], medical image analysis [227, 228], and image-based profiling assays in high-throughput imaging [229, 230].

The advances of many of these predictive tasks are founded on the extensive work done by international consortia to gather data on genetic and epigenetic annotations (Ensembl [231], ENCODE [232, 233], Roadmap Epigenomics [234]), protein sequences and structures (UniProt [235], PDB [236]), cancer data (TCGA [237]), and more. Due to the experimental and resource constraints intrinsic to biological and clinical experiments, collaborative efforts to gather large-scale high-quality datasets will play a pivotal role in advancing our understanding of human disease in the coming years.

The decoding of patterns in molecular biology is not the only appealing research direction predicated on the paradigm of machine learning. Disentangling and distilling information from complex biological data opens up interesting opportunities in healthcare, such as the identification of disease risk factors, predicting the response of a patient to specific therapies, and tailoring clinical treatments to maximize effectiveness. The stratification of patients into small groups—or even at the level of the single individual—is the core concept of the field of precision medicine [238, 239], which is founded on the hope that personalized treatment can significantly improve the effectiveness of healthcare and optimize resource allocation. The potential of integrating complex and heterogeneous sources of biological information was recognized early in the development of precision medicine [240], and the field has now moved to include diverse sources of data, from the so-called -omics sources (proteomics, metabolomics, epigenomics, and more), to electronic health records, and up to medical imaging [241].

Within the field of precision medicine specifically, recent developments in ML applications include methods for diagnosis and disease prediction [242, 243], the prediction of treatment outcomes and drug response [244], identification of biomarkers [245], clinical trial design [246], and drug discovery [247]. Importantly, the progress of machine learning models for healthcare is not restricted to theoretical *in silico* predictions, but it is leading to increasing translation into practical settings. For example, several of the advancements brought by ML have enabled improvements to

medical devices, with hundreds of them approved by the Federal Drugs Administration in the USA [248]. It is important to note, however, that the vast majority of these advancements have produced applications mainly in radiology and for the treatment of cardiovascular and neurological diseases, while many other areas of research have not led to the same level of success.

The translation of *in silico* results to clinical practice remains an enormously challenging task to undertake. The term "AI Chasm" aptly describes the large discrepancy between the development of innovative models and their application in practice [249]. Often, ML models that show good performance in an isolated study fail to translate to practical applications [250]. Emblematic is the case of diagnostic and prognostic models for COVID-19, which were developed in large numbers during the pandemic, but showed high risk of bias in two systematic reviews [251, 252].

The main factors that limit the translation of machine learning progress to clinical practice revolve around methodological errors and biases that arise in the data collection and model design. Such biases introduce confounding factors and spurious correlations that can seriously hinder the generalizability of the results obtained, for example leading to cases where a model trained within a specific hospital offers poor performance when ported to another. Some of the specific problematics that I will discuss in Section 3.2 are a direct cause of distribution shifts as examined in Chapter 2, while others present more broad issues for the biomedical applications of machine learning. The development of methodologies to ensure robustness within generalization settings will be crucial to overcome these challenges; their correct application will also require careful consideration and a more in-depth understanding of the biases present in biomedical data.

3.2 Sources of bias in the data and design

In this thesis, the label "biomedical data" is used as an umbrella term that refers to data that describes biological processes involved in human disease, as well as data that describes all the variables found in the clinical practice. As such, under this label I included a wide variety of typologies of data, like clinical records, molecular structures, sequencing data, and images.

Biomedical data is a rich source of information that can lead to actionable improvements in clinical decision-making [253], to the development of novel drugs and treatments [254], and to shaping policymaking [255]. However, biomedical data presents

several criticalities that need to be addressed; otherwise, the intentional or unintentional misuse of the findings derived from such data can lead to negative real-world consequences [256]. After all, it is well known that data quality is one of the most crucial limiting factors for machine learning tasks [257], and this holds doubly true for biomedical machine learning.

The data itself is, however, only one piece of the puzzle in the pursuit of a scientific inquiry; many other considerations have to be carefully explored to ensure the quality of the research outcomes. What tools are used to perform the analysis? How do we measure the performance of a model? What target is the model trained on, and is it a suitable task to then achieve the desired real-world application?

I will present here a discussion of some of these factors, as they are crucial for the development and translation into clinical practice of innovative applications of biomedical machine learning. The topic of validating the results of biomedical ML models is examined separately in Section 3.3, due to its critical importance for addressing the AI Chasm.

3.2.1 Selection bias

The first and most obvious challenge in developing robust research applications in biomedicine and healthcare is the issue of selection biases. This type of bias refers to systematic errors in the selection of samples in a study. This general definition can refer to the selection of patients in clinical studies, the outsized representation of certain categories in population studies, or even the biomolecules that are the focus of in-depth analysis.

A biased selection can severely affect the conclusions linking exposure and effect or the relationship between biological molecules. Essentially, non-random associations can muddle the causal connections that would be employed for decision-making, often by inducing or preserving confounding biases and spurious correlations [258].

As a result, medical studies can lead to the identification of erroneous risk factors [259] and to substantially biased estimates of associations [260]; due to the spurious correlations introduced by improper selection, it can even become complex to ascertain whether a study was properly randomized [261].

The selection of cohorts on which these studies are performed is often influenced by selective participation or attrition that distorts the associations analysed [262].

The data sources produced in this way can lead to an overwhelming representation of the population of certain countries—such as the United States of America—and can amplify systematic under-representation of other populations [263].

Finally, within the context of molecular biology, a biased selection of targets for research efforts can lead to skewed knowledge bases that can affect downstream analysis. For an example relevant to Chapter 4, the selective study of histone modifications may lead to biased datasets where certain mechanisms are mapped far more extensively than others. Some of these modifications (e.g., H3K4me3 and H3K27ac) are extensively studied, due to how evolutionarily well-conserved they are [264] and the availability of biochemical methods to study them [265]; however, many other less-studied modifications still play crucial roles in gene regulation [266, 267], and yet are poorly understood. The visualization of the data matrix used for the experiments in Chapter 4 (Supplementary Figure A.1) offers a clear example of this type of selection bias.

Epigenetic modifications are far from the only molecular mechanism affected by this kind of bias. For example, the uneven research effort dedicated to different proteins can lead to inherent biases in protein-protein interaction networks [268, 269], to limitations on our understanding of protein function due to an over-reliance on certain methodologies [270], and to biased semantic similarity measures for proteins [271].

3.2.2 Reporting and publication biases

A major issue in the assessment of healthcare interventions is caused by certain types of data or outcomes being disproportionately represented in training datasets, often due to selective reporting or publication bias [272]. For example, interventional studies with negative results are often not published, which skews the relevant literature to represent only a biased subset of the relevant evidence [273, 274].

This issue has been known for a long time [275] and addressing it is crucial for robust applications of machine learning. As ML models learn the biases of the data used to train them, they may encode skewed functional relationships, leading to inaccurate predictions or recommendations. In biomedical machine learning, this can manifest in several detrimental ways, with models that fail to generalize across diverse patient populations [276], overlook rare diseases [277], or exacerbate existing health disparities [278].

The reliability of biomedical machine learning applications hinges on the representativeness of the training data and the validity of the background domain knowledge used to select suitable inductive biases. Addressing publication biases and developing better standards for reducing overoptimistic reporting is a crucial challenge that needs to be addressed by the research community [279].

3.2.3 Non-stationary systems and data drift

One of the most common problems to tackle in biomedical ML is that the underlying systems rarely exhibit the stationary stability necessary to guarantee the IID assumption over time. To begin with, many biological systems themselves are complex dynamic entities that exhibit nonlinear fluctuations and systematic drifts [280]. This constantly evolving nature is a considerable challenge for many machine learning models, as their capacity for capturing this layer of information depends entirely on the data collection process and the inductive biases designed into the algorithms.

An emblematic example of these limitations is the AlphaFold2 model [222], considered one of the greatest successes of applied ML in biology. AlphaFold2 performs very well on frozen structures of proteins, but has to rely on contrived procedures to predict structural ensembles [281], and is completely unable to describe complex dynamic proteins such as intrinsically disordered proteins [282].

This issue extends beyond the small scale of molecular biology and typically affects also the clinical setting, where drifts in data distributions over time represent a major concern [283, 284]. The causes of such drifts are varied, from evolving clinical practices [285], to demographic shifts in patient populations, or changes in data acquisition methodologies and policy chances [286]; handling this concept drift is, therefore, the topic of considerable research efforts in ML [287].

Several mitigation strategies can be employed, although all approaches present practical constraints that may limit their efficacy. Among the strategies to ensure that the data drift does not lead to unacceptable clinical risks we find continuous monitoring of the models [283], periodic retraining [288], continual learning [289], and the use of generalization approaches such as those described in Section 2.6 to enhance OOD robustness.

3.2.4 Data aggregation and batch effects

Due to cost or technical constraints, the gathering of biomedical data often leads to limited datasets. This phenomenon is especially prevalent in clinical settings, where small datasets are a common occurrence for a wide variety of reasons, ranging from low prevalence of certain diseases, to ethical constraints and resource limitations [290–292].

A small sample size may hinder the robust development of ML models; it has been shown, for example, that insufficient data may lead to biased performance estimates when paired with unsuitable study design [293]. To overcome this issue, there is often the necessity of integrating similar datasets collected in different settings. However, combining datasets gathered in different conditions is far from trivial; challenges such as data heterogeneity, inconsistent data quality, and differences in population may introduce detrimental biases in the combined data. This bias can lead to negative consequences, as models trained on heterogeneous, inconsistent, or non-representative data are prone to poor generalization, leading to decreased performance when they are applied to new data from different sources.

Additionally, the aggregation of different datasets entails entirely new issues, such as ensuring privacy and security for the human participants involved [294, 295]. Several solutions are being developed to address these additional concerns of data integration, such as federated learning; however, these methodologies require careful consideration to account for the domain shifts between data sources [296].

Data aggregation is not just a challenge in the clinical setting, but also for biological data gathered in large-scale experiments, due to the presence of batch effects. Batch effects are systematic non-biological variations that arise from technical inconsistencies during data collection and processing; these variations can be attributed to factors such as differences in equipment, operator handling, time of experiment, environmental conditions, and sample processing protocols [297]. Such effects are pervasive across various biomedical domains, including genomics, proteomics, microbiome analysis, and imaging [298–300]. Batch effects pose significant challenges to the development and application of machine learning models in biomedical research, as they can lead to false positives or misleading patterns, suggesting biological relevance where none exists because of spurious learned correlations [301]. In turn, models trained on batch-affected data may fail to generalize to new datasets from different batches, limiting their usefulness.

3.2.5 The curse of dimensionality

The curse of dimensionality refers to the exponential increase in computational complexity and data sparsity as the number of features or dimensions in a dataset grows [31]. This phenomenon poses significant challenges in the field of biomedical machine learning, where high-dimensional data is commonplace due to the vast number of potential biomarkers, genetic variations, imaging features, and clinical variables involved.

In high-dimensional spaces, data points become increasingly sparse, and conversely the amount of data necessary to cover the feature space increases exponentially; this sparsity makes it difficult for machine learning models to find meaningful patterns and generalize from training data. In the clinical context, where sample sizes are often limited due to the cost and difficulty of data collection, this issue is particularly pronounced [302].

Reducing the dimensionality through the use of feature selection techniques can remove sources of noise and reduce the biases captured in a dataset [303, 304]; for some ML models, the increased difficulty in defining meaningful distances can lead to poor generalization performance that leads to inaccurate predictions [305]. Additionally, the high number of features typical of certain biomedical settings can make it harder to examine the workings of ML models, which can lead to biased interpretations of a model's outputs and hinder the identification of actionable insights.

A common strategy to mitigate the biases introduced by high-dimensional data in biomedical machine learning is the use of dimensionality reduction techniques to extract a lower-dimensional representation of the signal while discarding noise [306]. The use of linear dimensionality reduction techniques such as principal component analysis and singular value decomposition is prevalent in many bioinformatics workflows, due to their convenient statistical properties and computational efficiency [307, 308].

When the scale and structure of the data allow it, non-linear methods offer the opportunity to capture the structure of the data in a much richer manner. Models that learn lower-dimensional representations offer interesting possibilities in molecular biology and healthcare, due to their capability for handling complex data such as graphs [309], sequences [310], and electronic health records [311]. Other non-linear techniques are of special interest because they enable rich visualizations of the data despite all the limitations involved [312], which can aid the exploration of the data and improve the communication of research outcomes.

By addressing the challenges posed by the curse of dimensionality, researchers can develop more robust and reliable machine learning models that enhance our understanding of complex biomedical phenomena and improve clinical outcomes.

3.2.6 Missing data

Missing data is a prevalent issue in clinical datasets, stemming from various sources such as patient non-compliance, errors in data collection, and limitations in recording systems [313]. The presence of missing data can significantly impact the development and performance of machine learning models, and the robustness of statistical findings in general. Improper handling of missing data can be a major source of bias in clinical trials [314–316]. If the missing data is not properly accounted for, the model may be trained on a biased subset of the data, leading to skewed predictions.

To ensure that missing data does not negatively affect the results, it is important to consider the cause of the missing features. Different patterns of "missingness" can have considerably different consequences on a statistical analysis; a commonly used classification devised by Rubin [317] classifies missing data into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The non-random component of MNAR data can carry information in and of itself, and can be leveraged to improve predictive performance [318]; however, improper handling can lead to biased estimates [319] and loss of statistical power [320].

A common approach to account for missing data (MCAR or MAR) is to impute the unobserved values [321]; this method can prove quite effective, although researchers must exercise caution when considerable portions of the data are missing, as it could lead to biased estimates [322]. Additionally, a common pitfall when using imputation methods is the accidental introduction of information leakage between the training and test data [323].

An alternative methodology to account for missing data consists in relying on models that can handle missing data directly, such as gradient-boosted trees [324]; however, many state-of-the-art architectures such as deep neural networks cannot natively handle missing values.

By understanding the nature and extent of missing data, applying appropriate imputation methods or selecting suitable algorithms, and conducting thorough sensitivity analyses, researchers can mitigate the risks associated with missing data. This ensures

that the conclusions drawn from ML models are robust, generalizable, and reflective of the true biomedical phenomena under investigation.

3.2.7 Proxy measures: the lack of ground truth and gold-standards

The development and validation of machine learning models in biomedical research heavily depends on the availability of high-quality ground truth data or gold-standard datasets. This requirement is reflective of the broader issue of data quality in machine learning, as the presence of biases in datasets can introduce spurious correlations and severely impact applications of ML [325].

Often no such high-quality data is available, for practical or technical reasons, leading to the use of imperfect targets or proxy measures on which the models are trained and tested. ML models trained on such imperfect signal are generally optimized to perform well on this specific prediction task, which might not necessarily mean that they perform well on the actual target task.

The use of surrogate measures is quite common in clinical studies, and it requires an exceptional level of caution to ensure that the robustness of the resulting claims [326]. Additionally, when we lack an objective evaluation setting for ML models, it can be extremely challenging to compare different approaches and to evaluate their effectiveness [327].

For an example relevant to the content of Chapter 4, the enrichment of chemical modifications on histone proteins is measured through ChIP-seq experiments that have to be processed with a complex bioinformatics pipeline. Establishing a ground truth result in this case would require the analysis of biological samples of exceptional quality, for example by using a large sequencing depth and numerous replicates. In practice, data of such quality is rarely available, and the samples measured ultimately provide an indirect and noisy measure of the presence of histone modifications for most experiments that can be used to train a model. The models trained on these signals learn to predict the noisy signal rather than the local enrichment of certain histone modifications, hindering the generalizability of the results. This fact complicated considerably the evaluation of the 2019 ENCODE Imputation Challenge [328], a competition organized to develop novel epigenomic imputation models that led to the development of the eDICE model [23].

The impact of these proxy measures can lead to considerable issues in the use of ML models in clinical practice. For example, it has been shown that algorithms that approximate healthcare needs using patient costs can lead to the exacerbation of racial inequalities in the provision of clinical care [18]. In the context of biomedical machine learning, high-quality datasets that directly capture information on the underlying phenomena are critical for training algorithms, evaluating their performance, and ensuring their generalizability to real-world applications. The scarcity or absence of such datasets presents significant challenges that hinder the progress and reliability of ML models in this field (e.g., [329]).

For classification problems specifically, high-quality data often requires annotation from human experts [330], which is costly and time-consuming [331, 332]. This factor limits the amount of data that can be collected, hindering the validation and even more so the training of ML models. Some works explored the use of crowdsourced annotations [333], which can be a viable method in certain fields like biomedical images annotation [334]; additionally, the disagreements and inconsistencies introduced by the crowd may actually be a positive feature that better reflects the complexity of real data [335], although other works highlight the possible issues raised by inconsistent annotations [329].

Due to the recent progress in the field of large language models and other foundation models, the idea of automatic annotation powered by these algorithms is an appealing prospect that could scale up the annotation efforts by a huge margin [336, 337]. However, the robustness of these models is still a point of contention, and their use should still be adequately monitored to avoid issues such as hallucinations and random variations caused by the specific prompts used [338, 339].

An alternative solution explored in the literature is the creation of synthetic data on which to train and evaluate the models [340]. With this approach, the complete ground-truth of the data is available to exactly quantify the performance of the models. However, simulated data is not guaranteed to capture the full complexity and variability of real biological systems; a model trained on such data may perform well on simulated data, but poorly on real data [341]. While synthetic data offers interesting possibilities—especially for preserving privacy—research in the field is still at an early stage, and requires considerably more effort to reach the robustness required for applying biomedical ML at scale.

3.2.8 Complex multilayered workflows

Bioinformatics workflows have become indispensable in the analysis and interpretation of large-scale biological data. With the advent of increasingly refined experimental techniques, the processing of raw data to extract the desired information has grown in complexity to include steps such as quality control, filtering, alignment, functional annotations, taxonomic analysis, and integration with established databases.

These procedures are typically composed in a workflow consisting of a sequence of computational steps that integrate diverse tools and databases to process raw data into meaningful biological insights. However, the increasing complexity of these software stacks introduces potential biases that can affect the reliability and reproducibility of the results.

Firstly, each software tool is based on specific algorithms and heuristics, each of which can introduce its own specific biases. Even programs designed for the same functionality may rely on different algorithms, leading to differences in the output data [342], which complicates the comparison of results produced by different laboratories if they employ different workflows.

In addition, different versions of the same tool can produce different results due to changes in algorithms, default parameters, or bug fixes. And even further, variations in configuration settings, such as quality thresholds or alignment parameters, can lead to inconsistent results. Many bioinformatics tools rely on a chain of dependencies, including libraries and auxiliary tools; incompatibilities or bugs in these dependencies can propagate through the workflow, introducing subtle biases.

The result is that systematic errors introduced at various stages of the workflow can accumulate, potentially leading to incorrect biological interpretations. For example, biases in variant calling can affect downstream analyses such as association studies or functional predictions, impacting the reliability of the conclusions drawn from the data [343, 344].

To address these issues, it is necessary to develop and adhere to standardized protocols and best practices. Such practices can include the use of version control tools [345], the creation of extensive documentation [346], and the development of pipelines to ensure reproducibility with tools such as Galaxy [347], Nextflow [348], and Snakemake [349]. Such workflows improve the tracking of data provenance, increase portability and scalability, and increase efficiency by avoiding repeating redundant

steps [350]. As a final layer to ensure the encapsulation of all the tools needed, the containerization of a bioinformatics pipeline using tools such as Docker and Singularity can streamline reproducibility [351].

While the use of these best practices can ensure robustness and reproducibility, well established benchmarking tasks are crucial to ensure the highest standard in publicly available pipelines [350, 352, 353]. Such resources enable the comparison of different tools, and can allow researchers to test their models on data processed with different pipelines to examine the robustness to the biases introduced by the software stack.

3.3 Validating predictions in biomedical machine learning

The robustness of clinical studies relies on first formulating a hypothesis, then conducting a study with adequate sample size and a suitable population, and then assigning a level of significance to the rejection of a null hypothesis in the form of a p-value. Issues like p-hacking, insufficient statistical power, and confounding biases can hinder the quality of the results, and ultimately slow or prevent the translation of actionable results into clinical practice [354, 355], although in certain settings such as meta-analyses the conclusions may be affected to a lesser extent [356].

Machine learning models are based on the opposite paradigm, wherein the data is gathered first, and only afterwards we attempt to derive relationships from it; ML models, therefore, cannot be evaluated in the same manner as clinical studies. Within this approach, the poor performance of the analysis is not described by the lack of significance for a statistical hypothesis, but rather by overfitting and lack of generalization capabilities [357].

Due to this shift in paradigm, validating the results of a machine learning model in biomedicine can be an extremely challenging task. The highest standard of evaluation would be to use the trained ML model to conduct a prospective study, and obtain experimental confirmation of its predictive capabilities. This method, however, is severely limited by cost and resource constraints: in many situations, gathering a sufficiently high-quality test set would require enormous expenses and might not be feasible on short time-scales. It is no coincidence that, for many biomedical ML tasks, only a small minority of the studies conducted employ external validation (e.g., [358, 359]).

The most common approach adopted in machine learning more broadly, and also in many biomedical ML applications specifically, is some form of internal validation. In this case, the data is split using methods such as cross-validation or bootstrapping to evaluate informative performance metrics on hold-out datasets [360]. Where possible, the use of publicly available datasets constitutes a valuable tool for the validation of ML models [361], especially because the wide availability of these datasets offers a common benchmarking setup on which competing models can be evaluated fairly.

In high-stakes settings such as healthcare, however, external validation still constitutes one of the best tools to ensure the necessary robustness standards [362, 363]. Approaches to systematize the evaluation of robustness of machine learning models, such as meta-validation based on similarity between datasets [364, 365], provide crucial insights that will enable responsible and effective adoption of ML tools in clinical practice.

On a practical level, a synergistic combination of internal and external validation is the most effective path to translate models into clinical practice [366]; internal validation can provide a first filter to select promising methodologies, so that resources for the costly and time-consuming external validation can be properly allocated [367].

When validating predictions for ML models, it is crucial to select evaluation criteria that accurately reflect the ultimate goal of a research endeavour. It is not uncommon for ML practitioners to select evaluation criteria that are accurate on a technical level and yet do not fully capture the necessary complexity of the research question. A deeper collaboration with domain experts and clinicians might ensure a more fruitful scientific inquiry; however, these experts are often involved only in parts of the studies [368]. The development of interdisciplinary guidelines for the development of biomedical ML models [369], or protocols for the reporting and assessment of their performance such as TRIPOD and PROBAST [370, 371]—and their extensions to AI models [372]—can provide valuable tools to support good scientific practice and ensure fruitful collaboration across domains.

The path from ML development to implementation in clinical practice is rife with challenges, including experimental design, ethics, and regulation [373]. This process also requires ongoing monitoring after adoption of novel models in clinical practice, as it is crucial to follow-up on their effects through impact studies [374].

Overall, the topic of validation for ML models in biomedicine is of crucial importance to ensure a successful translation of cutting-edge techniques to clinical practice and

other biomedical settings; while external validation is the gold-standard for model evaluation, its extensive use is still limited by resource constraints. Complementary efforts such as the creation and publication of large-scale databases and the development of systematic benchmarking settings could help bridge the gap between *in silico* research and robust evaluation of trained algorithms.

Chapter 4

Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning

Epigenetic modifications are dynamic mechanisms involved in the regulation of gene expression. Unlike the DNA sequence, epigenetic patterns vary not only between individuals, but also between different cell types within an individual. Environmental factors, somatic mutations and ageing contribute to epigenetic changes that may constitute early hallmarks or causal factors of disease. Epigenetic modifications are reversible and thus promising therapeutic targets for precision medicine. However, mapping efforts to determine an individual's cell-type-specific epigenome are constrained by experimental costs and tissue accessibility. To address these challenges, we developed eDICE, an attention-based deep learning model that is trained to impute missing epigenomic tracks by conditioning on observed tracks. Using a recently published set of epigenomes from four individual donors, we show that transfer learning across individuals allows eDICE to successfully predict individual-specific epigenetic variation even in tissues that are unmapped in a given donor. These results highlight the potential of machine learning-based imputation methods to advance personalized epigenomics.

Declaration

This chapter is based on an updated version of the published manuscript [23]:

> *Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning*
> Alex Hawkins-Hooker*, Giovanni Visonà*, Tanmayee Narendra, Mateo Rojas-Carulla, Bernhard Schölkopf, Gabriele Schweikert
> Nature Communications (2023)
> * equal contribution

Compared to the published version, the figure panels have been split into individual figures or smaller panels for improved readability; the figure captions have consequently been adjusted and improved. Minor grammatical corrections have been made for clarity. Data and code availability have been moved to the Methods section. The supplementary material from the original paper is included in Appendix A.

Author contributions (CRediT)

Alex Hawkins-Hooker: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing
Giovanni Visonà: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization
Tanmayee Narendra: Methodology, Validation, Software, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing
Mateo Rojas-Carulla: Conceptualization, Supervision
Bernhard Schölkopf: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition
Gabriele Schweikert: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

4.1 Introduction

Epigenetic mechanisms play an essential role in developmental biology and human disease [375, 376]. They act at the intersection of genetic and environmental factors to control, regulate, and propagate cellular responses, significantly contributing to diverse cellular phenotypes. Importantly, their influence on gene activity is reversible without altering the underlying DNA sequence. Therefore, they provide unique diagnostic and therapeutic opportunities and offer promising targets for precision medicine approaches [377–379], with particular interest for applications in cancer treatment [380, 381]. Advances in epigenome editing technologies are paving the way for including epigenetic modifications not just as biomarkers, but also as direct intervention targets for novel treatments [382–384]. However, crucial challenges remain, mainly because epigenomes are cell-type specific and dynamically changing on different time scales, for example during the cell cycle, development, or ageing. Therefore, decoding epigenetic patterns is particularly laborious, expensive, and data-intensive.

A more in-depth understanding of epigenetic modifications has shed new light on the mechanisms involved in certain neurological and neurodegenerative diseases, developmental disorders, and some forms of cancer [377, 385–387]. Large-scale efforts to map the functional properties of human epigenomes proved essential for these developments and have provided a crucial resource to understand how the interplay between genetic and epigenetic factors affects cellular identity and function [388, 389]. While these projects aim to profile diverse cell types exhaustively using various epigenetic assays, the associated experimental costs impose constraints that lead to incomplete maps, with many cell types still sparsely analysed. This sparsity presents a particular challenge for the study of individual-specific epigenomic variation. Individual-specific epigenomic signatures have the potential to inform personalized predictions for risk stratification [390], drug resistance [391, 392], or personalized therapies [393]; however, producing comprehensive individual-specific epigenomic maps remains practically infeasible, not least because of the difficulty of obtaining samples from certain tissues.

As a result, computational approaches that can leverage existing epigenomic data to impute the results of as-yet unperformed assays are of considerable interest, particularly if they are able to predict individual-specific variation. As well as advancing overall understanding of the epigenomic landscape, effective imputation methods have the potential to play a role in the development of novel precision medicine workflows, for example by predicting the results of epigenetic assays in tissues that are difficult to

probe in living patients, or aiding in the prioritization of epigenomic measurements [394]. Previous work in epigenomic imputation has shown that machine learning models can be trained to exploit the correlations between sets of epigenomic marks within and between cell types to successfully predict missing measurements [395–397]. However, these studies have focussed on the imputation of reference epigenomes, and have not explored the use of imputation methods to generate individualized predictions.

In this work, we introduce eDICE (Figure 4.1), a Transformer-inspired imputation model, which is trained to impute missing epigenomic tracks given sets of observed tracks. eDICE learns to encode the epigenomic signal in a set of observed tracks into factorised local representations of each cell type and each assay, enabling imputations to be made for unseen combinations of cell type and assay by decoding from the appropriate representations. We first show that our architecture leads to improved imputation performance relative to previous methods on the reference Roadmap epigenomes, while displaying significant practical benefits. Next, we use recently published individual-specific epigenomes from EN-TEx [398] to test whether eDICE can be used to generate individualized epigenomic imputations. Inspired by precision medicine applications, we devise a task designed to assess the utility of imputation methods for predicting individualized epigenomes in hard-to-access tissues, and find that transfer learning across individuals allows eDICE to predict individual-specific epigenomic variation in this setting.

4.2 Results

4.2.1 eDICE and previous work on epigenomic imputation

In 2015, Ernst and Kellis pioneered work in the field of large-scale epigenomic imputation by introducing ChromImpute [395], an imputation strategy for the reference epigenomic datasets produced by the Roadmap and ENCODE projects [388, 389]. Given sets of reference epigenomes generated by performing various epigenomic assays in a set of cell-types, the epigenomic imputation task posed by ChromImpute is that of predicting epigenomic tracks representing combinations of cell-type and assay for which experimental data are not available, thereby "completing" the epigenomic map. To solve this problem, ChromImpute adopts a regression-based approach, requiring the training of a separate ensemble of models for each target track. While ChromImpute has shown effective performance, it relies on manual engineering of input feature sets

and the training of thousands of separate models, preventing the effective sharing of information across the highly related tasks of imputation of different tracks.

Subsequently, imputation strategies based on tensor factorization have been proposed as a way of reducing the complexity of ChromImpute. PREDICTD [396] generates predictions via a linear combination of learned factors representing cell type, assay, and genomic location. Avocado [397] replaces the linear combination of factors used in PREDICTD with a learned nonlinear operation, by passing concatenated embeddings corresponding to each factor through a neural network. Tensor factorization approaches have the appealing property that, given a learned set of factors, predictions can be generated at any genomic location for a track corresponding to any combination of one of the modelled cell-types and assays. Nonetheless, the performance of these approaches has only outstripped ChromImpute on a subset of metrics.

Seeking to combine the strengths of prior approaches, we developed a deep learning model, eDICE (epigenomic Data Imputation via Contextualized Embeddings), based on framing the epigenomic imputation problem as one of masked input reconstruction. During training, a random subset of the observed signal values for a set of epigenomic tracks at a single genomic position are masked out, and the model is tasked with learning to impute the masked values given the remaining observed values. Unlike in standard masked input reconstruction applications, the epigenomic imputation problem requires models to be capable of predicting signal values for tracks never seen during training, representing novel combinations of cell type and assay. To achieve this combinatorial form of generalization, eDICE encodes the input signal at the genomic position of interest into separate latent representations summarizing the local epigenomic state of each cell type and the local activity profile of each assay. The signal value in a masked track is then reconstructed by concatenating the representations for the relevant cell type and assay and passing them through a multi-layer perceptron (MLP) decoder. At test time, predictions for new tracks can be generated in the same way, by feeding the model with the signal values of a set of observed tracks at a genomic location of interest, and decoding from the representations of the target cell type and assay.

To implement the factorized encoding of local epigenomic signals, we developed a self-attentive neural network module (Figure 4.1), based on the Set Transformer architecture [95]. This module starts by independently encoding the signal in each cell type and each assay, then constructs contextualized representations of each cell type and each assay by transferring information among related cell types and related assays using self-

attention. By conditioning on observed signal values to build representations of the local epigenomic state, rather than learning location-specific embeddings, eDICE achieves the generalization capacity of tensor factorization models while offering substantial improvements in both training efficiency and performance. A full description of the architecture and training procedure are provided in the Methods section.

4.2.2 eDICE imputations are highly accurate on the reference epigenomes

For direct comparison with previous imputation work, we evaluated the accuracy of eDICE imputations on a dataset of epigenomic tracks collated by the Roadmap project [234] and used in previous studies [395–397]. This dataset consists of 1014 signal tracks from 24 epigenomic assays in 127 cell types. All but one of the assays target histone modifications, with the remaining assay profiling chromatin accessibility via DNase-seq. A core set of five assays (H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3) is available in most cell types, while coverage of the cell types with the remaining assays varies widely. We used the first train/test split defined by Durham et al. [396], which consists of 709 training tracks, 102 validation tracks, and 203 test tracks (Supplementary Figure A.1 and Supplementary Section A.3.1). To compare the performance of imputation methods, we report a series of metrics assessing the quality of imputations of the tracks in the test set by models trained on tracks in the training and validation sets (and optionally using these tracks to provide inputs at test time). The metrics are computed across chromosome 21 of the hg19 assembly, the smallest human chromosome, spanning approximately 48 million base pairs. As baselines, we report results for the prior methods ChromImpute, PREDICTD, and Avocado (Section 4.4.6). Finally, as a parameter-free baseline, we also report predictions made by averaging the signal of the target assay in all other cell types in the training data except the target cell type (AVG).

Previous studies of imputation methods have varied in the choice of the primary metrics by which to assess performance [395–397]. In an attempt to provide a balanced view of model quality, we report performance on a selection of metrics designed to capture three desirable characteristics of imputations: (i) global similarity between imputations and ground truth values (ii) similarity between imputations and ground truth values focusing on foreground (Fg) and background (Bg) bins, as determined by MACS2 [399] and (iii) discriminative accuracy for a peak vs non-peak classification task. The first two categories are assessed using mean-squared error and Pearson

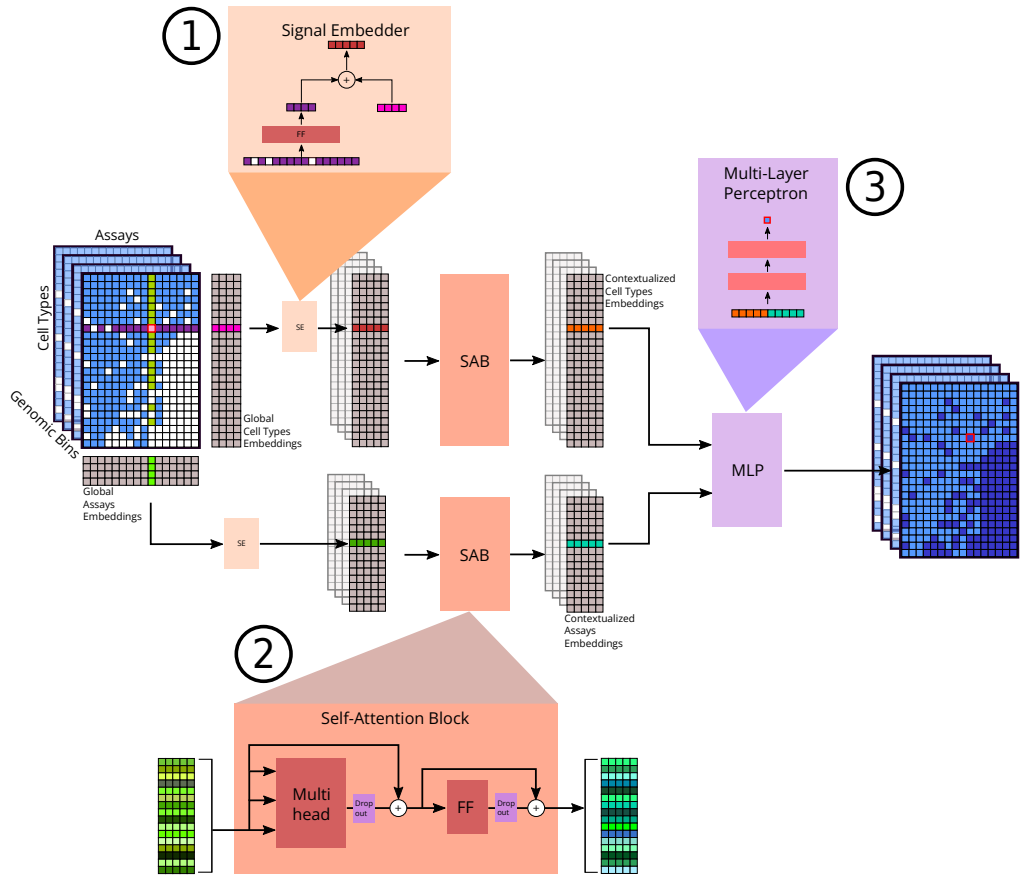


Fig. 4.1 **Schematic representation of the eDICE model.** For each cell type, we collect all measured signal values from assays performed in that cell type at the target bin, and project this set of values into a shared embedding space, where it is combined with a global embedding representing the cell type (1). We do likewise for assays, projecting the sets of values measured in different cell types from each assay into a distinct embedding space. We then apply self-attention over both sets of embeddings separately, allowing the network to capture relationships between cell types and between assays to produce contextualized latent embeddings which are functions of the local signal values in all observed tracks (2). Finally, a feed-forward neural network combines the contextualized embeddings for a target cell type-assay combination to generate a prediction for the local signal value (3).

correlation on the arcsinh-transformed signal values, while in the latter category we measure peak classification performance via the threshold-agnostic area under the precision-recall curve (AUPRC), as well as by precision and recall after calling peaks on the imputations using MACS2. Additional details on all metrics are found in Methods and Supplementary Section A.3.2.

The performance of the models is presented in Figure 4.2 and Supplementary Figures A.2-A.9, with numeric values reported in Supplementary Table A.2. eDICE outperforms PREDICTD and Avocado across all metrics, and ChromImpute across the majority, although ChromImpute shows strong performance for the prediction of peak height in the foreground (Figures 4.2 and 4.4, Supplementary Figures A.2 and A.3). eDICE's relative disadvantage here suggests a tendency to systematically underestimate the absolute signal values within peaks, which is exemplified in the trade-off between precision and recall compared to ChromImpute. However, it ranks peak and non-peak regions relative to each other more accurately than ChromImpute, as demonstrated by the fact that it outperforms all baselines on the AUPRC metric, thereby offering the best overall imputation in terms of global discriminatory power. We emphasize that while PREDICTD and Avocado generated imputations respecting the same data split used for eDICE, the ChromImpute imputations were produced in a leave-one-out fashion, so our model's improved performance comes despite a considerable handicap relative to ChromImpute in terms of the available training data.

Qualitatively, eDICE presents many of the characteristics that were present in its predecessors, such as a general smoothing of the imputed tracks, which is especially notable in the background regions (example in Figure 4.3). Additionally, the imputed tracks reduce the impact of outlier values, such as the extremely high peaks present in a few tracks for H3K4me3. Such peaks are not necessarily a direct representation of the high significance of the local enrichment but can be heavily affected by the coverage and quality of control samples, which, when low, can bias the estimated p-values towards extreme values (see further discussion in Supplementary Section A.3.1).

To confirm that these aggregate results were not unduly influenced by variation in the range of metric values across different types of assay, we also examined the metrics at the level of individual tracks (Figure 4.4 and Supplementary Figures A.4-A.6) and aggregated by assay (Figure 4.5). The track-level comparisons confirm that eDICE's performance improvements are consistent across different combinations of cell-type and assay. Grouping tracks by assay reveals significant differences in the performance of imputation methods depending on the type of epigenetic mark. For example, all models

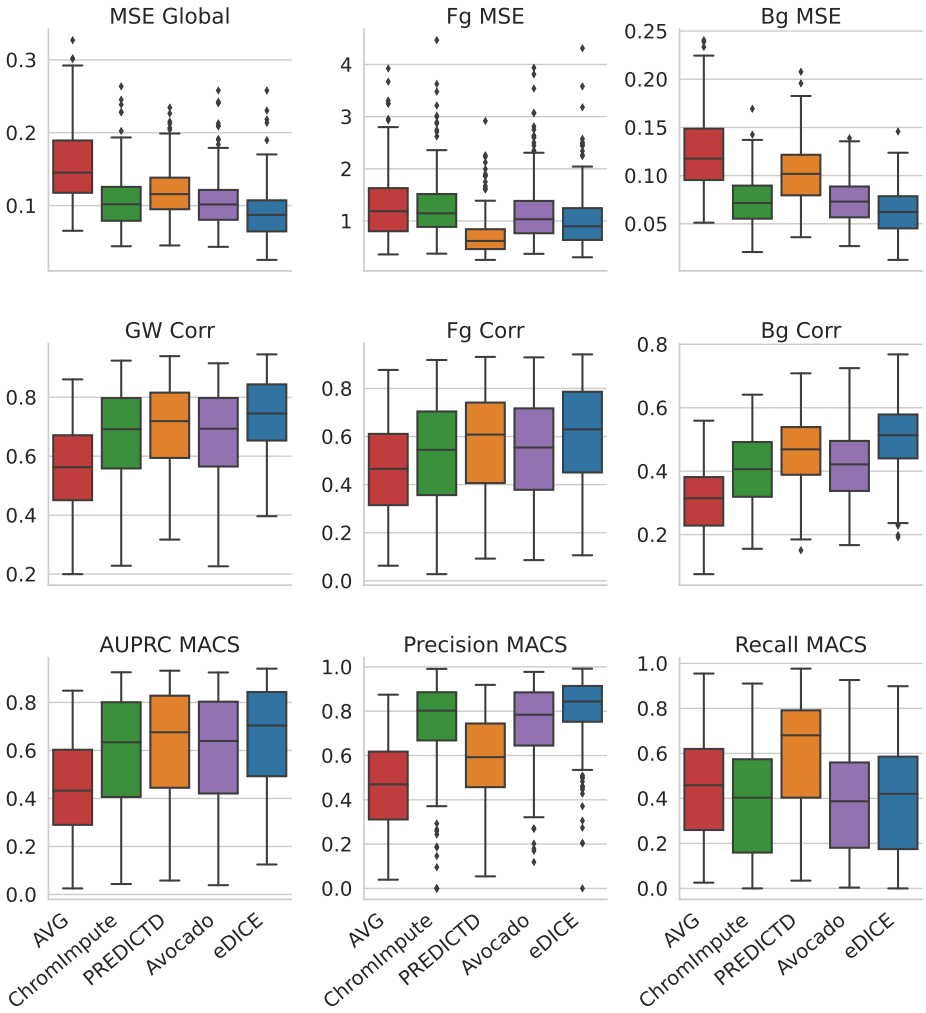


Fig. 4.2 **Comparison of imputation methods on the Roadmap reference epigenomes.** Performance metrics for the imputation of the n=203 test tracks on chromosome 21 for each model. Boxes represent the interquartile range (IQR), with the middle line representing the median; the whiskers represent points that lie within 1.5 IQRs of the lower and upper quartiles, while remaining outliers are explicitly displayed. Metrics presented include mean squared error (MSE) and Pearson correlation coefficient (Corr) for the genome-wide (GW/Global), foreground (Fg) and background (Bg) regions, as well as the area under the precision-recall curve (AUPRC), precision, and recall for the classification of peaks detected with MACS2.

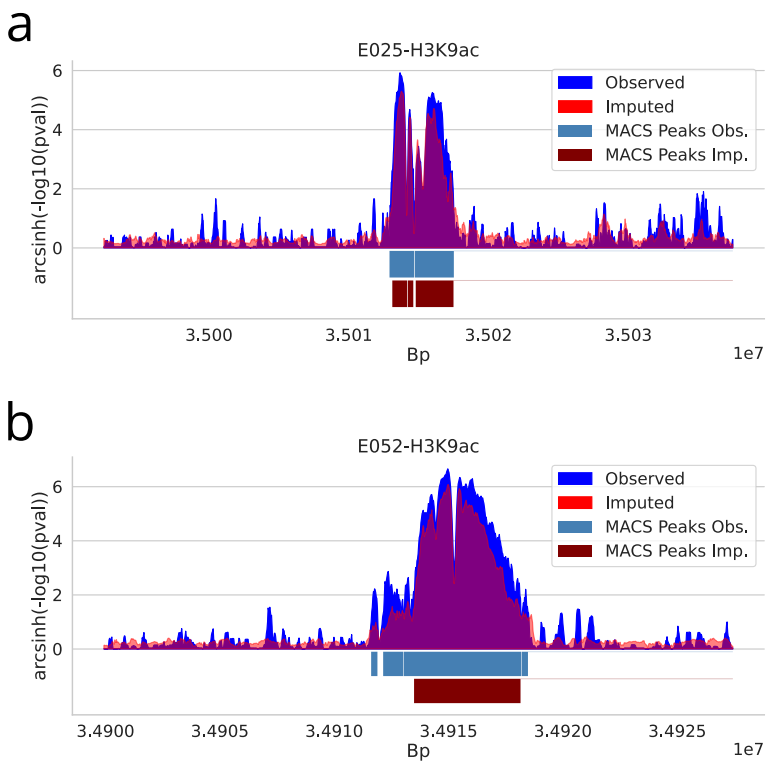


Fig. 4.3 **Signal reconstruction using eDICE.** Examples of observed epigenomic tracks with the signals imputed by eDICE for the assay H3K9ac in two selected tissues (E025 **(a)**, E052 **(b)**). Below the tracks, the peaks detected with MACS2 highlight how the imputations accurately capture enriched regions. The peaks were detected using a one-sided Poisson hypothesis test with Benjamini-Hochberg correction for multiple test correction and a cut-off value of 0.01.

tend to perform relatively poorly when predicting H3K27me3 and H3K9me3 (Figure 4.5 and Supplementary Figures A.7-A.9). Comparing the average assay-level performance of each model shows that the improvements brought by eDICE are consistent across the board despite these discrepancies between assays (Supplementary Figure A.2).

Finally, we explored whether differences in performance between types of assays could be related to differences in specific properties of the epigenetic marks. Some histone modifications can be classified as either narrow-peak (H3K27ac, H3K4me2, H3K4me3, H3K9ac) or broad-peak marks (H3K27me3, H3K36me3, H3K4me1, H3K79me2, H4K20me1). Comparing the performance of eDICE on test tracks across these two groups, we observed that performance tended to be higher on narrow-peak than on broad-peak marks for correlation and classification metrics (Figure 4.6a). Furthermore, a similar divide is observed when splitting histone modifications into repressive (H3K27me3, H3K9me3) and activating marks (the active promoter-associated H3K9ac, H3K4me2, H3K4me3, active enhancer-associated H3K4me1 and H3K27ac and DNase-seq, displayed in Figure 4.6b). As repressive marks are often linked to heterochromatin configurations, this discrepancy is possibly due to biases introduced by the processing pipelines because of systematic sequencing differences in these regions. However, as repressive marks also tend to display broad peaks, it is challenging to pinpoint the precise reason for the observed differences.

Importantly, the performance benefits of eDICE are coupled to an increased efficiency in the training procedure. Relative to ChromImpute, this is a result of training only a single model rather than a separate ensemble of models for each target track. Relative to Avocado and PREDICTD, eDICE can make accurate genome-wide predictions without needing to train on every genomic location, leading to major improvements in training efficiency. To highlight this, in Figure 4.7, we show the results of training eDICE on smaller subsets of the randomly selected genomic locations. Even when trained on a small fraction of the available genomics data, eDICE outperforms Avocado, suggesting that the tensor factorization models severely overparameterize the imputation problem by learning a representation for each genomic bin.

4.2.3 Imputations capture significant differences between tissues

Epigenomic patterns differ between cell types to control and register cell function and identity. It is critical that imputations accurately capture these differences if they are

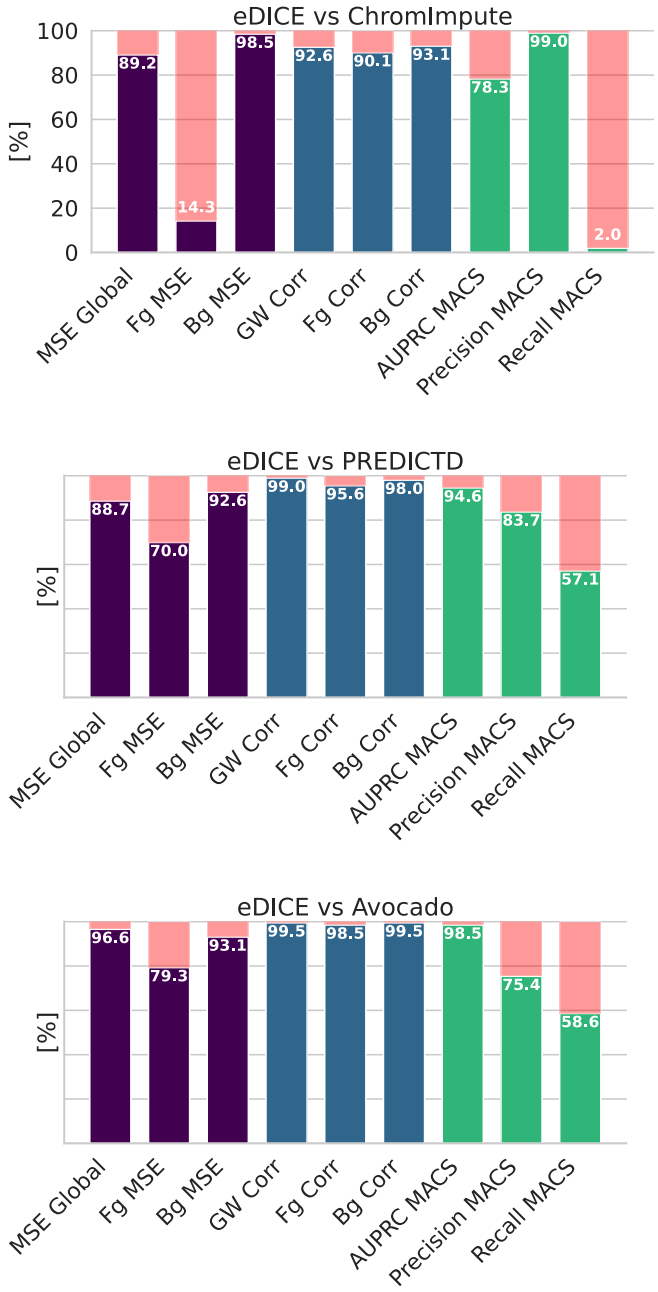


Fig. 4.4 **eDICE outperforms competitor models on most metrics.** Percentages of test tracks on which eDICE outperforms the baselines for each metric. ChromImpute shows good performance on tasks related to the height of the peaks, while eDICE outperforms PREDICTD and Avocado on all metrics.

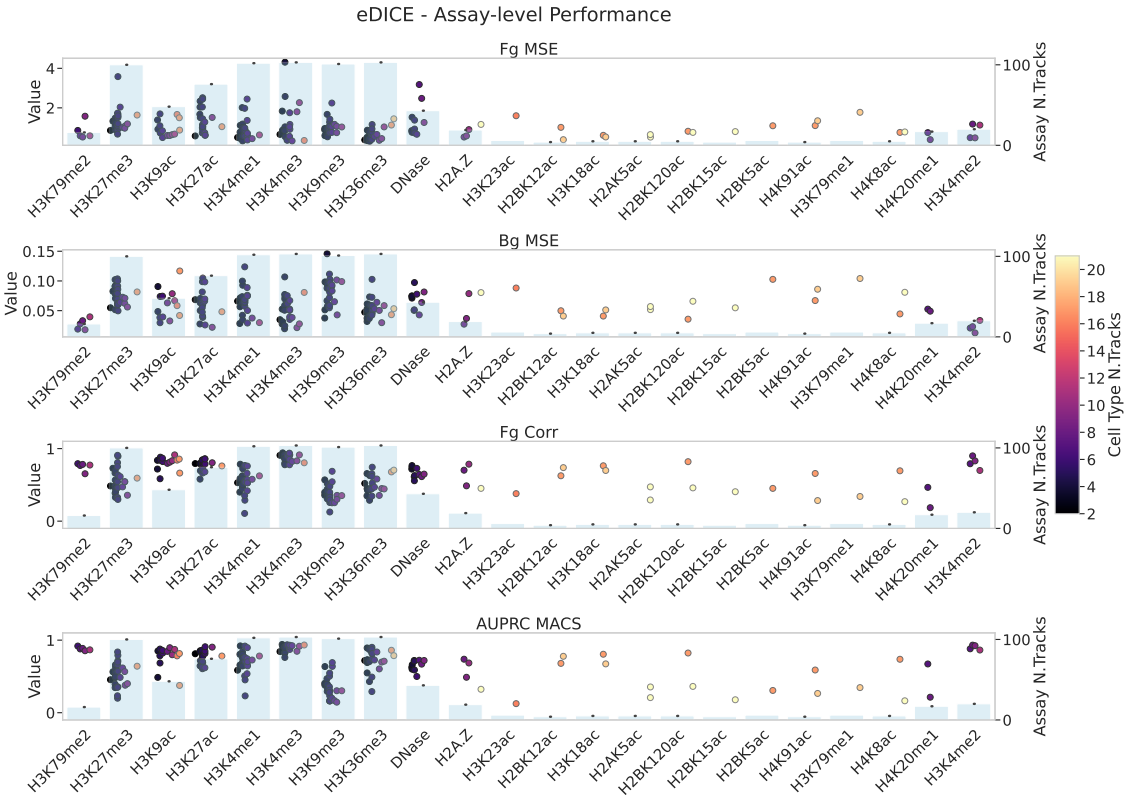


Fig. 4.5 **Imputation performance varies significantly between different assays.** Grouping the tracks by assay reveals considerable differences in the imputation performance. This phenomenon is observed in the previous models as well, indicating that it is most likely due to the nature of the specific modifications and the biases that their signal includes. The colour of each dot indicates the number of training tracks that share the cell type with that specific test track, while the light blue bars in the background show the number of training tracks that share the same assay.

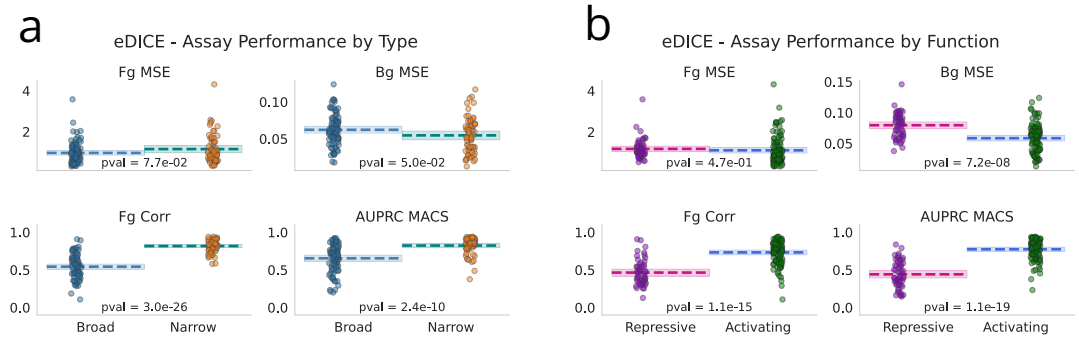


Fig. 4.6 **The type and function of histone modifications affects imputation performance. (a)** Assays split into broad- and narrow-peak marks show consistently different performance for the imputation task. For each metric, we performed a 2-sided Welch's t-test under the null hypothesis that both sets of metrics have the same mean, and reported the resulting p-value at the bottom of each plot. **(b)** Splitting the histone marks by functionality (repressive vs. activating) shows a similar bias as the comparison in (a). For each metric, we performed a 2-sided Welch's t-test under the null hypothesis that both sets of metrics have the same mean, and reported the resulting p-value at the bottom of each plot.

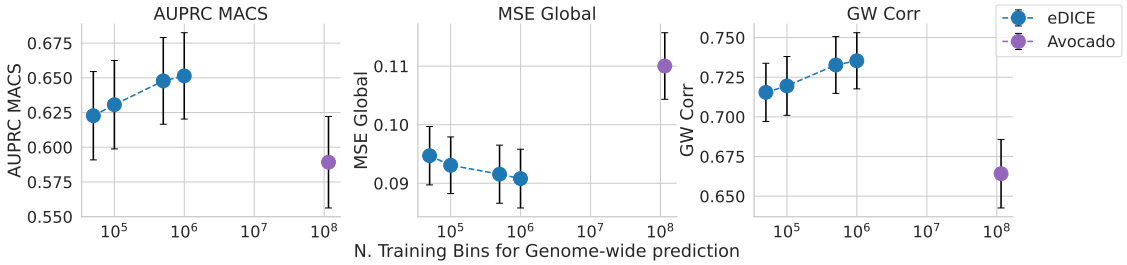


Fig. 4.7 **eDICE can be trained on a smaller set of genomic regions.** Learning curves that display several global performance metrics against the number of genomic positions used in training. Tensor factorization models such as Avocado need to be trained on the whole genome to make genome-wide predictions. eDICE, on the other hand, can be trained efficiently on a small subset of genomic regions and still obtain improved performance, suggesting that previous models severely overparameterized the imputation problem. Data are presented as mean ± 95% confidence interval for n=203 test tracks.

to constitute valuable resources of cell-type-specific epigenomic landscapes. However, within the scale of the whole genome, these cell-type-specific differences are subtle and global evaluation metrics such as those considered above are dominated by regions that have a shared functionality across cell types, such as large intergenic regions.

In the analysis of epigenetic modifications, it is crucial to capture not just a single instance of the local signal measured by experimental assays, but also the local variability which may characterize each tissue. To distinguish potentially functional differences from either technical or biological fluctuations, established experimental protocols explicitly require several biological replicates to estimate local variability. This is essential for robust statistical hypothesis testing [400]. On the other hand, experimental tracks are generally pooled for imputation tasks, and predictions thus constitute mean epigenomic tracks per cell type, where the inherent variability is lost. We present here a case study in which we estimated local variability on the training data and then generated simulated replicates from the mean epigenomic imputation. This strategy was used to predict and identify differential peaks in H3K9ac tracks across two tissues (corresponding to Roadmap cell types Adipose-Derived Mesenchymal Stem Cell Cultured Cells (E025) and Muscle Satellite Cultured Cells (E052)).

We highlight that the overall shape of individual peaks is remarkably conserved between individual experimental replicates for corresponding tracks (Figure 4.8a). In the case of tissue-specific peaks, on the other hand, the signal shapes are distinct between replicated measurements derived from different tissues (Figure 4.8b). We have previously exploited this observation for differential peak calling [401], where we considered the genomic region of the peak as a metric space and treated the pile-up of sequenced reads like a sample from a hidden probability distribution on that space. This strategy dramatically improves the test's statistical power compared to methods based on total counts alone. We also note that the shape differences are well captured by the mean signals (Figures 4.8a and 4.8b bottom panels). To quantify differences in peak shapes across the two cell types, we computed the Wasserstein (WS) distance between the pooled ground truth signals across the two cell types, and likewise between the imputed signals. Figure 4.8c shows that the distances in imputed and ground truth tracks are strongly correlated, indicating that the imputations accurately capture cell-type-specific differences in the shape of signal enrichment at peak regions.

As an independent analysis, we next took advantage of the robustness of the existing differential peak analysis method, DiffBind [402]. Since DiffBind requires replicates for statistical testing, we estimated the local variability of cell-type-specific

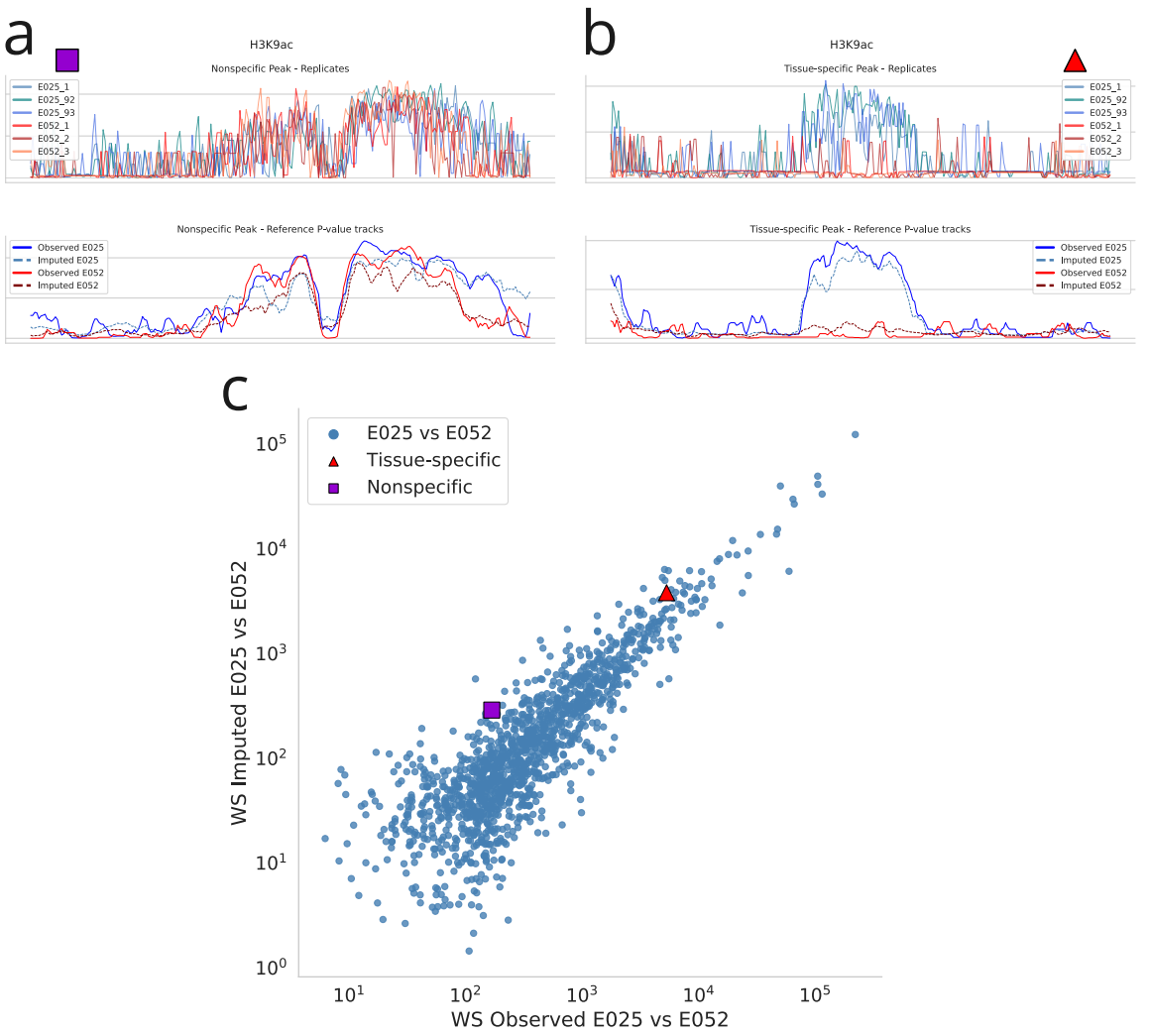


Fig. 4.8 **Imputations with eDICE capture variation between tissues. (a)** and **(b)** show examples of non-specific and tissue-specific peaks respectively for H3K9ac in the two chosen tissues (E025 and E052). The upper part shows the measured replicates, while the lower portions display the aggregate p-value tracks for the observations and the corresponding imputations. The aggregate tracks do not capture the information on the biological variability between samples. **(c)** A scatter plot of the Wasserstein distance between the signal in the two tissues, for each peak in the enriched peakset of E025. The x-axis displays the WS distance between observed signals, while the y-axis between imputed signals. The imputations retrieve most of the information contained in the measurements, especially for the stronger differences between tissues. We highlighted the two points corresponding to the peaks shown in (a) and (b).

test tracks (Supplementary Section A.3.4). Assuming a negative binomial distribution, the estimated variance parameters were subsequently used to simulate replicates from the imputed mean signal tracks on chromosome 21. While an arbitrary number of replicates can readily be generated in this way, we chose to use three to four simulated replicates, similar to typical experimental scenarios. Those tracks were fed into the standard differential analysis pipeline, and the outcome was compared with the results obtained from the corresponding analysis of actual replicated measurements. We emphasize that the simulation procedure employed only replicates from the training set and tissue-specific control samples in addition to the imputed tracks, and made no use of any information from the test set.

Employing the DiffBind library [402] we compared binding affinity scores, which are indicative of the strength of interaction between DNA and biomolecules (such as modified histones). Figure 4.9a shows a correlation heatmap for the similarity of affinity scores for different samples. The block structure highlights the expected relationship between the replicates derived from different tissues; however, the simulated replicates show high similarity across tissues, possibly due to the adopted procedure underestimating the biological variance between samples.

Within DiffBind, we used DESeq2 to identify peaks of differential enrichment with default parameters. Specifically, we used a 'glmGamPoi' fit type to estimate dispersion and used a Wald test for negative binomial distribution ('nbinomWaldTest') to identify statistical significant peaks. A total of 1165 and 1299 peaks were detected as differentially enriched in measurement and imputations, respectively (FDR threshold of 0.05). 855 peaks ($\sim 73\%$ of the measured peaks) are shared between the two sets, resulting in a Positive Predictive Value of 0.66 (Figure 4.9b). Binding affinity scores for each differentially enriched peak in the consensus peakset derived from imputations and measurements are shown in Figure 4.9c, where the block structure resulting from agglomerative clustering of the measurements (left side) is replicated in the imputations (right side).

The differential analysis procedure was repeated for all the models analysed, with eDICE outperforming Avocado and ChromImpute; the model PREDICTD showed comparable performance (Supplementary Figure A.23). In summary, we conclude that the imputations accurately capture cell-type-specific differences, both in terms of altered shapes of signal enrichment at peak regions and also regarding integrated total counts in the peak regions, when considering local variability. In general, increasing the number of replicates in a sequencing experiment leads to more robust results [403].

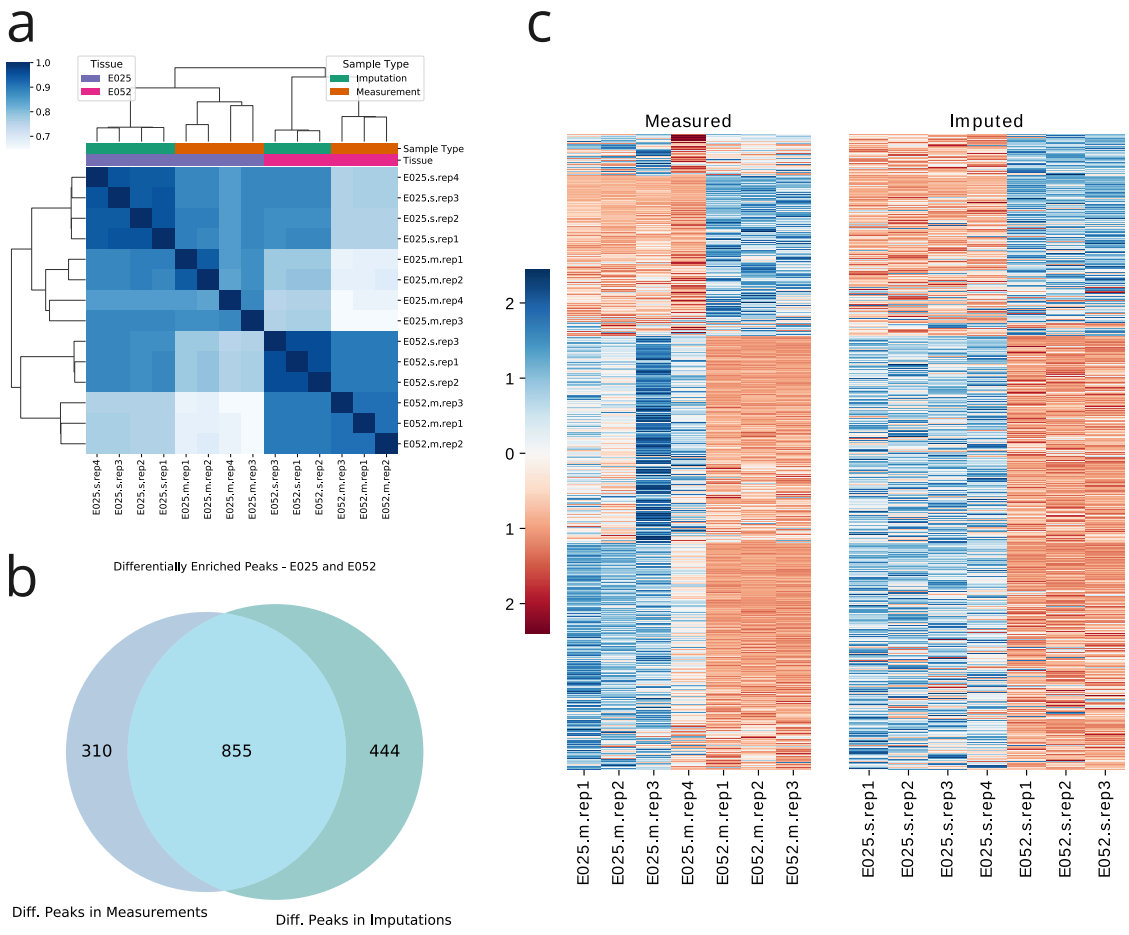


Fig. 4.9 **Differential peak analysis using imputed epigenomic tracks.** **(a)** Correlation heatmap of the affinity scores for different replicates. The simulated replicates correctly retrieve the expected relationships to the measured replicates, although they show a high degree of similarity between themselves, likely an artefact of the simulation procedure. **(b)** Venn diagram representing the peaks that are detected as differentially enriched between tissues using imputed and measured signals. The imputed signal retrieves 66% of the true peaks. **(c)** Binding affinity heatmaps for the measured replicates and the imputed pseudo-replicates. Each row corresponds to one of 1609 differentially enriched peaks detected in either of measurements and imputations. The imputed replicates display the same global block structure as the measurement replicates.

Therefore, a similar augmentation strategy could also be applied to complement certain existing experimental data sets with additional replicates from an imputed mean track.

4.2.4 eDICE accurately predicts personalized epigenomes in unseen tissues

Recent advances have highlighted the role that alterations in the epigenetic machinery play in human disease [404–406]. In the field of precision medicine, epigenetic mutations are currently examined mainly for their potential role in early detection and drug response prediction [407–409]. However, increasingly robust epigenome editing methods [382] open up exciting opportunities for direct interventions on the epigenome for the treatment of illnesses such as cancer [384]. Achieving a more in-depth understanding of individual- and cell-type-specific epigenetic patterns and their effect on the cellular machinery will be crucial to realizing the promise of such applications.

Recently, a collaboration between the ENCODE [388, 410] and the Genotype-Tissue Expression (GTEx) consortia created data sets that include extensive individual-specific histone modification measurements from four donors [398]. We decided to use this dataset to test whether eDICE could be applied to impute epigenomic measurements in an individual-specific manner. One particular use case for imputations in this setting could be to predict epigenomic measurements in otherwise hard-to-access tissues, potentially avoiding the need for invasive procedures. Motivated by this use case, we developed a task to test the prediction of epigenomic measurements in a particular individual in tissues for which no epigenomic information for that individual is available. Specifically, we aim to impute epigenomic tracks for a target tissue in one individual patient ("target individual"), by using other observations from the same individual, as well as a more complete set of observations for another individual ("training individual"), which include the target tissue. To adapt eDICE to this task, we adopt a transfer learning approach. We first train an eDICE model on the complete set of observations for the training individual. The model is then fine-tuned on the set of observations for the target individual that do not include the target tissue, before imputing the target observations (Figure 4.10). We employed an eDICE model with the same architecture as that used for Roadmap, but altered the masking process used during training to reflect the tissue-based prediction task, ensuring that the set of masked tracks at each genomic bin all belonged to a single randomly selected tissue.

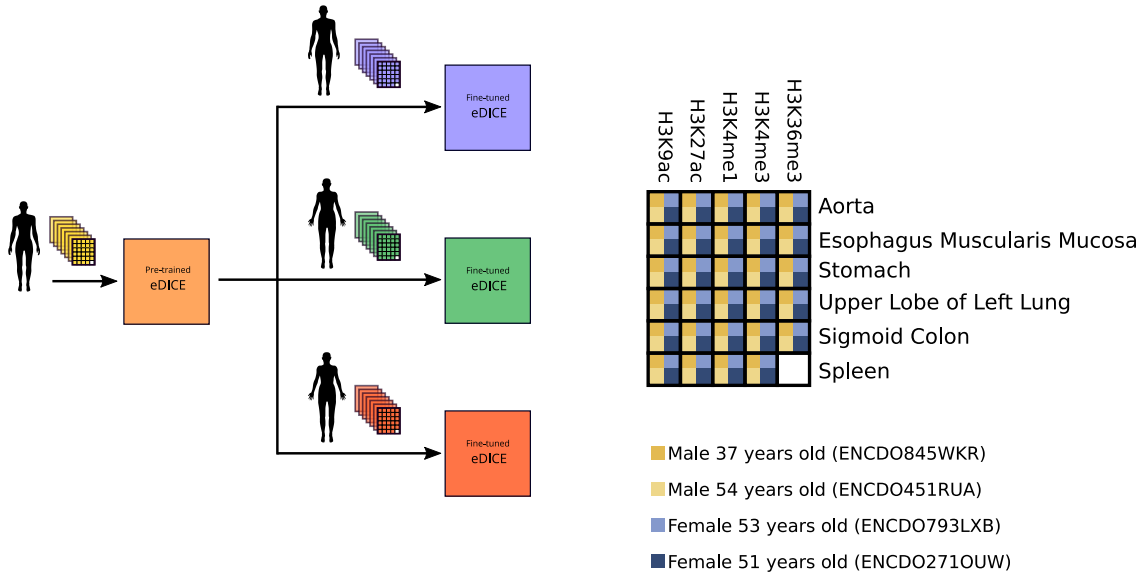


Fig. 4.10 **Transfer learning scheme for the imputation of unseen tissues in the EN-TEx dataset.** The full set of tracks for a training individual are used to pre-train an eDICE model so that it can learn to appropriately encode all the tissues represented in the dataset. Afterwards, the model is finetuned on a subset of tracks from the target individual that exclude a target tissue. Finally, the model is used to predict the epigenomic signal for the target tissue in the target individual.

To get a better understanding of this task, we performed an initial analysis of epigenomic variation between individuals in the EN-TEx dataset. The data includes measurements spanning 25 different tissues from two adult males, 37 and 54 years old, and two adult females, 53 and 51 years old. We selected for further study 29 tissue-assay combinations comprising measurements of histone modifications available for all four individuals (Supplementary Table A.3), focusing on chromosome 21 in all cases. Initial analysis of observed tracks revealed both a large degree of similarity in epigenomic signal across individuals in numerous instances (Figure 4.11), as well as the dominant role of tissue identity in determining epigenetic patterns, in particular for marks H3K27ac, H3K4me1, and H3K9me3 (Figure 4.12a).

Individual-specific peaks unique to only one or a subset of individuals are nonetheless observed, most notably for H3K9me3 (Figure 4.12b). Three-dimensional histograms of co-occurrences across tissues and individuals highlight that across all marks individual-specific peaks are typically also specific to one or a small number of tissues, and that the frequency of such peaks varies substantially between marks (Figure 4.12c and Supplementary Figures A.17-A.21). These personal epigenomic differences may either reflect underlying DNA sequence variants, in which case they may be observable across

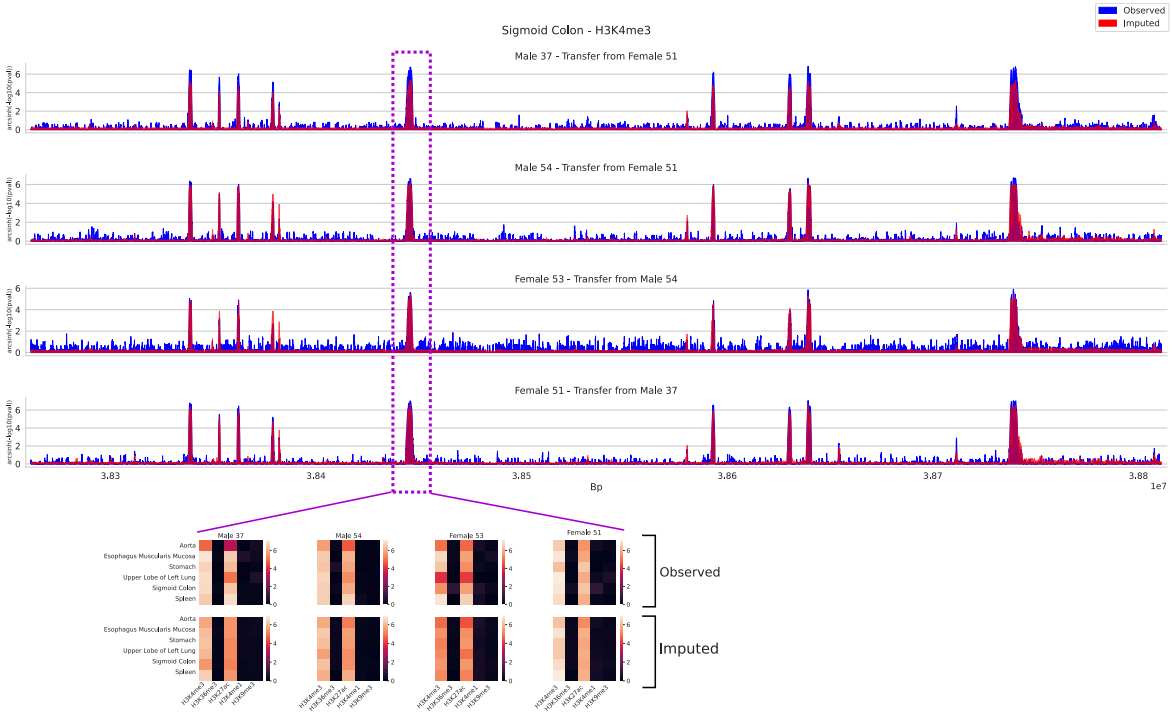


Fig. 4.11 **Patterns of epigenetic modifications are largely shared across individuals.** Sigmoid Colon-H3K4me3 track spanning 800kb and showing consistent patterns for all four individuals. For the central peak we display a slice across the epigenomic tensor demonstrating signal conservation across tissues and individuals.

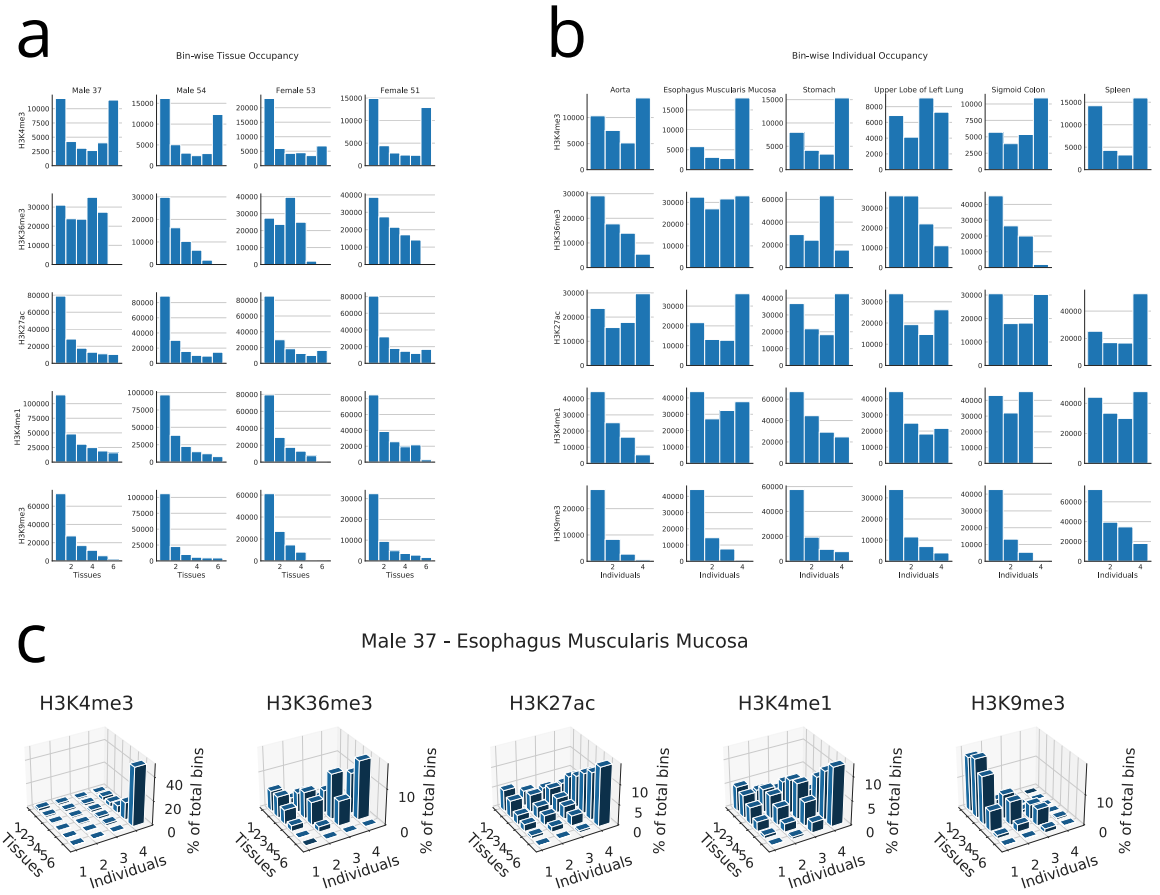


Fig. 4.12 **Occupancy histograms across tissues and individuals reveal high variation in the data.** **(a)** Occupancy histograms for the enriched bins across tissues. **(b)** Occupancy histograms for the overlap of enriched bins across individuals. **(c)** Occupancy across tissues and individuals for each enriched bins in the tracks for Male 37 for the Esophagus Muscularis Mucosa tissue.

different tissues of the same individual, or they may result as a consequence of ageing or due to interactions with external stimuli, potentially in a tissue-specific manner.

We next assessed the accuracy of eDICE imputations generated using the transfer learning scheme described above. We compared these predictions to imputations from two model-free baseline methods. The first method directly uses the corresponding track in the training individual as a prediction of a given track in a target individual. The second method generates a predicted track by averaging the tracks from the target assay in all tissues apart from the target tissue in the target individual (i.e., an individualized version of the AVG baseline used previously). For each method, we consider all possible combinations of target tissue, target individual, and training individual, and evaluate the resulting predictions. Using the transfer learning strategy

presented, eDICE produces imputations which are globally more accurate than either of the baseline methods, as measured by MSE and Pearson correlation (Figure 4.13), indicating that transfer learning successfully adapts eDICE to the context of a new individual, while retaining the understanding of tissue types inherited from the training individual to allow successful prediction in tissues without measurements in the target individual.

4.2.5 eDICE captures epigenetic variation between individuals

Finally, we used the same transfer learning framework to assess eDICE's ability to predict individual-specific epigenomic signatures. Defining such signatures is far from trivial; a robust analysis would require more than four individuals to properly understand the overlap of enriched regions and the external factors that influence them. As a working approximation, we define individual-specific peaks as those enriched regions detected from the measured samples that span at least 150bp (i.e. the approximate length of the DNA wrapped around a nucleosome) and which are present only in the target individual, and not the training individual. This definition aims to capture peaks such as the example shown in Figure 4.14a, where H3K4me3 is clearly found in one individual. This task presents significant challenges due to the relatively small portion of epigenetic enrichments that meaningfully differ between individuals and because of the complex epigenetic patterns that arise in these regions of variability, exemplified by the heatmaps displayed in the lower portion of Figure 4.14a. In these cases, local variability is observed not just between individuals, but also between tissues within the same individual (Figure 4.14b).

To assess the ability of eDICE to predict these individual-specific peaks, we compared imputations to test tracks after excluding enriched regions shared with the training individual. We additionally excluded enriched regions specific to the test individual but spanning less than 150 bp. Within the remaining regions, we assessed the extent to which the imputations successfully distinguished individual-specific peaks from the background using the area under the precision-recall curve (AUPRC). For completeness, we included the fraction of positive samples (Pos. Fraction) as a standard baseline for the AUPRC measure [411]. The results, presented in Figure 4.15a, show that eDICE improves the prediction of individual-specific enrichment compared to the model-free baselines. A track-level comparison of eDICE's improvement over the model-free baselines is shown in Figure 4.15b.

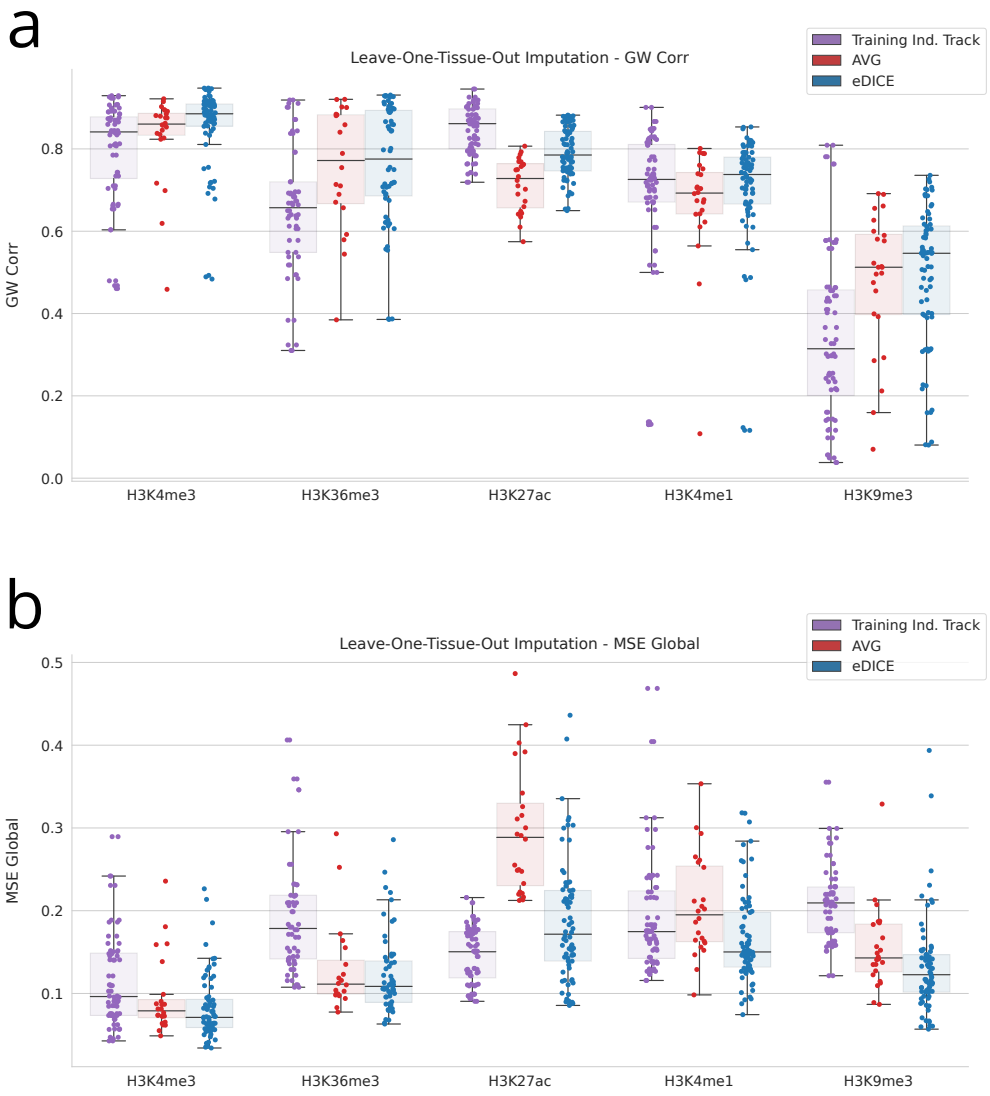


Fig. 4.13 **Imputation performance on the whole ENTEx tracks.** Pearson correlation **(a)** and MSE **(b)** for the leave-one-tissue-out imputation of chromosome 21 using transfer learning from one training individual to the target individual. n=72 imputed tracks for the 'Training Ind. Track' baseline and eDICE for each assay except H3K36me3, where n=60. n=24 for the "AVG" predictor for each assay except H3K36me3, where n=20. Boxes represent the IQR, with the middle line representing the median; the whiskers represent points that lie within 1.5 IQRs of the lower and upper quartiles.

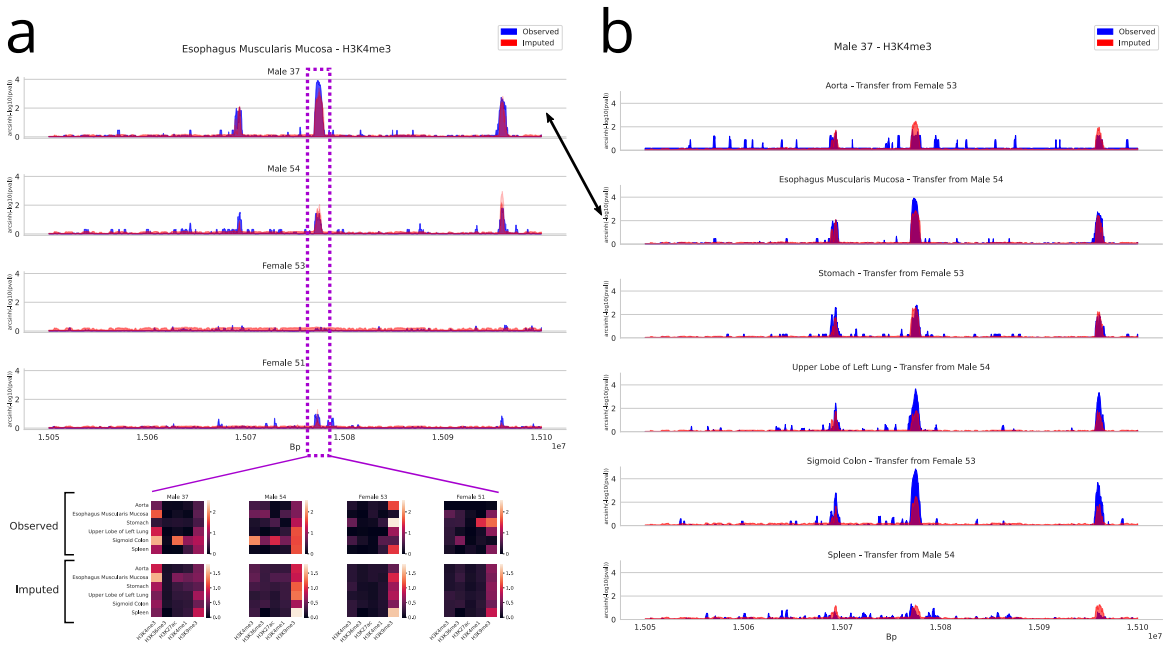


Fig. 4.14 **Individual-specific peaks.** **(a)** Individual-specific H3K4me3 enrichment for (Male 37) in Esophagus Muscularis Mucosa tissue. For this genomic location, we display a slice of the epigenomic tensor for each of the four individuals, highlighting the challenge of imputing these varied patterns. Peaks were detected with MACS2 using a one-sided Poisson hypothesis test with Benjamini-Hochberg correction for multiple test correction and a cut-off value of 0.01. **(b)** Observed and imputed tracks for the H3K4me3 assay in Male 37 across tissues in the same genomic region as (a).

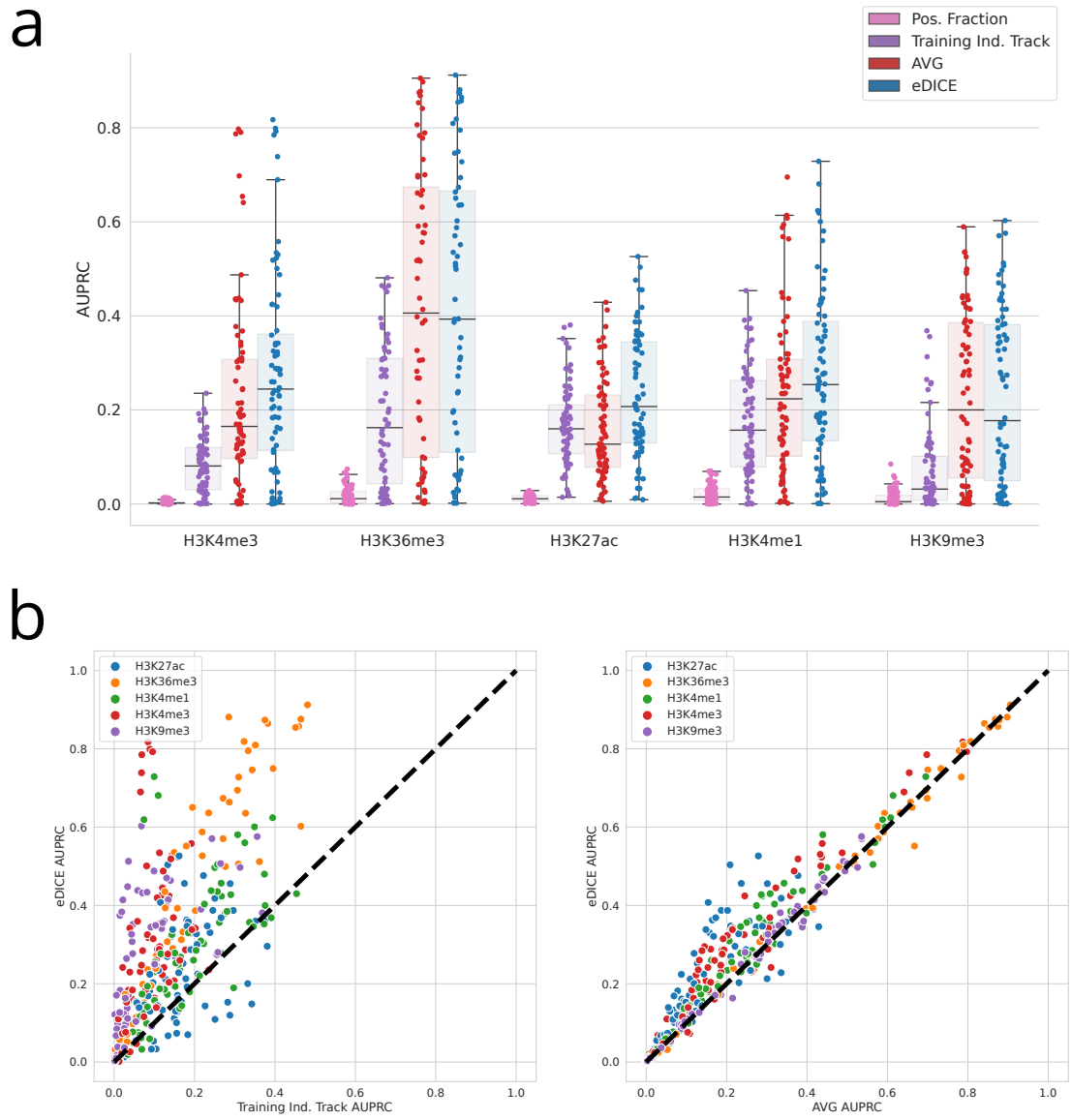


Fig. 4.15 **Imputation for precision epigenomics (a)** AUPRC for the prediction of individual-specific enrichment in the LOO EN-TEx imputations, where the peaks shared with other individuals have been masked out. n=72 imputed tracks for each model for all assays except H3K36me3, where n=60 for each model. Boxes represent the IQR, with the middle line representing the median; the whiskers represent points that lie within 1.5 IQRs of the lower and upper quartiles. **(b)** Track-level AUPRC for the prediction of individual-specific enriched bins.

Capturing individual-specific differences is crucial for the robust application of state-of-the-art machine learning models to epigenetics within a clinical context. The case study presented aims to be a guiding example for the development of better and more accurate models that may be included in a clinical workflow.

4.3 Discussion

We presented eDICE, a deep-learning-based epigenomic imputation framework which achieves high accuracy by combining the advantages of its predecessor models. Like ChromImpute, eDICE uses the local signal of observed tracks to encode information on genomic position, removing the need to learn explicit embeddings for each position. Similar to the tensor factorization models PREDICTD and Avocado, eDICE uses factorized representations to achieve combinatorial generalization, while drastically reducing the required parameter count (Supplementary Table A.1). On reference epigenomes, eDICE's performance is robust across a variety of metrics capturing different facets of imputation performance, surpassing all baselines across the majority of metrics, while offering significant practical benefits as a simple single-model approach that is efficient to train and run.

We emphasize the need for imputation models to be trained and designed with the aim of including imputations in established bioinformatics processes. As a case study, we explored the possibility of simulating biological replicates from the imputed data, which are then used for differential peak calling obtaining results compatible to the measured replicates. We pose that future developments in the field of epigenomic imputation should account for and predict not only the average value of measurements, but also the intrinsic biological variability of different samples. Explicitly modelling the variance of epigenomic measurements would allow for more robust analysis to distinguish the differences caused by fluctuations due to the natural variability of the samples from the true differences between tissues and marks that encode the functional variations of cell profiles.

Finally, we demonstrated the possibility of imputing personalized epigenomic tracks, showing how eDICE can be adapted to generate imputations for unseen tissues that outperform those from model-free baselines. The transfer learning approach adopted allows the model to learn representations for all tissues from a training individual, which can then be transferred to a target patient, enabling accurate imputations in tissues where no data for the target individual is available. While our results offer a proof of

concept for the direct applicability of methods originally designed for the imputation of reference epigenomes in this setting, individual-specific imputation presents additional challenges which future works might seek to address directly. In particular, further enhancements in accuracy might be unlocked by incorporating information from DNA sequence, as well as information aggregated across other individuals and from reference datasets to augment the relatively limited number of measurements in any single individual. In order to fully leverage the promise of transfer learning across epigenomic datasets, further consideration should be paid to the role of systematic biases introduced by differences in experimental methodology and bioinformatics pipelines used to process sequencing data, as highlighted by the findings of the ENCODE imputation challenge [328], to which we submitted a prize-winning entry using a predecessor of eDICE. We believe that our results offer a strong indication that machine learning methods are well-placed to address these challenges, and, in so doing, to help overcome the experimental constraints that limit our understanding of epigenetic variation.

4.4 Methods

4.4.1 Data

A dataset of epigenomic measurements from the Roadmap Consortium [389] was selected to allow direct comparison with prior imputation methods. The Roadmap dataset consists of 1014 signal tracks from 24 types of epigenomic assay in 127 cell types. Each signal track is obtained by mapping a set of sequence reads to a genome to form a genome-wide activity profile. All but one of the assays target histone modifications, with the remaining assay profiling chromatin accessibility via DNase-seq. A core set of five assays, targeting H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3, is available in each cell type, while coverage of the cell types with the remaining assays varies widely. We use the first train/test split defined by [396], which consists of 709 training tracks, 102 validation tracks, and 203 test tracks. Supplementary Figure A.1 gives an overview of the data splits over training, validation, and test.

Following previous imputation work, we work with signals in the form of negative logarithms of p-values ($-\log_{10}$ p-value) tracks, which indicate the statistical significance of a mark at each genomic position, and seek to impute the average $-\log_{10}$ p-value within each non-overlapping 25 base pair interval in a given subset of the genome. We additionally preprocess the $-\log_{10}$ p-value signal using an arcsinh transform, which

reduces the impact of outliers and differences in distribution between different types of assay, again inspired by prior work [396, 397, 412].

The EN-TEx dataset contains the results of a variety of functional genomic assays in 25 tissue types from four donors (in the main text, for consistency with prior publications, we use 'tissues' to refer to biosamples in EN-TEx and 'cell type' to refer to biosamples in Roadmap; for the purposes of the imputation method the two terms should be treated as interchangeable). We selected 116 histone modification tracks common to all four individuals (Supplementary Table A.3). These tracks were processed in the same manner as those from Roadmap. The tracks measured for 'thoracic aorta' and 'ascending aorta' were merged to cover all four individuals.

4.4.2 Enrichment detection and evaluation metrics

We used MACS2 [399] to detect peaks in observed tracks, using a one-sided Poisson hypothesis test with Benjamini-Hochberg correction for multiple test correction and a cut-off value of 0.01. We refer to the 25-bp genomic bins belonging to the peaks detected by MACS2 for a given track as 'enriched bins' or 'foreground regions' for that track. Enriched bins detected in this way were used to define evaluation metrics for the tasks of both reference epigenome imputation and individual-specific imputation, as described in detail below and in Supplementary Section A.3.2.

Roadmap imputation metrics

The global quality of imputations was measured using the mean squared error (MSE) and Pearson correlation coefficient applied to imputed and ground-truth tracks. These metrics were also evaluated separately on foreground and background regions. Recovery of enriched bins (i.e. bins occurring in MACS2 peaks) was measured using the threshold-agnostic area under the precision-recall curve (AUPRC). Finally, MACS2 was applied to imputed tracks to generate a set of predicted peaks using the same fixed parameters as used to call peaks on the observed tracks. The resulting peaksets were then compared with the peaksets returned from the observed tracks using precision and recall.

Individual-specific imputation metrics

Global imputation performance was measured as above. For prediction of individual-specific peaks, we used the AUPRC to compare imputed tracks and MACS2 peaks, after

excluding all genomic regions containing peaks conserved across individuals involved in the transfer learning and spurious individual-specific peaks of less than 150bp.

4.4.3 Tensor factorization

Given a set of observed tracks that are the result of performing at least one of a set of n_a assays $(a_1, \ldots, a_{n_a})$ in each of a set of n_c cell types $(c_1, \ldots, c_{n_c})$, the goal is to generate imputations for all assay-cell type combinations which are not represented by tracks in the observed set. The complete set of possible measurements (all assays in all cell types at all genomic locations) can be represented as a rank-3 tensor $\mathcal{Y}$, with $\mathcal{Y}_{ijk}$ the signal observed at the k^{th} genomic position when performing the j^{th} assay in the i^{th} cell type.

Tensor factorization approaches model entries in the tensor as interactions between separate representations for each dimension. In PREDICTD and Avocado, learned cell type embeddings, $\mathbf{c}$, assay type embeddings, $\mathbf{a}$, and genomic bin embeddings, $\mathbf{b}$, are combined via a parametric function g_θ to reconstruct or impute tensor elements:

$$\hat{\mathcal{Y}}_{ijk} = g_\theta(\mathbf{c}_i, \mathbf{a}_j, \mathbf{b}_k) \,. \tag{4.1}$$

The embeddings are learned to optimally reconstruct the observed tensor entries. Crucially, the use of a factorized functional form allows such models to generate predictions for arbitrary combinations of cell-type, assay and genomic location, meaning that missing values in the tensor can be straightforwardly imputed given the learned embeddings.

4.4.4 eDICE model

Given that individual epigenomic tracks are either completely observed or completely missing, to impute a particular missing entry $\mathcal{Y}_{ijk}$ corresponding to the signal value in a missing track at a particular genomic location k, the most important source of information is the observed values of other tracks at the same location. Let Y^k represent the partially observed $(n_c \times n_a)$ matrix corresponding to taking a slice of the tensor at a particular genomic position. Our strategy is to learn to impute masked subsets of entries in Y^k given the remaining entries, by learning a factorized regression function:

$$\hat{Y}_{ij}^{k} = g_{\theta}\left(\mathbf{c_i}(\tilde{Y}^{k}),\ \mathbf{a_j}(\tilde{Y}^{k})\right) . \tag{4.2}$$

Here $\tilde{Y}^k$ is the matrix of local signal values in which a subset of tracks has been masked by setting the corresponding entries of the matrix Y^k to 0. All missing tracks likewise have their values set to 0. Factorization is achieved by encoding the matrix of local signal values into cell-type- and assay-type- specific representations, $\mathbf{c}_i(\tilde{Y}^k)$ and $\mathbf{a}_j(\tilde{Y}^k)$. These representations are thus directly conditioned on the local signal, unlike in the case of tensor factorization approaches, where cell-type and assay representations are global and parameterized directly.

The model produces the local embeddings $\mathbf{c}(\tilde{Y}^k)$ and $\mathbf{a}(\tilde{Y}^k)$ via separate cell and assay encoders. First, these encoders embed the local signal in each cell and each assay, then use self-attention to produce cell and assay embeddings that are informed by the signal in related cells and assays, respectively.

The model is trained by minimizing the mean squared error of the predictions of signal values in masked tracks in expectation over masks. The loss for a single genomic location is then:

$$\mathcal{L}(\theta, \phi, Y^k) = \frac{1}{|M|} \sum_{\{ij\} \in M} (\hat{Y}_{ij}^{k}(\theta, \phi) - Y_{ij}^{k})^2 . \tag{4.3}$$

The mean squared error is minimized with respect to the parameters of the encoders ϕ and decoder θ over a fixed training set of randomly selected genomic locations. At each iteration, a single mask M is drawn at random for each location and used to compute a Monte Carlo estimate of the loss. In practice, we mask 120 tracks at a time ($|M| = 120$), and use the remaining tracks as 'context' to predict the masked values. This training objective can be seen as a kind of self-supervised learning, similar to that employed by denoising autoencoders [413], but differing in the use of a factorized encoder and decoder. At test time, all tracks from the training set are used as inputs to predict the values of held out tracks.

Cell encoder

Let $\mathbf{y}_{c_i}^{(k)}$ denote a partially observed signal vector characterizing the signal in tracks across all assays in cell type c_i in the k^{th} bin (i.e. the size of this vector is n_a, where n_a is the total number of assays, some of which may be missing, and therefore set

to 0 for the cell type in question). This cell-specific local signal vector is mapped to an embedding space through a non-linear function $\mathbf{f}_{\phi_C}$, shared by all cell types, and implemented through a fully connected layer with parameters ϕ_C and a ReLU activation function. To allow the network to combine the local signal representation with knowledge of the global properties of the cell type, we add to the local signal embedding a learned global cell type embedding $\mathbf{u}_c$, which plays the role of a position embedding in the standard Transformer architecture.

$$\mathbf{h}^k_{c_i} = \mathbf{f}_{\phi_C}(\mathbf{y}^k_{c_i}) + \mathbf{u}_{c_i} \tag{4.4}$$

The cell encoder then applies a Transformer-style self attention block to the resulting embeddings:

$$\mathbf{c}_1(\tilde{Y}^k), ..., \mathbf{c}_{n_c}(\tilde{Y}^k) = \text{SAB}(\mathbf{h}^k_{c_1}, ..., \mathbf{h}^k_{n_c}) \tag{4.5}$$

The self-attention block (SAB) is identical to a standard self-attentive Transformer layer [89], except for the removal of Layer Normalisation, which we did not find important in our shallow networks.

To account for differences in the number of observed entries across cell types, a scaling step is applied in the signal embedding. This step involves multiplying the activations of the fully connected layer ϕ_C by a factor $\frac{1}{n_{obs}}$, where n_{obs} is the number of observed assays in the cell type, in an attempt to account for the uneven mapping of the epigenome, similar to the activation scaling used in Dropout [414].

Assay encoder

The assay encoder operates analogously to the cell encoder, taking as inputs assay signal vectors whose entries are the local signal values observed when performing a given assay in each cell type:

$$\mathbf{h}^k_{a_j} = \mathbf{f}_{\phi_A}(\mathbf{y}^k_{a_j}) + \mathbf{u}_{a_j} \tag{4.6}$$

$$\mathbf{a}_1(\tilde{Y}^k), ..., \mathbf{a}_{n_a}(\tilde{Y}^k) = \text{SAB}(\mathbf{h}^k_{a_1}, ..., \mathbf{h}^k_{a_{n_a}})\,. \tag{4.7}$$

Signal Decoder

The result of the factorized self-attention is a set of cell representations $(\mathbf{c_1}^k, \ldots, \mathbf{c_{n_c}}^k)$ and a set of assay representations $(\mathbf{a_1}^k, \ldots, \mathbf{a_{n_a}}^k)$, each of which is a function of the identity of the particular entity being represented and the full set of local signal values in all observed tracks at the k-th genomic bin $\left(\mathbf{c}_i^k \equiv \mathbf{c}_i\left(c_i, Y_{obs}^{(k)}\right)\right.$ and $\left.\mathbf{a}_j^k \equiv \mathbf{a}_j\left(a_j, Y_{obs}^{(k)}\right)\right)$. Given these representations, the prediction for a given cell type-assay pair is obtained by passing the corresponding contextual cell type and assay representations through the fully connected neural network g_θ (Equation 4.2).

4.4.5 Hyperparameters and training details

The model uses cell and assay embeddings of dimension 256 at all stages in processing. Within the self-attention block, we use 4 attention heads, whose output is concatenated and fed to a feed-forward neural network with a single hidden layer with 128 neurons and a 256-dimensional output. Finally, the combination of cell and assay representations are fed to a multilayer perceptron with 2 hidden layers with ReLU activations and 2048 neurons per layer. During training, Dropout with rate 0.3 is applied to each hidden layer in the output MLP.

The model used to analyse the eDICE performance in the Results section was trained on the union of the training and validation set for 50 epochs, using the Adam optimizer with a learning rate of 3×10^{-4}, and masking 120 randomly selected tracks to use as imputation targets for each training bin. Hyperparameters for this model were manually adjusted to maximise performance on the validation set.

For the EN-TEx imputations, the reconstruction task is modified so that the masked values belong to the same cell type in each individual bin, which closer mimics the generalization task analysed. The EN-TEx models have a reduced number of parameters in the embedding layers (128-dimensional) and the MLP hidden layers (512-dimensional), to account for the smaller dataset size. The transfer learning procedure involves training on one individual for 30 epochs, followed by 15 epochs of fine-tuning on the target individual with a reduced learning rate 3×10^{-5}.

4.4.6 Baselines

ChromImpute and PREDICTD imputations were downloaded directly from the resources accompanying their respective publications [395, 396]. In the case of ChromImpute, these imputations were generated in a leave-one-out manner, while PREDICTD's

imputations for tracks in our test set were generated by models respecting the same train-test split used to train eDICE, and our thus directly comparable to our results. Avocado's publicly available imputations, on the other hand, were generated by a model trained on the full Roadmap dataset (i.e. on all tracks, including the tracks in our test set), and therefore cannot be used to compare performance with other models. We therefore retrained an Avocado model from scratch to respect the data splits used here. To achieve this, we followed the two-stage procedure from [397], first training all parameters on chromosome 4, then freezing all parameters other than the genomic location embeddings, and fitting these for chromosome 21, to allow the generation of predictions for the test tracks on this chromosome. All results for Avocado refer to imputations made using this re-trained model.

4.4.7 Data and code availability

The Roadmap dataset is available at http://www.roadmapepigenomics.org/

The epigenomic tracks for the 4 individuals part of the EN-TEx dataset can be found on the portal for the ENCODE project https://www.encodeproject.org/. The accession codes used for the EN-TEx analysis are listed in Supplementary Table A.3.

The processed HDF5 files containing the training bins and chromosome 21 for the Roadmap dataset, and chromosome 21 for the selected tracks of the EN-TEx dataset can be found online at on Edmond, the open research data repository of the Max Planck Society [415].

Source code for eDICE [416] can be found at https://github.com/alex-hh/eDICE.

Chapter 5

Multimodal learning in clinical proteomics: enhancing antimicrobial resistance prediction models with chemical information

Large-scale clinical proteomics datasets of infectious pathogens, combined with antimicrobial resistance outcomes, have recently opened the door for machine learning models which aim to improve clinical treatment by predicting resistance early. However, existing prediction frameworks typically train a separate model for each antimicrobial and species to predict a pathogen's resistance outcome, resulting in missed opportunities for chemical knowledge transfer and generalizability.

We demonstrate the effectiveness of multimodal learning over proteomic and chemical features by exploring two clinically relevant tasks for our proposed deep learning models: drug recommendation and generalized resistance prediction. By adopting this multi-view representation of the pathogenic samples and leveraging the scale of the available datasets, our models equalled or outperformed the previous single-drug and single-species predictive models. We extensively validated the multi-drug setting, highlighting the challenges in generalizing beyond the training data distribution, and quantitatively demonstrate how suitable representations of antimicrobial

drugs constitute a crucial tool in the development of clinically relevant predictive models.

Declaration

This chapter is based on an updated version of the published manuscript [24]:

> *Multimodal learning in clinical proteomics: enhancing antimicrobial resistance prediction models with chemical information*
> Giovanni Visonà*, Diane Duroux*, Lucas Miranda, Emese Sükei, Yiran Li, Karsten Borgwardt, Carlos Oliver
> Bioinformatics (2023)
> * equal contribution

Compared to the published version, the section analysing the comparison of the ResMLP model to the single-drug single-species baselines was updated to clarify the effects of specific data-generating distributions on the ResMLP model. Additionally, subfigure (c) of the recommendation task results was adjusted to use the formulation of Average Precision at n that averages over all indexes from 1 to n, as it offers a more comprehensive description of the overall ranking performance. Figure captions and related discussions in the text have been adjusted accordingly. Minor grammatical corrections have been made for clarity. Data and code availability have been moved to the Methods section. The supplementary material from the original paper is included in Appendix B.

Author contributions (CRediT)

Giovanni Visonà: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization
Diane Duroux: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Lucas Miranda: Conceptualization, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Emese Sükei: Conceptualization, Validation, Writing - Original Draft, Writing - Review & Editing
Yiran Li: Conceptualization, Investigation
Karsten Borgwardt: Conceptualization, Resources, Writing - Review & Editing,

Supervision, Project administration, Funding acquisition
Carlos Oliver: Conceptualization, Validation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Visualization

5.1 Introduction

Antimicrobial resistance (AMR) poses a significant threat to human health worldwide. Based on recently published predictive statistical models, an estimated 4.95 million (3.62–6.57) deaths were associated with bacterial AMR in 2019, including 1.27 million (95% UI 0.911–1.71) deaths attributable to bacterial AMR [417]. Effective prevention strategies are urgently needed to stall AMR emergence and dissemination.

With a detailed understanding of the potential resistance mechanisms of the pathogen, clinicians can select specific antimicrobials with a higher chance of success. In this regard, disk diffusion and microdilution antibiograms are still the references for determining AMR [418]. While effective, these approaches are too cumbersome and time-consuming to enable the rapid selection of an adequate targeted antimicrobial treatment [419–421].

The emergence of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) provides a fast and cost-effective method for analysing bacterial strains. This technology is predominantly used as an analytical tool to identify and understand the structure of unknown biomolecules [422–424], and it has been used as an antimicrobial resistance detection tool in the clinic [425]. However, the usefulness of MALDI-TOF as a data source for machine learning AMR detection has only recently garnered interest in research [426–428]. These studies have mainly focused on creating models for specific combinations of antimicrobials and pathogens.

Many state-of-the-art (SOTA) tools such as CARD-RGI [429], AMRFinder [430], and SARGFAM [431] use variants of alignment-based methods like BLAST [432]. More recently, deep learning-based techniques have shown SOTA performance. Using similarity features to compare the query sequence to existing antimicrobial resistance gene (ARG) databases, DeepARG [433] was developed by building on a multi-layer perceptron model. Li et al. [434] proposed a multitask deep learning framework called HMD-ARG that first predicts whether the input sequence is an ARG and then predicts the resistant antimicrobial family, resistance mechanism, and gene mobility. Many published studies used pathogens such as *Staphylococcus aureus* and the β-lactam antimicrobial family [435–437]. Other relevant clinical pathogens, such as quinolones and macrolides, were studied in [438, 439]. Feucherolles et al. [421] showed that MALDI-TOF MS combined with ML provides a useful tool for AMR screening in the case of *C. coli* and *C. jejuni*.

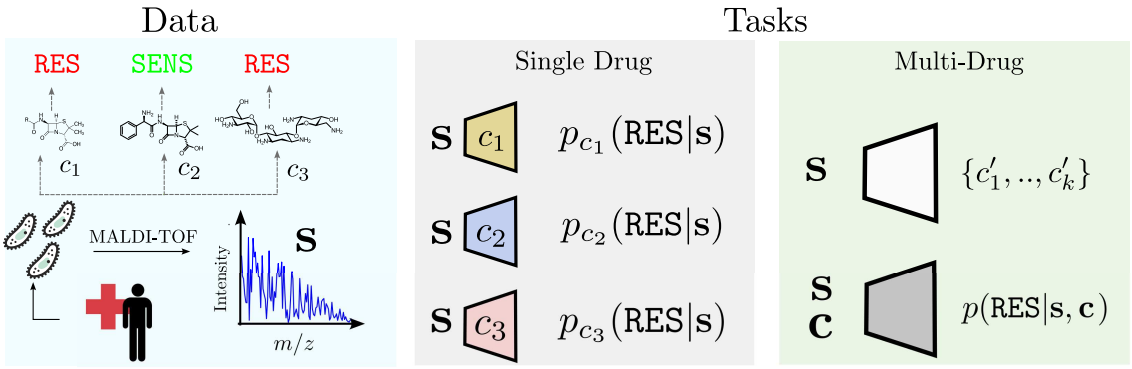


Fig. 5.1 **Description of antimicrobial resistance prediction tasks.** The dataset consists of MALDI-TOF mass spectra for bacterial samples of hospital patients treated for infection. For each sample, a set of compounds c_1, c_2, c_3 is annotated as inducing a sensitive or resistant outcome in the bacterial sample (left panel). From this data, we construct two new tasks extending the previous setting (middle panel) where each compound gives rise to a single binary classification task of resistance versus sensitivity for a given spectrum $\mathbf{s}$. In this work, we introduce two tasks (right panel), which are to predict resistance given a drug-spectrum pair and to recommend antimicrobials for a given observed spectrum.

In 2022, Weis et al. [440] developed a large (over 700,000 resistance labels and 300,000 MALDI-TOF spectra) *Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra (DRIAMS)* and utilized ML models to predict the resistance of significant pathogens like *Staphylococcus aureus*, *Escherichia coli*, and *Klebsiella pneumoniae*. The study concluded that focusing on predicting resistance for specific species-drug pairs improved classifier accuracy, likely due to the complexity of resistance mechanisms.

Drug recommendation is another ML application that has been gaining significant interest, particularly in cancer research. Various solutions have emerged, including Kernelized Bayesian Multi-Task Learning [441], which learns the relationships between different drugs during training. This algorithm, along with random forest, was the best-performing approach in a challenge-based competition on a breast cancer dataset [442]. Another promising approach is Kernelized Rank Learning [443], which focuses on providing a ranked list of drugs instead of exact sensitivity values and was specifically designed to handle sparse training datasets. Along these lines, recommendation models could assist in maintaining or enhancing infection coverage rates while employing fewer broad-spectrum antimicrobials than current practices [444].

Although previous works show promising results, the current SOTA AMR prediction methods based on proteomics do not integrate multiple relevant data sources, such as the chemical composition of antimicrobials alongside MALDI-TOF spectra obtained from pathogenic samples. Instead, a separate model is trained for each antimicrobial and pathogen species combination, limiting the potential for knowledge transfer, generalizability to new drugs, and deciphering the underlying resistance mechanisms. Developing such general-purpose models could enhance patient care in a robust and highly adaptable way.

To address this problem, our work focuses on incorporating chemical data into antimicrobial resistance prediction using mass spectrometry pathogen profiles (Figure 5.1). This learning framework has been successfully applied in predicting cell line response to cancer drugs, with some models proving successful [445, 446]. We propose several prediction and evaluation settings for antimicrobial resistance where chemical information can be utilized and demonstrate increased prediction performance and generalizability compared to single-drug models. Furthermore, we define direct drug recommendation models to predict drugs with a high chance of sensitivity or resistance for unseen spectra and thoroughly evaluate their feasibility and performance.

5.2 Methods

Using the DRIAMS dataset (Section 5.2.1) and molecular fingerprinting (Section 5.2.1), we explore two major prediction settings which leverage chemical information: drug recommendation (Section 5.2.2) and resistance prediction (Section 5.2.3). Through these settings, we test whether chemical information can be used to improve resistance prediction SOTA and propose new model development avenues. These workflows are illustrated in Figure 5.1.

5.2.1 Dataset

MALDI-TOF mass spectra dataset

This study utilized the publicly accessible DRIAMS dataset [447], a comprehensive resource comprising MALDI-TOF mass spectra obtained from hospital patients across four Swiss diagnostic labs during the period spanning 2016 to 2018. The dataset encompasses 303,195 mass spectra and 768,300 antimicrobial resistance labels, covering 803 different bacterial and fungal pathogen species. The dataset has been meticulously organized into four distinct sub-collections, each representing different hospital sites.

Each data point contains a mass spectrum from a patient sample, complemented by annotations denoting its susceptibility or resistance to as many as 71 antimicrobials. In our analysis, we harnessed the 6000-dimensional binned mass spectra vector representation, aligning with the methodology proposed by Weis et al. [440].

Chemical feature extraction

Molecular fingerprinting [448, 449] is a popular method for encoding chemical information into numerical features for ML models. It represents a molecule as a series of bits that encode the presence or absence of certain substructures. This technique captures important information about the molecular structure, including topological, physiochemical, and structural properties. The use of molecular fingerprints is prevalent in chemical informatics and drug discovery and has been shown to be effective in many applications [450].

We tested three standard techniques, namely the Molecular ACCess Systems keys (MACCS) [451] (166-bit keyset), the PubChem Fingerprints (PubChemFP) [452] (881-bit long keys), and the 1024-bit long Morgan fingerprints [453]. We obtained the fingerprints for the antimicrobial drugs in the DRIAMS dataset using RDKit [454] and PubChemPy [455], two open-source Python packages. Certain treatments present in the dataset consist of mixtures of compounds; since it is impossible to associate a fingerprint representation in such cases, they have been excluded from our analysis.

5.2.2 Drug recommendation

We first examine the interaction between clinical proteomics and chemical features through the task of drug recommendation. In the recommendation setting, a model directly suggests a set or a ranking of potentially suitable drugs for a query spectrum. To perform this search, we test various explicit and learned functions of spectrum and chemical similarity for each query spectrum, returning n recommendations.

We evaluate the effectiveness of five personalized treatment recommendation methods focusing on the impact of incorporating various levels of information into the drug ranking process, including the pathogen species, spectra, and drug features:

1. Random baseline: randomly select k samples from the training set for a query sample and return the drugs that most frequently elicited a sensitive reading among the k samples.

2. Baseline species: randomly select k samples from the training set which correspond to the same pathogen species as the query, and again return the drug which most frequently results as effective.

3. Spectrum similarity: given a similarity function between spectra, select the k most similar spectra to the query. We test multiple measures of similarity between spectra, namely cosine similarity, correlation, Euclidean, Manhattan and Wasserstein distances.

4. Siamese networks: learn joint embeddings of the drugs and spectra and use logistic regression (LR) on the embeddings to rank drugs for recommendations based on the resulting probabilities. Siamese networks [456] contain two identical subnetworks with shared weights and work in tandem on two input vectors composed of the MALDI-TOF mass spectra and the chemical fingerprints to minimize the difference between the actual and predicted similarity between pairs of observations (Supplementary Figure B.3a).

5. ResMLP: train a classification multi-layer perceptron with residual skip-connections [457] to predict the probability of resistance for drug-spectrum pairs. Each drug is then ranked according to the predicted resistance likelihood. This model uses skip-connections that provide a path for data to reach deeper layers in the network by skipping some layers (Supplementary Figure B.3b), generally improving the training procedure. To account for the different number of features in the MALDI-TOF mass spectrum and the chemical fingerprint, the model first projects each to the same dimension before concatenating the two vector representations and using them as input for the feed-forward network.

We test multiple values of $k \in \{1, ...100\}$ and study their impact on performance to determine the optimal threshold. Then, we use majority voting: among the drugs with known response values for the test sample, we recommend the drug that results most often sensitive across the chosen k samples. If multiple drugs have the highest occurrence, we select all as recommended treatments. If there are no common drugs between the drugs tested for a specific sample in the test and the drugs we recommend, then we do not compute the performance.

Evaluation

The test set consists of a random selection of 20% of the samples and all associated tested drugs to ensure that all observations related to a spectrum are in the same set.

Additionally, we impose a constraint that each spectrum in the testing set must be associated with at least one resistant and one sensitive outcome.

We conduct our recommendation analyses on the DRIAMS-B dataset, with the training set containing 1907 pathogen samples and the test set containing 477 pathogen samples, and report multiple measures to evaluate the performance of each approach derived from the literature on information retrieval, namely precision P, precision at cutoff n $P@n$, and the mean average precision at cutoff n $mAP@n$ (additional details in Supplementary Section B.4).

5.2.3 Generalized antimicrobial resistance prediction

In the resistance prediction task, each observation corresponds to a biological sample and a drug to which it was exposed. The aim is to associate with each sample-drug pair an outcome that estimates the likelihood of the sample being resistant to the drug.

The task can be formalized as learning a mapping $f : \mathcal{X} \times \mathcal{C} \to [0, 1]$ where $\mathcal{X}$ is the space of bacterial samples and $\mathcal{C}$ is the space of chemical compounds. Each biological sample is represented by the measured MALDI-TOF spectrum, while the chosen molecular fingerprints represent the antimicrobial drugs. With this formalism, we model the output $f(x_i, c_j)$ as the probability that the sample corresponding to the mass spectrum x_i is resistant to the antimicrobial drug represented by c_j. This formalism generalizes the original prediction setting introduced in [440], which learns one predictor for each compound and only uses the spectrum as input.

We compared three machine-learning-based approaches to model the resistance prediction function:

1. Baseline: principal component analysis (PCA) with LR. As the dimensions of the mass spectra and the chemical fingerprints are of different scales, applying PCA projects them to lower and comparable dimensions while preserving 95% of the variance of the original variables. These embeddings were then concatenated and used as input for the LR model.

2. Siamese networks: similarly to the drug recommendation case, we use the learnt joint representations from the Siamese networks as input to LR to yield the resistance predictions.

3. ResMLP: train a classification ResMLP to predict the probability of resistance (we use the same configuration as the recommendation task; see Section 5.2.2).

We designed three data splits to reflect different data-generating processes to examine the prediction capabilities of the previously described machine learning models.

1. Random split: the observations in each DRIAMS dataset are randomly sampled to create training, validation, and test sets with a partitioning of 70%, 10%, and 20%, respectively. This data split corresponds to the IID setting, where all the sets are drawn from the same joint distribution.

2. Species-drug zero-shot split: the test set contains novel pairs of species and drugs. Given the finite size of the datasets, we used a heuristic to randomly select species-drug combinations that account for approximately 20% of the data and ensured that the species s and the drug d are not present in any observations of the training set. The remaining data is randomly split into training and validation sets that contain approximately 70% and 10% of the dataset, respectively.

3. Drug zero-shot split: we hold out as a test set all the observations corresponding to the target drug d and test how accurate the predicted resistances are for a compound that the model has never seen in training.

We report three standard classification metrics for imbalanced data: area under the precision-recall curve (AUPRC), balanced accuracy, and Matthews correlation coefficient (MCC). To analyse the importance of chemical features in the AMR prediction task, we employed SHAP [458], a framework rooted in game theory that is among the most popular post-hoc interpretation methodologies (additional details in Supplementary Section B.5).

5.2.4 Data and code availability

The DRIAMS dataset is publicly available online[1]. The code used to produce the results presented in this work is available at https://github.com/BorgwardtLab/MultimodalAMR.

5.3 Results

In this section, we ask whether (i) the drug recommendation setting using DRIAMS contains useful and non-trivial spectrum-drug associations (Sections 5.3.1, 5.3.2, 5.3.5), and (ii) whether generalized AMR prediction models are comparable to SOTA

[1]https://datadryad.org/stash/dataset/doi:10.5061/dryad.bzkh1899q

single-drug models and how they behave under different data-generating distributions (Sections 5.3.3, 5.3.4).

5.3.1 Model-free approaches offer strong baselines for recommending drugs

We first analyse the use of 3 model-free recommendation approaches to select drugs suitable for treating clinical patients (Section 5.2.2). The random baseline, baseline species, and spectrum similarity methodologies rely on similarities between a test sample and samples from the training set. In these set-ups, we select the top-k similar samples and recommend the drugs that result as effective most often in the selected set.

For all three methods, the test precision quickly increases up to $15 \leq k \leq 30$, then stabilizing or showing small changes (Figure 5.2 (a)). Therefore, selecting a high k is preferred over including only a few samples. Based on these results, in the following analyses, we used $k = 30$.

The performance based on the random baseline set-up is the lowest among the three methods. This indicates that the other two methods incorporate additional information beyond recommending drugs based on the highest occurrence across samples. The spectrum similarity approach led to comparable performance to the baseline species method, suggesting that spectra similarity is insufficient to capture significant additional information compared to the species. In the baseline species, the performance is only computed when more than k available samples correspond to the same species in the training set. This could introduce a bias when k increases if the number of species in the training set is not random but can be accounted for by external variables or properties. For instance, the performance could be deflated if the drug sensitivity is more homogeneous for species that only appear a few times in the dataset.

We evaluated the effect of increasing the number of top similar samples on the number of recommended drugs. Indeed, if multiple drugs have the same highest occurrence across the top-similar samples, it leads to the recommendation of several drugs. We found that as the number of samples used for the majority vote increases, the likelihood of obtaining a ranking with no similar occurrence also increases (Supplementary Figure B.1a). The baseline species set-up resulted in the highest number of recommendations (ranging from 12 drugs with $k = 1$ to 2 drugs with $k = 100$) while the random baseline

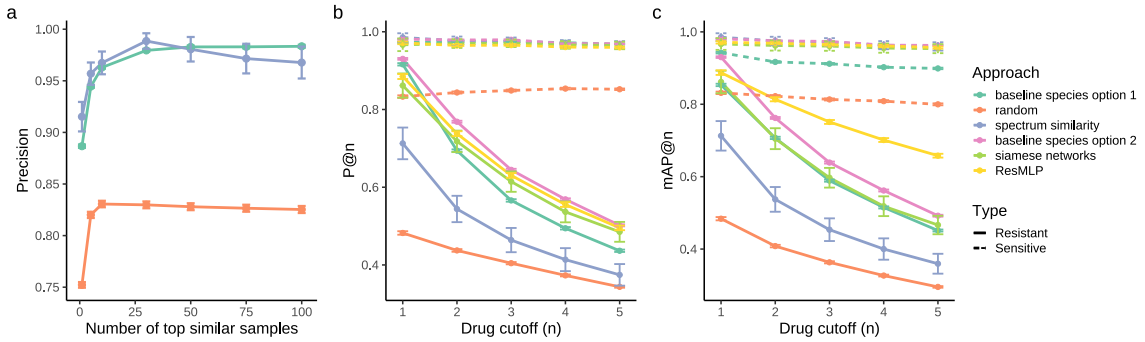


Fig. 5.2 **Performance metrics for the recommendation task. (a)** Recommendation performance for the most frequently recommended drugs based on the top-k similar samples, with 95% Confidence Intervals. The top-k most similar samples to the new observation are selected; drugs are ranked according to the number of similar samples sensitive to the drugs. The drug that exhibits the highest frequency of sensitivity is identified. If multiple drugs show comparable sensitivity properties across similar samples, we include them all. The precision obtained from these drugs for the new sample is reported. Three approaches to assess the similarity between samples are compared: random, baseline species option 1, and spectrum similarity (details in Section 5.2.2). Precision **(b)** and mean average precision **(c)** at multiple cut-offs n across all samples in the test set for the different recommendation methods. For the random baseline, baseline species, and spectrum similarity option 1 approaches, the number of top neighbours k is set to 30. In the spectrum similarity option 2 set-up, k is the maximum number of samples available in the training set that correspond to the same species that the sample investigated in the testing set. The dashed lines in (b) and (c) represent recommendations that aim to assign high ranks to drugs to which the sample is sensitive, while the continuous lines aim to rank the drugs to which the sample is resistant.

set-up had the lowest number of recommendations (ranging from 6 drugs with $k = 1$ to 1 drug with $k = 100$).

5.3.2 Beyond sensitivity: the challenge of targeting resistance in drug recommendation systems

After examining the model-free baselines, we tested their performance against the Siamese network and ResMLP models by producing ranked recommendations. The recommendations targeted both sensitivity and resistance. They were evaluated with precision at cutoff n and the mean average precision at cutoff n. In this context, sensitivity and resistance refer to the pathogen's response to the effects of a drug, either

by being susceptible to its therapeutic action or by having mechanisms to withstand its impact.

Fig 5.2b illustrates the precision at cut-offs 1 to 5 for both sensitivity and resistance. For the baseline species set-up, we also consider the additional option of setting k to the maximum number of samples available in the training set that correspond to the same species that the sample investigated in the testing set (baseline species option 2). With this approach, k changes from one test sample to another. In the sensitivity recommendation setting, the performance of the random baseline is again the lowest, with the other methods yielding comparable performance. The spectrum similarity approach already achieves a very high mean precision (0.97).

Overall, integrating drug fingerprints in the models produced results similar to those from the baseline species approaches for the recommendation task. A limitation of the approaches based on the top-k neighbours (including the baseline species set-up) is that we cannot evaluate the precision for drugs not tested in the top-k neighbours. The precision decreases when the drug cut-off increases for all methods in the resistance setting. This is likely due to the low numbers of drugs to which samples in the testing set are resistant (3.2 on average versus 12.3 for sensitivity).

In general, the precision at cut-offs 1 to 5 of the drug sensitivity recommendation (dashed lines) is overall higher than the corresponding precision for the drug resistance recommendation. However, this could be due to the metric used to evaluate the recommendation system. Indeed, the resistance precision at cut-off n may decrease due to the test sample being resistant to very few drugs.

To address this, we also evaluated the truncated version of precision at cut-off n (Supplementary Figure B.2), confirming that resistance is more challenging than sensitivity as a recommendation target.

Finally, we determined the mean average precision at cutoff n, which considers not only the number of correct predictions but also the associated ranking (Figure 5.2c). For the identification of the sensitivity, the random baseline and random baseline option 1 still lead to the lowest performance. The other methods give very close results. However, for the identification of the resistance, from cutoff $n = 2$, the ResMLP model performs better than all the other approaches. Hence, while most approaches are able to recommend a sufficient number of sensitive drugs, the ResMLP model demonstrates greater consistency in identifying the most resistant drugs, leading to the highest $mAP@n$ overall. This result highlights the value of using the ResMLP

model and, more generally, the inclusion of the drug chemical features in the resistance prediction task. Overall, Figures 5.2b, 5.2c, and Supplementary Figure B.2 show that precise drug recommendation offers promising opportunities but also highlights the complexity of the task. These analyses motivate further research on the methodological developments of MALDI-TOF mass spectra and drug molecular fingerprinting for antimicrobial recommendation.

5.3.3 Joint modelling of chemical and proteomics information subsumes single-species single-drug classifiers

To evaluate the effectiveness of joint multimodal learning, we compared our deep learning model to the more restricted machine learning-based approaches proposed in [440] where a single model is trained for each drug-species combination.

We selected the same drug-species combinations described in the paper, and we present here 3 of them. The full set of the drug-pathogens combinations showcased in Figure 2 of [440] can be found in the Appendix B (Supplementary Figures B.4 and B.5). We specifically focus on a marker for methicillin-resistant *Staphylococcus aureus* (MRSA) [459] by analysing resistance to Oxacillin, and a marker for resistance against broad-spectrum beta-lactam antibiotics by examining the resistance of *Escherichia coli* and *Klebsiella pneumoniae* samples to Ceftriaxone.

We adopted a 5-fold validation scheme to estimate the test performance using the area under the receiver operating characteristic curve (AUROC) and AUPRC as metrics. For each combination, the target samples are held out from the DRIAMS A set and split into 5 folds. A ResMLP is trained on all the remaining DRIAMS A samples that do not include the target spectra to obtain a pre-trained network, using the configuration described in Supplementary Section B.3. Finally, each of the 5 test splits is selected as target, and the remaining 4 splits for the target combination are used to fine-tune the last 2 layers of the model with a reduced learning rate before outputting the predictions for the test fold.

In addition to this fine-tuned model, we also trained a ResMLP by including in the pre-training phase samples corresponding to the target spectra, but in combination with drugs other than the target drug. This effectively changes the prediction task to be closer to a few-shot learning setting, where some of the resistance labels for a target spectrum are given, and we look to impute the resistance to the target drug.

The results of these experiments are shown in Figure 5.3, where the fine-tuned ResMLP model displays performance comparable or slightly improved compared to the baseline model. Interestingly, the extended setting in which the model learns from the additional resistance labels shows considerably better results, highlighting how effective this transfer of information is. This result is expected, as the DRIAMS A dataset contains samples for antibiotics with related modes of action (such as Cefepime and Ampicillin for the beta-lactam antibiotics). To corroborate this observation, we can highlight how beta-lactam antibiotics constitute a sizable portion of the DRIAMS A dataset ($\sim 23\%$ of observations), and there is no lack of samples for which multiple resistance outcomes are available, hence the large improvement for Oxacillin and Ceftriaxone. In contrast, fusidic acid is a drug that inhibits the bacterial elongation factor G (EF-G) during protein synthesis; no other drug with the same mode of action is present in the DRIAMS A dataset, which is likely a determining factor for the poor predictive performance and basically non-existent improvements for the ResMLP model on the prediction of resistance to fusidic acid in *Staphylococcus aureus* samples (Supplementary Figures B.4 and B.5).

This large improvement based on the added pre-training data suggests that more research is needed to examine the effects of data availability on the predictive power of the model, which can further direct data collection efforts. Additionally, it is not uncommon for infections to reoccur in the same individual; in this case, access to the clinical history of the patient can provide valuable information that has been shown to be relevant in certain cases [460].

Overall, the multimodal ResMLP architecture displays comparable or often slightly better performance than the single-species single-drug baselines (Supplementary Figures B.4 and B.5), while enabling the use of the model for additional tasks such as drug recommendation (Section 5.3.2).

5.3.4 Deep learning enables accurate predictions of antimicrobial resistance in the IID setting

The performance of AMR prediction models can vary significantly depending on the data-generating process of the target prediction task.

We performed a set of experiments to analyse the predictive performance of our models in the three data splits described in the Methods section. The random, species-drug zero-shot, and drugs zero-shot splits correspond to the IID setting, the

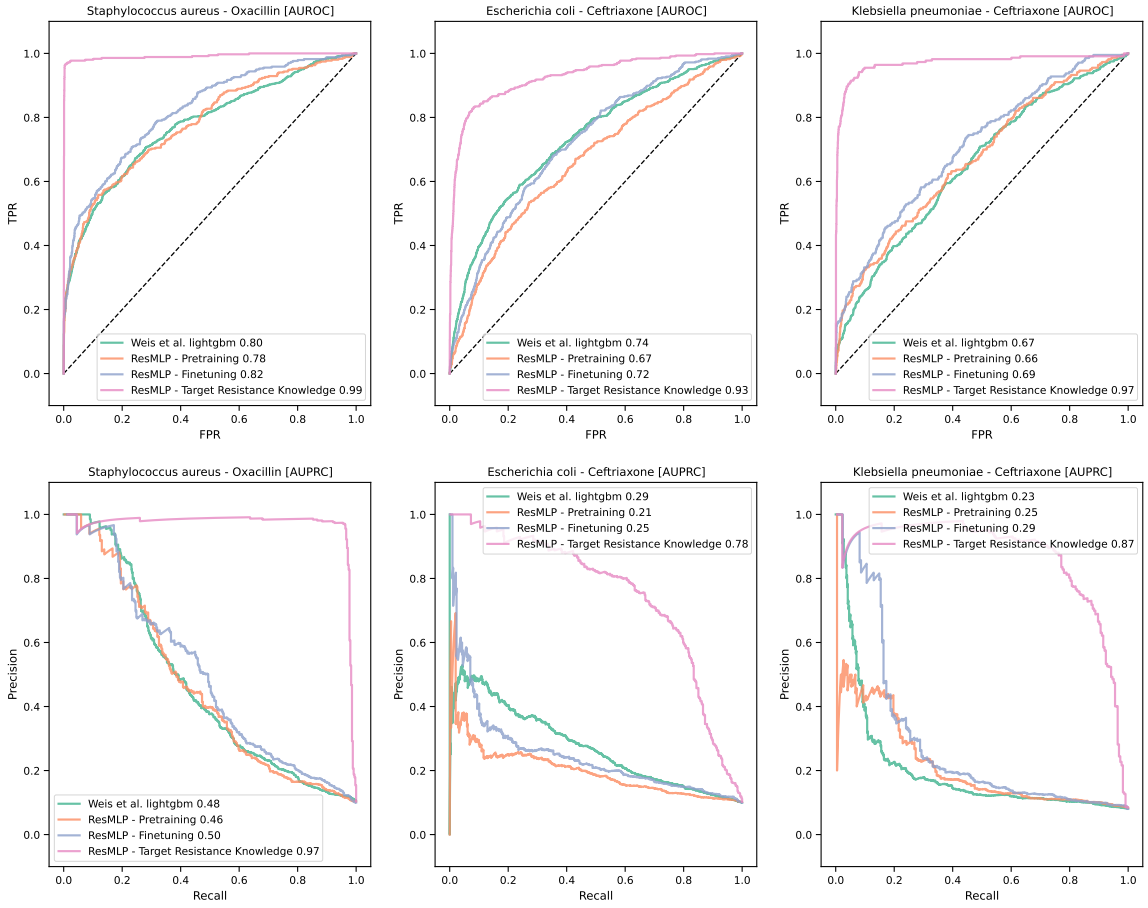


Fig. 5.3 **Comparison with the LightGBM baseline model from [440] for three significant drug-species combinations.** On this subset, the ResMLP model performs comparably to the original baseline, with the fine-tuned ResMLP offering some performance advantages on average. With a model that has access to other resistance labels for the target spectra, the performance improves by a large margin. This is most likely due to the presence of other antimicrobials with similar mode of action in the training set. Numerical values of AUROC and AUPRC are reported from the average performance over 5 train-test splits for the target drug-pathogen combination.

Table 5.1 **Direct AMR prediction results with multi-drug models on the DRIAMS B dataset.** The average metrics are reported together with their standard deviation that is obtained by repeating the analysis over multiple randomization seeds for the random and species-drug zero-shot splits, and for each held-out drug in the drug zero-shot split. In bold, we highlighted the best average metric across models for a specific split.

Split type	**Model**	**Cross-validation performance score - mean (SD)**		
		AUPRC	**Bal. accuracy**	**MCC**
Random	PCA + LR	0.64 (0.02)	0.705 (0.007)	0.51 (0.02)
	Siamese + LR	0.49 (0.01)	0.76 (0.01)	0.53 (0.02)
	Sp-ResMLP	0.35 (0.04)	0.59 (0.03)	0.21 (0.05)
	ResMLP	**0.87 (0.02)**	**0.90 (0.01)**	**0.79 (0.02)**
Species-drug zero-shot	PCA + LR	0.44 (0.04)	0.63 (0.02)	0.30 (0.04)
	Siamese + LR	0.42 (0.01)	0.664 (0.004)	**0.40 (0.01)**
	Sp-ResMLP	0.52 (0.04)	0.62 (0.02)	0.30 (0.04)
	ResMLP	**0.54 (0.04)**	**0.70 (0.02)**	0.39 (0.03)
Drug zero-shot	PCA + LR	0.33 (0.25)	0.57 (0.12)	0.12 (0.16)
	Siamese + LR	0.18 (0.16)	0.52 (0.05)	0.08 (0.14)
	Sp-ResMLP	0.17 (0.16)	0.50 (0.12)	0.01 (0.17)
	ResMLP	**0.47 (0.31)**	**0.71 (0.15)**	**0.35 (0.28)**

generalization to novel species-drug combination, and the generalization to new drugs, respectively.

The best results obtained by each model, shown in Table 5.1 for the dataset DRIAMS B and in Supplementary Table B.2 for all collection sites, reveal several interesting aspects of the AMR resistance prediction task. The IID setting of the random split allows the models to produce the best possible results, while the out-of-distribution splits pose a considerable challenge for obtaining accurate predictions.

The ResMLP model outperformed the other approaches in several prediction settings by significant margins. This model represents the largest of the methods tested, with $\sim 8.9M$ trainable parameters in the final configuration adopted, and required a much higher computational cost with training that included up to several hundred epochs of optimization (with a certain amount of variability due to the use of early stopping). This result suggests that the AMR prediction task may benefit from using large-scale deep learning models, whose success is predicated on collecting large quantities of data.

The species-drug zero-shot split leads to a noticeable degradation in performance for all models except the ablation experiment Sp-ResMLP (Section 5.3.5), suggesting that training the model on data that captures the interaction between specific pathogenic samples and antimicrobial drugs is crucial to leverage the information contained in the MALDI-TOF spectra.

The drug zero-shot prediction task was a difficult challenge for all models, as indicated by the large standard deviations in the measured metrics (Table 5.1). The high variability in performance can be attributed in part to the heterogeneous test splits for this task. Unlike the other two test settings, where the overall class balance from the dataset can be maintained with stratified splits, the class imbalance in the test set can vary significantly depending on the held-out target drug (see Supplementary Figure B.7). Additionally, we speculated that the test performance for a held-out drug could depend on its similarity to the remaining compounds in the training set. However, further analysis in this direction failed to reveal any direct correlation (see Supplementary Figure B.1).

The full set of plots showcasing the test AUPRC in the drug zero-shot split is available in Supplementary Figure B.9.

5.3.5 Ablation experiments and feature importance show the value of combining MALDI-TOF spectra with chemical features

We evaluated various configurations and design options for each model. These included early integration of the MALDI-TOF and chemical fingerprint, dimensionality reduction with a single PCA projection, and the use of different chemical fingerprints and classifiers for the PCA and Siamese methods. However, none of these design choices yielded results that surpassed those of the deep-learning-based ResMLP model.

To determine the value added by the MALDI-TOF spectra in predicting AMR compared to considering only the species of the bacterial samples, we trained a ResMLP model by replacing the input of the MALDI-TOF spectra with a simple 1-hot encoding of the species. The results, as shown in Table 5.1 under the label Sp-ResMLP, indicated a significant decline in performance in most test scenarios.

During our experimentation, we tested the use of different molecular fingerprinting methods. The use of feature importance analysis revealed the value of using chemical

fingerprinting methods. However, no specific fingerprint class emerged as consistently superior to the others (Supplementary Table B.3, Supplementary Figure B.8). Where it is not otherwise specified, we made use of the 1024-dimensional Morgan fingerprints (also known as ECFP4), which we chose since it is one of the most popular molecular representations used for small molecule screening, which has demonstrated robust performance in several tasks.

Finally, we utilized SHAP values [458] to quantify the contributions of the sets of spectral and chemical features for AMR in a ResMLP model trained using the MACCS chemical fingerprints. Analysing the feature importance grouped by data type (Supplementary Figure B.10) and the most important features (Supplementary Figure B.11) showed that both the spectrum and fingerprint features played an important role in the final prediction, corroborating our design choices.

Mapping back the highlighted features to the input MACCS fingerprints uncovered intriguing patterns related to well-known AMR mechanisms [461–463]. Specifically, our findings demonstrated that, for beta-lactam antimicrobials, the beta-lactam ring was a critical feature, especially in penicillins. The top features of aminoglycoside antimicrobials included amine or alcohol groups from sugar rings. Chloramphenicol and macrolide antimicrobials also displayed significant chemical features that align with known resistance mechanisms. These insights may inform the design of novel antimicrobials with improved resistance profiles. A visual representation of these findings can be seen in Figure 5.4, where antimicrobial structures are displayed with highlighted atoms corresponding to the discussed chemical features.

5.4 Discussion

This study explored the integration of chemical and proteomics data to predict antimicrobial resistance outcomes. By employing deep learning models, we examined the benefits of combining patient MALDI-TOF mass spectra with chemical fingerprints, which can be collected in a time-sensitive manner, making it highly relevant for various clinical applications. Our results indicate that combining information from multiple drugs and species outperforms baseline methods and demonstrates the potential of transferring chemical knowledge to improve antimicrobial resistance predictions. Moreover, we showed the potential for generalizability of our approach by incorporating drug information and evaluating its effectiveness on unseen compounds.

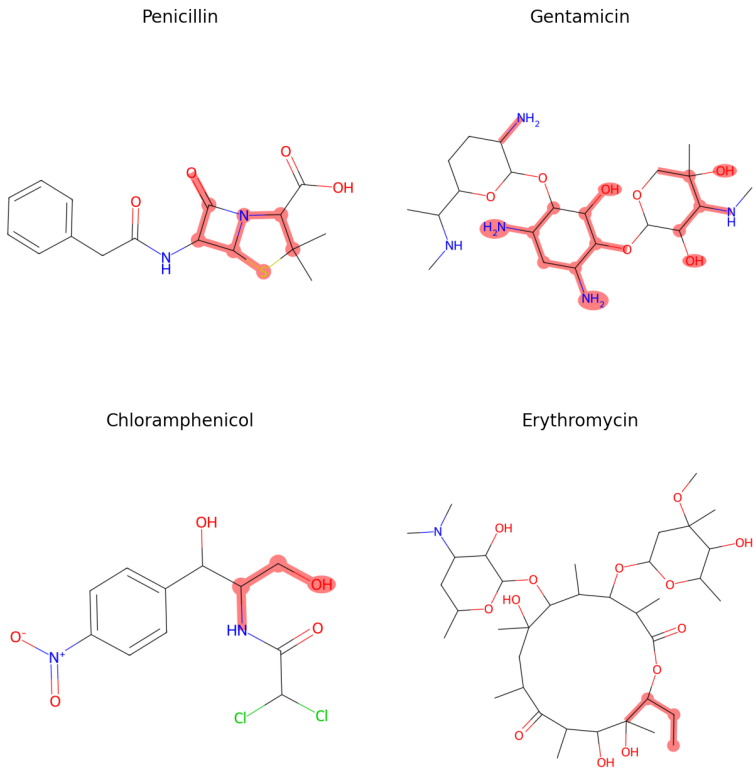


Fig. 5.4 **Chemical structures of representative drugs in four common antibiotic families**. The beta-lactam ring in **Penicillin** (highlighted in red) is a key structural feature in beta-lactam antibiotics such as Penicillin, where resistance is frequently conferred by beta-lactamase enzymes, which hydrolyse the amide bond, rendering the antibiotic inactive. The beta-lactam ring ranks among the first in the SHAP feature importance analyses presented in this paper for most Penicillin antibiotics (such as Amoxicillin, Oxacillin, Ampicillin and Benzylpenicillin). Interestingly, antibiotics that are not susceptible to beta-lactamases (such as Cefepime and Aztreoman) do not follow this trend. **Gentamicin** is an aminoglycoside antibiotic, as are Amikacin and Tobramycin. Although other resistance mechanisms involving these drugs have been reported, by far the most prevalent involve aminoglycoside-modifying enzymes that target the glycoside rings and their aglycone components, which matches the top-ranked features by the provided SHAP analysis (highlighted in red). **Chloramphenicol** encounters high-level bacterial resistance due to the enzyme chloramphenicol acetyltransferase. This enzyme mediates the transfer of an acetyl group from acetyl CoA to the primary hydroxyl group within the chloramphenicol molecule, which ranks as the top chemical feature in the provided SHAP analysis. **Erythromycin** is a macrolide (like Azithromycin), a family of antibiotics where drug modification is also the prevalent resistance mechanism in place. Among others, hydrolysation of the ester group by esterases in particular acts in the ester group next to the atoms highlighted in red (which rank first in the provided feature importance analysis for both mentioned macrolide antibiotics).

The use of deep learning in antimicrobial resistance prediction is not new. In contrast to other methods that require genetic sequencing [433], we rely on publicly available information on drug structure and on mass spectrometry data, which is already a routine for species identification, preceding cell culture to identify the best antimicrobial treatment [440]. We expect that this work will offer new insights to the AMR community in the direction of unifying knowledge and eventually deciphering the relationship between pathogen composition and chemical features of treatments.

Exploring further the idea of reasoning over chemical space for AMR prediction, we proposed recommendation systems that can robustly predict drugs with a high chance of sensitivity or resistance for unseen spectra. By reducing the application of generic, non-specific medications in most cases, machine learning models could also help prevent antimicrobial overuse.

Proteomics and genomics have both been used in AMR prediction [464, 427, 465, 421]. Proteomics, notably MALDI-TOF MS, offers a rapid and cost-effective diagnostic method for infectious diseases in clinical settings. It provides almost real-time insights into organism responses to antibiotics and functional information, aiding in tracking adaptive responses and discovering resistance mechanisms. Genomics, on the other hand, provides stable DNA data for consistent comparisons and identifies intrinsic resistance mechanisms like gene mutations. Molecular diagnostics like PCR swiftly detect resistance genes, but often target single genes and lack comprehensive insight into non-genetically mediated resistance mechanisms.

Although our primary focus is on the application of MLPs in AMR prediction, there is also potential in multi-label classification approaches [466]. For instance, ensemble methods [467] could offer a promising avenue for further investigation. Moreover, the astonishing progress in graph neural networks [468] makes it attractive to represent compounds as attribute-rich networks, and representation learning attempts have already shown that these approaches can successfully generate molecules with specific properties [469]. A follow-up in this direction could help increase performance even further by representing drugs more efficiently. As data collection efforts grow across multiple sites worldwide, the prospect of training large-scale representation learning models, akin to foundation models [86], for AMR prediction appears increasingly attainable. Additionally, evaluating the cross-site generalizability of our models is paramount to ensure the robustness and applicability of our findings across diverse healthcare settings, ultimately enhancing the potential impact of our research.

Nonetheless, our approach represents a novel and more flexible development over the previous state-of-the-art. It constitutes a step towards building technologies that can leverage as much information as possible from different relevant modalities. This advancement holds the promise of enhancing patient care through more precise and adaptable predictive tools.

Chapter 6

Conclusions

6.1 Summary and discussion

The transformative potential of machine learning applied to biomedicine is a key driver of novel developments and opens up unprecedented opportunities for decoding biological mechanisms and improving clinical care. Different interests come into play in a complex picture that raises crucial challenges to ensure that this research benefits all parties involved.

The focus of this thesis was the analysis of the technical hurdles in enabling biomedical machine learning models to perform well outside the training distribution. The task of generalizing to new domains or data-generating distributions is a crucial part of ML research (Chapter 2); while machine learning—and deep learning in particular—performs incredibly well in the IID regime, difficulties arise when the test setting differs significantly from the training data (Section 2.1). Within the biomedical field, this issue is especially relevant: the variety and frequency of biases in the processes to gather and analyse biomedical data inject spurious correlations and confounding biases that obfuscate the underlying biological mechanisms (Section 3.2).

So the fundamental question at the heart of this thesis arises: how do we train biomedical ML models to generalize outside the training distribution, and in doing so, achieve personalized or fine-grained predictions that enable precision medicine? Understanding the sources of bias and accounting for them in the development of ML models is vital for achieving the best possible outcome when translating the results of ML research into practical applications.

To this end, this thesis examined various paradigms that have been developed over the years to correct for data shifts, transfer knowledge, or learn representations that are less affected by domain changes (Section 2.6). These methodologies, in particular transfer learning and zero-shot learning, played a central role in the two projects presented in Chapters 4 and 5.

The TL scheme adopted with the eDICE model, for example, enabled the pre-trained model to learn representations for all the tissues available for the training individual; with the added fine-tuning step, the model was able to impute individual-specific peaks (Section 4.2.4), a first case study for the application of deep learning to personalized epigenomics.

The zero-shot learning prediction tasks designed to test the ResMLP model, on the other hand, allowed us to delve into the workings of the ResMLP model developed to predict resistance to antimicrobials from MALDI-TOF spectra (Chapter 5). A careful combination of ablation experiments and tests of the model under different data-generating distributions revealed several insights that can guide the future developments of this approach, ranging from the inclusion of the patient's clinical history, which has already been shown to be relevant for certain infections [460], to the need for more informative representations for chemical compounds.

Overall, generalization approaches are a key piece of biomedical machine learning (as the two projects presented demonstrated), and it needs to be front and centre in the priorities of the research community. Increasing the reliability and robustness of ML models, excluding biases from the data, and carefully combining computational results with domain knowledge will be necessary for facilitating the translation of machine learning into real-world results in biology and healthcare.

It is important to remark, however, that the technical aspects of robustness and generalization are only one piece of the puzzle in the development of biomedical machine learning. Some of the additional challenges that we need to consider are social in nature, and speak of our relationship with automated systems. Given what is at stake in the clinical setting, how do we know that we can trust the predictions of some cutting-edge algorithm? Building such trust is a daunting exercise in good validation practices, the use of explainable machine learning techniques, and clear communication that can involve all the stakeholders to clarify the nuances of the matter. Explaining and justifying the workings of a model has become increasingly important in the biomedical field [192], and it is likely that this trend will continue; ultimately, a robust

explanation of a machine learning model can be significantly more convincing than a direct quantitative evaluation, and it can be communicated far more effectively to policymakers and clinicians. Obtaining such a robust explanation is, of course, an extremely complex undertaking.

When interpretability is not an option, it may be sufficient to perform extensive validation of the models to offer convincing evidence for its translation into practice. Part of this thesis explored why validation is such a challenge in the current landscape of biomedical machine learning research (Section 3.3). With the considerable constraints on gathering and sharing many types of biomedical data, external validation is adopted in a small minority of the biomedical ML publications, despite being the most robust form of evaluation.

The necessity of demonstrating the performance of a model and engendering trust in its robustness is deeply linked to regulatory and socioeconomic factors that must be addressed for the fruitful translation of biomedical ML to the clinical practice. Such factors include the protection of privacy [470], legal liability [471], regulatory approval by the relevant organizations [472], fairness [473], and building trust through transparency with all the relevant parties [474].

Overcoming the challenges presented will require a convergence of efforts, ranging from the development of best practices for *in silico* validation, to the improvement of international collaborations that are the source of many large-scale databases which play a crucial role in biomedical ML.

6.2 Key contributions and limitations

The main research contributions presented in this thesis constitute Chapters 4 and 5. The remaining parts of the thesis aim to present a high-level picture of the challenges involved both on the side of machine learning and the biomedical domain.

6.2.1 Epigenomic imputation

In Chapter 4, I presented the development and extensive validation of the eDICE model. Within the broad context of the Roadmap reference epigenomes, eDICE proved remarkably successful, learning contextualized representations of epigenomic data and outperforming competitors on most measures. The use of self-attention blocks enabled the models to include interactions between tissues and between epigenetic modifications,

which allowed eDICE to produce high-quality imputations for unperformed assay even with a reduced amount of training data. The case study on the EN-TEx dataset explored how transfer learning can be used for the imputation of individual-specific peaks; by having the model learn a representation for all tissues from a source patient, we can perform zero-shot predictions on a target tissue for a target patient outperforming model-free baselines. To the best of my knowledge, this is the first application of deep learning to personalized epigenomics, a novel research direction that is sure to see increased interest in the coming years.

The experiments performed with eDICE present obvious limitations, which relate mostly to the quantity and quality of data available. For starters, the aggregation of epigenomic measurements in the Roadmap and ENCODE datasets aims to reconstruct a reference landscape of the human epigenome, akin to the reference human genomes to which DNA sequences are mapped. However, epigenetic marks are affected by various factors, including the genetics of the donor, their age, environmental exposure, and other hereditary factors. As such, a proper picture of the human epigenome would need to rely on more robust sets of experiments that explicitly account for all these additional factors. In this sense, the EN-TEx dataset is actually a more consistent set of measurements, given the numerous samples taken from each donor; however, the fact that it only includes four individuals is a strong limitation on the robustness of the conclusions that we can obtain.

Such limitations are expected when analysing biological mechanisms of such exquisite complexity as the epigenetic regulation of the genome. Large-scale efforts to gather epigenetic data have started only around 2003 [475, 232], with the technologies involved undergoing considerable improvements over the two decades passed since then. I expect that the issues caused by limited data will be overcome in time, with collaborative efforts to gather additional data to train even better models.

6.2.2 Antimicrobial resistance prediction

The analysis presented in Chapter 5 provides a strong argument for the continued exploration of MALDI-TOF spectrometry for the timely prediction of antimicrobial resistance. MALDI-TOF spectrometry has many appealing qualities, as experiments with this technology are cost-effective and scalable. Additionally, MALDI-TOF spectrometry has seen considerable growth in its use for pathogen identification [476–478], meaning that the instruments to perform this type of mass spectrometry are already available in many clinical settings. Most important of all is the fact that obtaining

information on the proteomic content of a pathogen through mass spectrometry is considerably faster than alternative workflows that include genetic analyses.

My colleagues and I have demonstrated how joint multimodal learning over chemical features and mass spectra enables the use of machine learning to recommend effective drugs, a task in which the proposed ResMLP model shows promising performance. The extensive analysis of how different data-generating processes affect our models, paired with ablation experiments and explainability techniques, offered insights into the challenges faced by our models in different generalization setups. Such insights can be leveraged to improve the model architecture in future iterations, and to guide data collection efforts to obtain highly informative samples for improving machine learning approaches for AMR prediction.

All the aforementioned advantages of MALDI-TOF spectrometry, however, come at the cost of the quality of the information that can be obtained on a pathogenic sample. A MALDI-TOF spectrum carries information about the proteins and peptides contained in the sample; however, this information comes in the form of a histogram of the mass-to-charge ratio of these biomolecules. As such, we obtain a "low-resolution" picture of the proteomics content of the pathogen, which cannot cover all the relevant information.

This limitation, in turn, means that a MALDI-TOF spectrum is suitable to recognize only certain pathways of resistance, as the wide array of resistance mechanisms includes complex processes that can have different effects on the proteome of the pathogen [479, 480]. Some resistance mechanisms, such as changes in the activity of beta-lactamases, can produce a clear signal in the mass spectra [481], and are particularly suitable for MALDI-TOF detection. Other mechanisms might not affect significantly the mass profile of the proteome of the pathogen, and therefore they might be difficult to identify through MALDI-TOF spectrometry; for example, upregulation of efflux pumps is not generally detectable in MALDI-TOF spectrometry [482], and the modification of the target sites of antibiotics [483] might lead to subtle changes that do not affect the mass of an enzyme but severely impact the effectiveness of the drug.

Aside from these intrinsic limitations of the methodology, our study focused on certain aspects of the prediction of antimicrobial resistance (such as the recommendation of effective drugs, or the impact of different data-generating processes), while other avenues of research were not yet analysed. Primarily the analysis of the generalization performance between different collection sites, or the aggregations of the separate

datasets into a cohesive whole, have not yet been examined with the required attention. Additionally, we relied on a relatively simplistic encoding of antimicrobial drugs by using molecular fingerprints. Foundation models for learning rich representations of chemical compounds are becoming increasingly available, and open up new opportunities to improve the effectiveness of models such as the ResMLP. These promising research directions are currently being explored.

Overall, MALDI-TOF spectrometry appears to be a valuable tool to combat the growing issue of antimicrobial resistance. While it cannot offer a comprehensive solution on its own, I expect that the integration of MALDI-TOF spectrometry into multimodal systems can be a key contributor to developing robust solutions for combatting antimicrobial resistance.

6.3 Personal reflections

When I started my Ph.D., the field of machine learning was well on its way to showing great potential. Starting from the early successes of models such as SVMs and LSTMs just three decades ago, machine learning models begun to show a glimpse of just how powerful the modelization of data could be to solve very different problems. Afterwards, with the rise of deep learning (starting from AlexNet and VAEs, to GANs, AlphaGo, and BERT), there was an expectation that machine learning would really be the innovative paradigm that could fully leverage our technological advancements.

It was with high expectations, then, that I walked through the doors of the Max-Planck Institute on the 1st of December 2019, ready to make my mark. However, even my excitement in joining the machine learning research community could not predict the pace of developments that happened since. The advancements of models like AlphaFold, CLIP, and GPT have transformed the landscape of research during my doctoral studies, opening up unprecedented opportunities and raising critical challenges.

My motivation to work on biomedical research using machine learning led me to glimpse the sheer possibilities at hand: just what could we achieve, if we learn to precisely model the most complex diseases, if we could perfectly tailor therapies to each individual, if we could decode the biological machinery of the cell? The more I learned about the biological sciences and their stunning complexity, however, the more I felt lost in Borges' Library of Babel, a world of uncountable possibilities where only a few make sense. As the author writes: "The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible, seemed almost

intolerable." Similarly, in biological systems, the vastness of the space of possible states seems like an insurmountable challenge. For instance, there is an inconceivable number of possible sequences of amino acids, some of which represent revolutionary drugs that may completely alter our lives; among the countless variations of the human genome, some can explain the mechanisms of certain diseases.

The ever-accelerating development of machine learning might just provide some of the tools we need to navigate the uncountable configurations of molecular biology; perhaps, through it, we will be able to find some of the "meaningful books," which will take the form of an RNA sequence that prevents a terrible infectious disease, a chemical formula that enables cheap and equitable access to effective medications, or maybe a biomarker that can be screened for to prevent enormous suffering.

With the current pace of progress in machine learning, it is difficult to predict just how successful the endeavours of biomedical AI will be, but I have become increasingly optimistic in this regard. At the same time, I realized first-hand the challenges involved, which made me appreciate even more the complex ecosystem that supports these advancements.

In biology—and healthcare in particular—collaborative efforts to gather and process data are absolutely vital to enabling any kind of advanced machine learning. When I also considered the multidisciplinary interactions needed between domain experts and machine learning researchers, I came to realize just how much the progress of biomedical machine learning is dependent on collaboration. I have high expectations for the progress that we can make in biomedicine through the use of machine learning; however, it is just as important that we face all the challenges involved, and we do not ignore that research is ultimately conducted by humans. We need to foster collaboration, encourage open research and good scientific practice, ensure equitable access to the fruits of these developments, and demand fairness in the socioeconomic impacts of machine learning. It will require considerable efforts from all of us, but only in this way we can try to realize the lofty promises of AI for a better future.

I look forward to being surprised again and again.

References

[1] Knorr W. R. On the early history of axiomatics: The interaction of mathematics and philosophy in greek antiquity. In *Theory Change, Ancient Axiomatics, and Galileo's Methodology: Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science Volume I*, pages 145–186. Springer, 1980.

[2] McCaskey J. P. *Induction*, pages 1–6. Springer International Publishing, Cham, 2016.

[3] Stadler F. *Induction and Deduction in the Philosophy of Science: A Critical Account since the Methodenstreit*, pages 1–15. Springer Netherlands, Dordrecht, 2004.

[4] Mill J. S. A system of logic. In *Arguing About Science*, pages 243–267. Routledge, 2012.

[5] Popper K. R. Science as falsification. *Conjectures and Refutations*, 1(1963):33–39, 1963.

[6] Jordan M. I. and Mitchell T. Machine learning: Trends, perspectives, and prospects. *Science*, 349:255 – 260, 2015.

[7] Kennedy B., Tyson A., and Saks E. Public awareness of artificial intelligence in everyday activities, 2023.

[8] Badue C., Guidolini R., Carneiro R. V., et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

[9] Wu S., Wang C., Cao H., et al. Crime prediction using data mining and machine learning. *Advances in Intelligent Systems and Computing*, 2018.

[10] Sarker I. H., Furhad M. H., and Nowrozy R. AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, 2, 2021.

[11] Varshney K. R. and Alemzadeh H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, 5(3):246–255, 2017.

[12] Jaeger P. F., Lüth C. T., Klein L., et al. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Goetz L., Seedat N., Vandersluis R., et al. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *npj Digital Medicine*, 7(1):126, 2024.

[14] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[15] Jensen A. R. The relationship between learning and intelligence. *Learning and Individual Differences*, 1:37–62, 1989.

[16] Schölkopf B. and von Kügelgen J. From statistical to causal learning. In *Proceedings of the International Congress of Mathematicians*, page 1, 2022.

[17] Evans H. and Snead D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *npj Digital Medicine*, 7(1):89, 2024.

[18] Obermeyer Z., Powers B. W., Vogeli C., et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447 – 453, 2019.

[19] Ellahham S., Ellahham N., and Simsekler M. C. E. Application of artificial intelligence in the health care safety context: opportunities and challenges. *American Journal of Medical Quality*, 35(4):341–348, 2020.

[20] Pearson K. National life from the standpoint of science. In *Scientific and Medical Knowledge Production, 1796-1918*, pages 281–287. Routledge, 2023.

[21] Bodmer W., Bailey R., Charlesworth B., et al. The outstanding scientist, RA Fisher: his views on eugenics and race. *Heredity*, 126(4):565–576, 2021.

[22] Sun W., Lee J., Zhang S., et al. Engineering precision medicine. *Advanced Science*, 6, 2018.

[23] Hawkins-Hooker A., Visonà G., Narendra T., et al. Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning. *Nature Communications*, 14(1):4750, 2023.

[24] Visonà G., Duroux D., Miranda L., et al. Multimodal learning in clinical proteomics: enhancing antimicrobial resistance prediction models with chemical information. *Bioinformatics*, 39(12):btad717, 2023.

[25] Vapnik V. N. Principles of risk minimization for learning theory. In *Neural Information Processing Systems*, 1991.

[26] Vapnik V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 5:988–99, 1999.

[27] Liu J., Shen Z., He Y., et al. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[28] Huang Y. and Gottardo R. Comparability and reproducibility of biomedical data. *Briefings in Bioinformatics*, 14:391 – 401, 2012.

[29] Zhang A., Xing L., Zou J., et al. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6:1330 – 1345, 2022.

[30] Bond-Taylor S., Leach A., Long Y., et al. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2021.

[31] Hastie T., Tibshirani R., Friedman J. H., et al. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[32] Baxter J. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12:149–198, 2000.

[33] Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.

[34] James W. and Stein C. *Estimation with Quadratic Loss*, pages 443–460. Springer New York, New York, NY, 1992.

[35] Moradi R., Berangi R., and Minaei B. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53:3947 – 3986, 2019.

[36] Ganjisaffar Y., Caruana R., and Lopes C. V. Bagging gradient-boosted trees for high precision, low variance ranking models. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[37] Langley P. Machine learning as an experimental science. *Machine Learning*, 3(1):5–8, 1988.

[38] Drummond C. Machine learning as an experimental science (revisited). In *AAAI Workshop on Evaluation Methods for Machine Learning*, pages 1–5. AAAI Press Menlo Park, CA, USA, 2006.

[39] Fradkov A. L. Early history of machine learning. *IFAC-PapersOnLine*, 2020.

[40] McCulloch W. S. and Pitts W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[41] Rosenblatt F. *The perceptron, a perceiving and recognizing automaton (Project Para)*. Cornell Aeronautical Laboratory, 1957.

[42] Hubel D. H. and Wiesel T. N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[43] Bozinovski S. and Fulgosi A. The influence of pattern similarity and transfer learning upon training of a base perceptron B2. In *Proceedings of Symposium Informatica*, volume 3, pages 121–126, 1976.

[44] Pratt L. Y. Discriminability-based transfer between neural networks. In *Neural Information Processing Systems*, 1992.

[45] Thrun S. and Pratt L., editors. *Learning to learn*. Springer, New York, NY, Oct. 2012.

[46] Galanti T., Wolf L., and Hazan T. A theoretical framework for deep transfer learning. *Information and Inference: A Journal of the IMA*, 5:159–209, 2016.

[47] McNamara D. and Balcan M.-F. Risk bounds for transferring representations with and without fine-tuning. In *International Conference on Machine Learning*, 2017.

[48] Tripuraneni N., Jordan M. I., and Jin C. On the theory of transfer learning: The importance of task diversity. In Larochelle H., Ranzato M., Hadsell R., et al., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[49] Cody T. and Beling P. A. A systems theory of transfer learning. *IEEE Systems Journal*, 17:26–37, 2021.

[50] Ioffe S. and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[51] Bjorck N., Gomes C. P., Selman B., et al. Understanding batch normalization. *Advances in Neural Information Processing Systems*, 31, 2018.

[52] Smith L. N. and Topin N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications*, volume 11006, pages 369–386. SPIE, 2019.

[53] Luo P., Wang X., Shao W., et al. Towards understanding regularization in batch normalization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[54] Santurkar S., Tsipras D., Ilyas A., et al. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 31, 2018.

[55] Vapnik V. N. *Statistical learning theory*. Wiley, 1998.

[56] Kearns M. J. and Vazirani U. *An Introduction to Computational Learning Theory*. The MIT Press, 08 1994.

[57] Vapnik V. N. and Chervonenkis A. Y. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pages 11–30. Springer International Publishing, Cham, 2015.

[58] Smith A. F. M. and Spiegelhalter D. J. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society Series B - Methodological*, 42:213–220, 1980.

[59] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[60] Schwarz G. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[61] Bartlett P. L., Harvey N., Liaw C., et al. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:63:1–63:17, 2017.

[62] Dziugaite G. K., Drouin A., Neal B., et al. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

[63] Bartlett P. L. and Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[64] Valiant L. G. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984.

[65] Bartlett P. L., Foster D. J., and Telgarsky M. J. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[66] Bousquet O. and Elisseeff A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[67] Feldman V. and Vondrak J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Beygelzimer A. and Hsu D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 25–28 Jun 2019.

[68] Arora S., Ge R., Neyshabur B., et al. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[69] Zhou W., Veitch V., Austern M., et al. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2019.

[70] Lotfi S., Finzi M., Kapoor S., et al. PAC-Bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

[71] Sefidgaran M., Gohari A., Richard G., et al. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In Loh P. and Raginsky M., editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4416–4463. PMLR, 2022.

[72] Park S., Bastani O., Matni N., et al. PAC confidence sets for deep neural networks via calibrated prediction. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[73] Kawaguchi K., Bengio Y., and Kaelbling L. *Generalization in Deep Learning*, pages 112 – 148. Cambridge University Press, 2022.

[74] Geiger M., Petrini L., and Wyart M. Landscape and training regimes in deep learning. *Physics Reports*, 924:1–18, 2021.

[75] Bach F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[76] Choromanska A., Henaff M., Mathieu M., et al. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204. PMLR, 2015.

[77] Li Y. and Liang Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In Bengio S., Wallach H. M., Larochelle H., et al., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8168–8177, 2018.

[78] Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

[79] Geman S., Bienenstock E., and Doursat R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[80] OpenAI. GPT-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[81] Belkin M., Hsu D. J., Ma S., et al. Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849 – 15854, 2018.

[82] Dar Y., Muthukumar V., and Baraniuk R. A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. *ArXiv*, abs/2109.02355, 2021.

[83] Singh S. P., Lucchi A., Hofmann T., et al. Phenomenology of double descent in finite-width neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[84] Du S. S., Zhai X., Póczos B., et al. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[85] Nakkiran P., Venkat P., Kakade S. M., et al. Optimal regularization can mitigate double descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[86] Bommasani R., Hudson D. A., Adeli E., et al. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[87] Krizhevsky A., Sutskever I., and Hinton G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[88] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[89] Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[90] Bahdanau D., Cho K., and Bengio Y. Neural machine translation by jointly learning to align and translate. In Bengio Y. and LeCun Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[91] Patwardhan N., Marrone S., and Sansone C. Transformers in the real world: A survey on NLP applications. *Information*, 14(4):242, 2023.

[92] Dosovitskiy A., Beyer L., Kolesnikov A., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[93] Wen Q., Zhou T., Zhang C., et al. Transformers in time series: A survey. In Elkind E., editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6778–6786. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Survey Track.

[94] Dong L., Xu S., and Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[95] Lee J., Lee Y., Kim J., et al. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[96] He K., Zhang X., Ren S., et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[97] Orhan E. and Pitkow X. Skip connections eliminate singularities. In *International Conference on Learning Representations*, 2018.

[98] Zhou Z., Siddiquee M. M. R., Tajbakhsh N., et al. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39:1856–1867, 2019.

[99] Liu F., Ren X., Zhang Z., et al. Rethinking skip connection with layer normalization. In *International Conference on Computational Linguistics*, 2020.

[100] Oyedotun O. K., Ismaeil K. A., and Aouada D. Training very deep neural networks: Rethinking the role of skip connections. *Neurocomputing*, 441:105–117, 2021.

[101] Furusho Y. and Ikeda K. Theoretical analysis of skip connections and batch normalization from generalization and optimization perspectives. *APSIPA Transactions on Signal and Information Processing*, 9:e9, 2020.

[102] He F., Liu T., and Tao D. Why ResNet works? Residuals generalize. *IEEE Transactions on Neural Networks and Learning Systems*, 31:5349–5362, 2019.

[103] Ruder S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[104] Boyd S. P. and Vandenberghe L. *Convex optimization*. Cambridge University Press, 2004.

[105] Hochreiter S. and Schmidhuber J. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[106] Keskar N. S., Mudigere D., Nocedal J., et al. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[107] Huang W. R., Emam Z., Goldblum M., et al. Understanding generalization through visualizations. In Zosa Forde J., Ruiz F., Pradier M. F., et al., editors, *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pages 87–97. PMLR, 12 Dec 2020.

[108] Chaudhari P., Choromanska A., Soatto S., et al. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[109] Dziugaite G. K. and Roy D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Elidan G., Kersting K., and Ihler A., editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

[110] Hardt M., Recht B., and Singer Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[111] Kuzborskij I. and Lampert C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

[112] Bassily R., Feldman V., Guzmán C., et al. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[113] Martin C. H. and Mahoney M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *J. Mach. Learn. Res.*, 22:165:1–165:73, 2018.

[114] Advani M. S. and Saxe A. M. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428 – 446, 2017.

[115] Cao Y. and Gu Q. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

[116] Opper M. and Kinzel W. *Statistical Mechanics of Generalization*, pages 151–209. Springer New York, New York, NY, 1996.

[117] Jacot A., Gabriel F., and Hongler C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[118] Arora S., Du S. S., Hu W., et al. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[119] Canatar A., Bordelon B., and Pehlevan C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021.

[120] Lee J., Xiao L., Schoenholz S., et al. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

[121] Lampinen A. K. and Ganguli S. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2019.

[122] Wenzel F., Dittadi A., Gehler P., et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

[123] Rahimian H. and Mehrotra S. Frameworks and results in distributionally robust optimization. *Open J. Math. Optim.*, 3:1–85, 2019.

[124] Mustafa W., Lei Y., and Kloft M. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, 2022.

[125] Settles B. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[126] Pan S. J. and Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[127] Weiss K., Khoshgoftaar T. M., and Wang D. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

[128] Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[129] Tahmoresnezhad J. and Hashemi S. Visual domain adaptation via transfer feature learning. *Knowledge and Information Systems*, 50:585–605, 2017.

[130] Redko I., Morvant E., Habrard A., et al. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

[131] Chapelle O., Schölkopf B., and Zien A. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[132] McCloskey M. and Cohen N. J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[133] Zhang W., Deng L., Zhang L., et al. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10:305–329, 2020.

[134] Malik H., Farooq M. S., Khelifi A., et al. A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging. *IEEE Access*, 8:139367–139386, 2020.

[135] Bois M. D., El-Yacoubi M. A., and Ammi M. Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine*, 199:105874, 2020.

[136] Wan Z., Yang R., Huang M., et al. A review on transfer learning in EEG signal analysis. *Neurocomputing*, 421:1–14, 2021.

[137] Valverde J. M., Imani V., Abdollahzadeh A., et al. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of Imaging*, 7, 2021.

[138] Schweikert G. B., Widmer C., Schölkopf B., et al. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Neural Information Processing Systems*, 2008.

[139] Lu J., Behbood V., Hao P., et al. Transfer learning using computational intelligence: A survey. *Knowl. Based Syst.*, 80:14–23, 2015.

[140] Zhuang F., Qi Z., Duan K., et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[141] Iman M., Arabnia H. R., and Rasheed K. A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40, 2023.

[142] Blanchard G., Lee G., and Scott C. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, 24, 2011.

[143] Muandet K., Balduzzi D., and Schölkopf B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[144] Wang J., Lan C., Liu C., et al. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2022.

[145] Zhou K., Liu Z., Qiao Y., et al. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

[146] Hendrycks D. and Dietterich T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[147] Caruana R. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[148] Maurer A., Pontil M., and Romera-Paredes B. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[149] Gulrajani I. and Lopez-Paz D. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[150] Li H., Pan S. J., Wang S., et al. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[151] Li Y., Tian X., Gong M., et al. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[152] Kim D., Park S., Kim J., et al. SelfReg: Self-supervised contrastive regularization for domain generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9599–9608, 2021.

[153] Ballas A. and Diou C. Towards domain generalization for ECG and EEG classification: Algorithms and benchmarks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8:44–54, 2023.

[154] Zhang L., Wang X., Yang D., et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39:2531–2540, 2020.

[155] Li H., Wang Y., Wan R., et al. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020.

[156] Zhang H., Dullerud N., Seyyed-Kalantari L., et al. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 279–290, 2021.

[157] Jose S. T. and Simeone O. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23, 2020.

[158] Hospedales T. M., Antoniou A., Micaelli P., et al. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:5149–5169, 2020.

[159] Vanschoren J. *Meta-Learning*, pages 35–61. Springer International Publishing, Cham, 2019.

[160] Breiman L. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[161] Schapire R. E. and Freund Y. Boosting: Foundations and algorithms. *Kybernetes*, 42(1):164–166, 2013.

[162] Finn C., Abbeel P., and Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[163] Santoro A., Bartunov S., Botvinick M. M., et al. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, 2016.

[164] Ortega P. A., Wang J. X., Rowland M., et al. Meta-learning of sequential strategies. *ArXiv*, abs/1905.03030, 2019.

[165] Ji K., Lee J. D., Liang Y., et al. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

[166] Yin M., Tucker G., Zhou M., et al. Meta-learning without memorization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[167] Rajendran J., Irpan A., and Jang E. Meta-learning requires meta-augmentation. In Larochelle H., Ranzato M., Hadsell R., et al., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[168] Choe S. K., Mehta S. V., Ahn H., et al. Making scalable meta learning practical. In Oh A., Naumann T., Globerson A., et al., editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[169] Mahajan P., Uddin S., Hajati F., et al. Ensemble learning for disease prediction: A review. *Healthcare*, 11, 2023.

[170] Zhang X. S., Tang F., Dodge H. H., et al. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[171] Wang J., Zheng S., Chen J., et al. Meta learning for low-resource molecular optimization. *Journal of Chemical Information and Modeling*, 2021.

[172] Tian Y., Zhao X., and Huang W. Meta-learning approaches for learning-to-learn in deep learning: A survey. *Neurocomputing*, 494:203–223, 2022.

[173] Wang W., Zheng V. W., Yu H., et al. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[174] Pourpanah F., Abdar M., Luo Y., et al. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4051–4070, 2022.

[175] Chao W.-L., Changpinyo S., Gong B., et al. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 52–68. Springer, 2016.

[176] Rezaei M. and Shahidi M. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review. *Intelligence-based Medicine*, 3:100005, 2020.

[177] Xian Y., Lampert C. H., Schiele B., et al. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018.

[178] Hayat N., Lashen H., and Shamout F. E. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In *Machine Learning for Healthcare Conference*, pages 461–477. PMLR, 2021.

[179] Mahapatra D., Bozorgtabar B., and Ge Z. Medical image classification using generalized zero shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3344–3353, 2021.

[180] Sivarajkumar S. and Wang Y. HealthPrompt: A zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, page 972. American Medical Informatics Association, 2022.

[181] Kulmanov M. and Hoehndorf R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics*, 38(Supplement_1):i238–i245, 2022.

[182] Schölkopf B. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022.

[183] Gopnik A., Glymour C., Sobel D. M., et al. A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, 111(1):3, 2004.

[184] Zhang C., Zhang K., and Li Y. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33:289–301, 2020.

[185] Pearl J. *Causality*. Cambridge University Press, 2 edition, 2009.

[186] Rubin D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[187] Schölkopf B., Janzing D., Peters J., et al. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1255–1262, New York, NY, USA, 2012. Omnipress.

[188] Peters J., Janzing D., and Schölkopf B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[189] Parascandolo G., Kilbertus N., Rojas-Carulla M., et al. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[190] Rojas-Carulla M., Schölkopf B., Turner R. E., et al. Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, 19:36:1–36:34, 2015.

[191] Sanchez P., Voisey J. P., Xia T., et al. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

[192] Malinverno L., Barros V., Ghisoni F., et al. A historical perspective of biomedical explainable AI research. *Patterns*, 4(9), 2023.

[193] Guo R., Cheng L., Li J., et al. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

[194] Schölkopf B., Locatello F., Bauer S., et al. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[195] Lv F., Liang J., Li S., et al. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022.

[196] Deleu T., Góis A., Emezue C., et al. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pages 518–528. PMLR, 2022.

[197] Scherrer N., Goyal A., Bauer S., et al. On the generalization and adaption performance of causal models. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

[198] Kilbertus N., Parascandolo G., and Schölkopf B. Generalization in anti-causal learning. In *NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning*, Dec. 2018.

[199] Cheng L., Guo R., Moraffah R., et al. A practical data repository for causal learning with big data. In *BenchCouncil International Symposium*, 2019.

[200] Pavlović M., Al Hajj G. S., Kanduri C., et al. Improving generalization of machine learning-identified biomarkers using causal modelling with examples from immune receptor diagnostics. *Nature Machine Intelligence*, 6(1):15–24, 2024.

[201] Lopez R., Tagasovska N., Ra S., et al. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023.

[202] Feuerriegel S., Frauen D., Melnychuk V., et al. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[203] Seeman N. C., Rosenberg J. M., and Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences*, 73(3):804–808, 1976.

[204] D'haeseleer P. What are DNA sequence motifs? *Nature Biotechnology*, 24(4):423–425, 2006.

[205] Nagy G. and Nagy L. Motif grammar: The basis of the language of gene expression. *Computational and Structural Biotechnology Journal*, 18:2026–2032, 2020.

[206] Sasse A., Laverty K. U., Hughes T. R., et al. Motif models for RNA-binding proteins. *Current Opinion in Structural Biology*, 53:115–123, 2018.

[207] Desaphy J., Raimbaud E., Ducrot P., et al. Encoding protein-ligand interaction patterns in fingerprints and graphs. *Journal of Chemical Information and Modeling*, 53 3:623–37, 2013.

[208] Danilova N. The evolution of immune mechanisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306(6):496–520, 2006.

[209] Flajnik M. F. and Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, 2010.

[210] Rath D., Amlinger L., Rath A., et al. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*, 117:119–128, 2015.

[211] Swarts D. C., Makarova K., Wang Y., et al. The evolutionary journey of Argonaute proteins. *Nature Structural & Molecular Biology*, 21(9):743–753, 2014.

[212] Amarante-Mendes G. P., Adjemian S., Branco L. M., et al. Pattern recognition receptors and the host cell death molecular machinery. *Frontiers in Immunology*, 9:417707, 2018.

[213] Herzog S., Reth M., and Jumaa H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. *Nature Reviews Immunology*, 9(3):195–205, 2009.

[214] Rossjohn J., Gras S., Miles J. J., et al. T cell antigen receptor recognition of antigen-presenting molecules. *Annual Review of Immunology*, 33:169–200, 2015.

[215] Ferapontov A., Omer M., Baudrexel I., et al. Antigen footprint governs activation of the B cell receptor. *Nature Communications*, 14(1):976, 2023.

[216] Rosenblatt A. D. and Thickstun J. T. Intuition and consciousness. *The Psychoanalytic Quarterly*, 63(4):696–714, 1994.

[217] Newen A., Welpinghus A., and Juckel G. Emotion recognition as pattern recognition: the relevance of perception. *Mind & Language*, 30(2):187–208, 2015.

[218] Cavalli G. and Heard E. Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766):489–499, 2019.

[219] Bird A. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, 2002.

[220] Lennartsson A. and Ekwall K. Histone modification patterns and epigenetic codes. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1790(9):863–868, 2009.

[221] Sapoval N., Aghazadeh A., Nute M. G., et al. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, 2022.

[222] Bryant P., Pozzati G., and Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1):1265, 2022.

[223] Lin Z., Akin H., Rao R., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[224] Kulmanov M., Khan M. A., and Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2018.

[225] Kulmanov M. and Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

[226] Subramanian I., Verma S., Kumar S., et al. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14:1177932219899051, 2020.

[227] Ker J., Wang L., Rao J. P., et al. Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9389, 2018.

[228] Cuocolo R., Cipullo M. B., Stanzione A., et al. Machine learning applications in prostate cancer magnetic resonance imaging. *European Radiology Experimental*, 3, 2019.

[229] Scheeder C., Heigwer F., and Boutros M. Machine learning and image-based profiling in drug discovery. *Current Opinion in Systems Biology*, 10:43 – 52, 2018.

[230] Simm J., Klambauer G., Arany A., et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chemical Biology*, 25(5):611–618, 2018.

[231] Harrison P. W., Amode M. R., Austine-Orimoloye O., et al. Ensembl 2024. *Nucleic Acids Research*, 52:D891 – D899, 2023.

[232] Feingold E. A., Good P. J., Guyer M., et al. The ENCODE (encyclopedia of DNA elements) project. *Science*, 306:636 – 640, 2004.

[233] Harrow J. L., Frankish A., Gonzalez J. M., et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22:1760 – 1774, 2012.

[234] Bernstein B. E., Stamatoyannopoulos J. A., Costello J. F., et al. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

[235] Consortium T. U. UniProt: a hub for protein information. *Nucleic Acids Research*, 43:D204 – D212, 2014.

[236] Berman H. M., Westbrook J., Feng Z., et al. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[237] Tomczak K., Czerwińska P., and Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19:A68 – A77, 2015.

[238] König I. R., Fuchs O., Hansen G., et al. What is precision medicine? *European Respiratory Journal*, 50(4), 2017.

[239] Kosorok M. R. and Laber E. B. Precision medicine. *Annual Review of Statistics and its Application*, 6:263–286, 2019.

[240] Council N. R. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies Press, Washington, DC, 2011.

[241] MacEachern S. J. and Forkert N. D. Machine learning for precision medicine. *Genome*, 64(4):416–425, 2021.

[242] Ahsan M. M., Luna S. A., and Siddique Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 2022.

[243] Uddin S., Khan A., Hossain M. E., et al. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19, 2019.

[244] Adam G., Rampášek L., Safikhani Z., et al. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precision Oncology*, 4, 2020.

[245] Al-Tashi Q., Saad M. B., Muneer A., et al. Machine learning models for the identification of prognostic and predictive cancer biomarkers: A systematic review. *International Journal of Molecular Sciences*, 24, 2023.

[246] Kavalci E. and Hartshorn A. S. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Scientific Reports*, 13, 2023.

[247] Dara S., Dhamercherla S., Jadav S. S., et al. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, 55:1947 – 1999, 2021.

[248] Joshi G., Jain A., Araveeti S. R., et al. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: An updated landscape. *Electronics*, 13(3):498, 2024.

[249] Keane P. A. and Topol E. J. With an eye to AI and autonomous diagnosis. *npj Digital Medicine*, 1(1):40, 2018.

[250] Navarro C. L. A., Damen J. A., Takada T., et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*, 375, 2021.

[251] Roberts M., Driggs D., Thorpe M., et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[252] Wynants L., Van Calster B., Collins G. S., et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*, 369, 2020.

[253] Shamout F. E., Zhu T., and Clifton D. A. Machine learning for clinical outcome prediction. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2020.

[254] Carracedo-Reboredo P., Liñares-Blanco J., Rodriguez-Fernandez N., et al. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19:4538 – 4558, 2021.

[255] Salas-Vega S., Haimann A., and Mossialos E. Big data and health care: Challenges and opportunities for coordinated policy development in the EU. *Health Systems & Reform*, 1:285 – 300, 2015.

[256] Hoffman S. and Podgurski A. The use and misuse of biomedical data: Is bigger really better? *American Journal of Law & Medicine*, 39:497 – 538, 2013.

[257] Jain A., Patel H., Nagalapatti L., et al. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562, 2020.

[258] Hernán M. A., Hernández-Díaz S., and Robins J. M. A structural approach to selection bias. *Epidemiology*, 15:615–625, 2004.

[259] Choi H. K., Nguyen U.-S. D. T., Niu J., et al. Selection bias in rheumatic disease research. *Nature Reviews Rheumatology*, 10:403–412, 2014.

[260] Munafo M. R., Tilling K., Taylor A. E., et al. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47:226 – 235, 2016.

[261] Kahan B. C., Rehal S., and Cro S. Risk of selection bias in randomised trials. *Trials*, 16, 2015.

[262] Schoeler T., Speed D., Porcu E., et al. Participation bias in the uk biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227, 2023.

[263] Celi L. A., Cellini J., Charpignon M., et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1, 2022.

[264] Woo Y. and Li W.-H. Evolutionary conservation of histone modifications in mammals. *Molecular Biology and Evolution*, 29 7:1757–67, 2012.

[265] Chatterjee C. and Muir T. W. Chemical approaches for studying histone modifications. *The Journal of Biological Chemistry*, 285:11045 – 11050, 2010.

[266] Yun M., Wu J., Workman J. L., et al. Readers of histone modifications. *Cell Research*, 21:564–578, 2011.

[267] Suganuma T. and Workman J. L. Signals and combinatorial functions of histone modifications. *Annual Review of Biochemistry*, 80:473–99, 2011.

[268] Schaefer M. H., Serrano L., and Andrade-Navarro M. A. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Frontiers in Genetics*, 6:137790, 2015.

[269] Gillis J. A., Ballouz S., and Pavlidis P. Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *Journal of Proteomics*, 100:44–54, 2014.

[270] Schnoes A. M., Ream D. C., Thorman A. W., et al. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLOS Computational Biology*, 9, 2013.

[271] Wang J., Zhou X., Zhu J., et al. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11:290 – 290, 2010.

[272] Mcgauran N., Wieseler B., Kreis J., et al. Reporting bias in medical research - a narrative review. *Trials*, 11:37 – 37, 2010.

[273] Sterne J. A., Egger M., and Moher D. Addressing reporting biases. *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*, pages 297–333, 2008.

[274] Dwan K., Altman D. G., Arnaiz J. A. S., et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLOS ONE*, 3, 2008.

[275] Easterbrook P. J., Gopalan R., Berlin J., et al. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.

[276] Abbasi-Sureshjani S., Raumanns R., Michels B. E. J., et al. Risk of training diagnostic algorithms on data with demographic bias. In Cardoso J., Van Nguyen H., Heller N., et al., editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192, Cham, 2020. Springer International Publishing.

[277] Schaefer J., Lehne M., Schepers J., et al. The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15, 2020.

[278] Rajkomar A., Hardt M., Howell M. D., et al. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169:866–872, 2018.

[279] Boulesteix A.-L. Ten simple rules for reducing overoptimistic reporting in methodological computational research, 2015.

[280] Peng C.-K., Costa M., and Goldberger A. L. Adaptive data analysis of complex fluctuations in physiologic time series. *Advances in Adaptive Data Analysis*, 1 1:61–70, 2009.

[281] Stein R. A. and Mchaourab H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLOS Computational Biology*, 18(8):e1010483, 2022.

[282] Wright P. E. and Dyson H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular Cell Biology*, 16(1):18–29, 2015.

[283] Sahiner B., Chen W., Samala R. K., et al. Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology*, page 20220878, 2023.

[284] Jung K. and Shah N. H. Implications of non-stationarity on predictive modeling using EHRs. *Journal of Biomedical Informatics*, 58:168–174, 2015.

[285] Kukar M. Drifting concepts as hidden factors in clinical studies. In *Conference on Artificial Intelligence in Medicine in Europe*, 2003.

[286] Finlayson S. G., Subbaswamy A., Singh K., et al. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

[287] Lu J., Liu A., Dong F., et al. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31:2346–2363, 2019.

[288] Davis S. E., Greevy Jr R. A., Fonnesbeck C., et al. A nonparametric updating method to correct clinical prediction model drift. *Journal of the American Medical Informatics Association*, 26(12):1448–1457, 2019.

[289] Lee C. S. and Lee A. Y. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

[290] Billingham L., Malottki K., and Steven N. M. Small sample sizes in clinical trials: a statistician's perspective. *Clinical Investigation*, 2:655–657, 2012.

[291] Hee S. W., Willis A., Smith C. T., et al. Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? An analysis of data from the aggregate analysis of clinicaltrials.gov. *Orphanet Journal of Rare Diseases*, 12, 2017.

[292] Kianifard F. and Islam M. Z. A guide to the design and analysis of small clinical studies. *Pharmaceutical Statistics*, 10, 2011.

[293] Vabalas A., Gowen E., Poliakoff E., et al. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[294] Karp D. R., Carlin S., Cook-Deegan R. M., et al. Ethical and practical issues associated with aggregating databases. *PLOS Medicine*, 5, 2008.

[295] Abouelmehdi K., Hssane A. B., and Khaloufi H. Big healthcare data: preserving security and privacy. *Journal of Big Data*, 5, 2018.

[296] Huang W., Ye M., Shi Z., et al. Rethinking federated learning with domain shift: A prototype view. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322, 2023.

[297] Leek J. T., Scharpf R. B., Bravo H. C., et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[298] Kothari S., Phan J. H., Stokes T. H., et al. Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 18:765–772, 2014.

[299] Lazar C., Meganck S., Taminau J., et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics*, 14 4:469–90, 2013.

[300] Wang Y. and Cao K.-A. L. Managing batch effects in microbiome data. *Briefings in Bioinformatics*, 2019.

[301] Soneson C., Gerster S., and Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLOS ONE*, 9, 2014.

[302] Berisha V., Krantsevich C., Hahn P. R., et al. Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4, 2021.

[303] Chandrashekar G. and Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[304] Cai J., Luo J., Wang S., et al. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[305] Klawonn F., Höppner F., and Jayaram B. What are clusters in high dimensions and are they difficult to find? In *International Workshop on Clustering High-Dimensional Data*, 2012.

[306] Ray P., Reddy S. S., and Banerjee T. S. Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, 54:3473 – 3515, 2021.

[307] Ma S. and Dai Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, 12 6:714–22, 2011.

[308] Wall M. E., Rechtsteiner A., and Rocha L. M. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*, pages 91–109. Springer, 2003.

[309] Li M. M., Huang K., and Zitnik M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6:1353 – 1369, 2022.

[310] Iuchi H., Matsutani T., Yamada K., et al. Representation learning applications in biological sequence analysis. *Computational and Structural Biotechnology Journal*, 19:3198–3208, 2021.

[311] Si Y., Du J., Li Z., et al. Deep representation learning of patient data from electronic health records (EHR): A systematic review. *Journal of Biomedical Informatics*, 115:103671, 2021.

[312] Kobak D. and Berens P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10, 2018.

[313] Nielson J. L., Cooper S. R., Seabury S. A., et al. Statistical guidelines for handing missing data in traumatic brain injury clinical research. *Journal of Neurotrauma*, 2020.

[314] Wood A. M., White I. R., and Thompson S. G. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368 – 376, 2004.

[315] Diaz-Ordaz K., Kenward M. G., Cohen A., et al. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*, 11:590 – 600, 2014.

[316] Bell M. L., Fiero M. H., Horton N. J., et al. Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, 14, 2014.

[317] Rubin D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[318] Lin J.-H. and Haug P. J. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, 41 1:1–14, 2008.

[319] Sun B., Liu L., Miao W., et al. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28 4:1965–1983, 2016.

[320] Pereira R. C., Abreu P. H., and Rodrigues P. P. Partial multiple imputation with variational autoencoders: Tackling not at randomness in healthcare data. *IEEE Journal of Biomedical and Health Informatics*, 26:4218–4227, 2022.

[321] Graham J. W. Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60:549–76, 2009.

[322] Eekhout I., de Vet H. C. W., Twisk J. W. R., et al. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, 67 3:335–42, 2014.

[323] Kuhn M. and Johnson K. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[324] Perez-Lebel A., Varoquaux G., Morvan M. L., et al. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11, 2022.

[325] Paullada A., Raji I. D., Bender E. M., et al. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.

[326] DeMets D. L. Statistical issues in interpreting clinical trials. *Journal of Internal Medicine*, 255, 2004.

[327] Hoffmann F., Bertram T., Mikut R., et al. Benchmarking in classification and regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1318, 2019.

[328] Schreiber J., Boix C., wook Lee J., et al. The ENCODE Imputation Challenge: a critical assessment of methods for cross-cell type imputation of epigenomic profiles. *Genome Biology*, 24(1):79, Apr. 2023.

[329] Sylolypavan A., Sleeman D., Wu H., et al. The impact of inconsistent human annotations on AI driven clinical decision making. *npj Digital Medicine*, 6(1):26, 2023.

[330] Muller M., Wolf C. T., Andres J., et al. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[331] Baumgartner Jr W. A., Cohen K. B., Fox L. M., et al. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, 2007.

[332] Grouin C., Lavergne T., and Névéol A. Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 54–58, 2014.

[333] Albarqouni S., Baur C., Achilles F., et al. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35:1313–1321, 2016.

[334] Maier-Hein L., Mersmann S., Kondermann D., et al. Crowdsourcing for reference correspondence generation in endoscopic images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*, pages 349–356. Springer, 2014.

[335] Aroyo L. and Welty C. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[336] Goel A., Gueta A., Gilon O., et al. LLMs accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, 2023.

[337] Moor M., Banerjee O., Abad Z. S. H., et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[338] Bolton W. J., Poyiadzi R., Morrell E., et al. RAmBLA: A framework for evaluating the reliability of LLMs as assistants in the biomedical domain. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

[339] Karabacak M. and Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

[340] Chen R. J., Lu M. Y., Chen T. Y., et al. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[341] Rankin D., Black M. M., Bond R. R., et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, 8, 2020.

[342] Hwang S., Kim E., Lee I., et al. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5(1):17875, 2015.

[343] Hou L., Sun N., Mane S. M., et al. Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genetic Epidemiology*, 41:152 – 162, 2017.

[344] Yan Q., Chen R., Sutcliffe J. S., et al. The impact of genotype calling errors on family-based studies. *Scientific Reports*, 6, 2016.

[345] Krafczyk M., Shi A., Bhaskar A., et al. Scientific tests and continuous integration strategies to enhance reproducibility in the scientific software context. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, pages 23–28, 2019.

[346] Coelho L. P. For long-term sustainable software in bioinformatics. *PLOS Computational Biology*, 20(3):e1011920, 2024.

[347] Goecks J., Nekrutenko A., Taylor J., et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11:1–13, 2010.

[348] Di Tommaso P., Chatzou M., Floden E. W., et al. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[349] Köster J. and Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[350] Wratten L., Wilm A., and Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, 18:1161 – 1168, 2021.

[351] Kadri S., Sboner A., Sigaras A., et al. Containers in bioinformatics: applications, practical considerations, and best practices in molecular pathology. *The Journal of Molecular Diagnostics*, 24(5):442–454, 2022.

[352] Weber L. M., Saelens W., Cannoodt R., et al. Essential guidelines for computational method benchmarking. *Genome Biology*, 20:1–12, 2019.

[353] Angers-Loustau A., Petrillo M., Bengtsson-Palme J., et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Research*, 7, 2018.

[354] Button K. S., Ioannidis J. P. A., Mokrysz C., et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365–376, 2013.

[355] Bruns S. B. and Ioannidis J. P. A. p-curve and p-hacking in observational research. *PLOS ONE*, 11, 2016.

[356] Head M. L., Holman L., Lanfear R., et al. The extent and consequences of p-hacking in science. *PLOS Biology*, 13, 2015.

[357] Peng J., Jury E. C., Dönnes P., et al. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Frontiers in Pharmacology*, 12:720694, 2021.

[358] Stafford I. S., Kellermann M., Mossotto E., et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *npj Digital Medicine*, 3(1):30, 2020.

[359] Groot O. Q., Bindels B. J. J., Ogink P. T., et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthopaedica*, 92:385 – 393, 2021.

[360] Steyerberg E. W., Harrell Jr F. E., Borsboom G. J., et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–781, 2001.

[361] Riley R. D., Ensor J., Snell K. I., et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj*, 353, 2016.

[362] Ho S. Y., Phua K., Wong L., et al. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, 1(8), 2020.

[363] Ramspek C. L., Jager K. J., Dekker F. W., et al. External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1):49–58, 2021.

[364] Cabitza F., Campagner A., Soares F., et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208:106288, 2021.

[365] Debray T. P., Vergouwe Y., Koffijberg H., et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3):279–289, 2015.

[366] Steyerberg E. W. and Harrell Jr F. E. Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, 69:245, 2016.

[367] Martens F. K., Kers J. G., and Janssens A. C. J. External validation is only needed when prediction models are worth it (letter commenting on: J Clin Epidemiol. 2015; 68: 25-34). *Journal of Clinical Epidemiology*, 69:249–250, 2016.

[368] Bracher-Smith M., Crawford K., and Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry*, 26(1):70–79, 2021.

[369] Luo W., Phung D., Tran T., et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of Medical Internet Research*, 18(12):e323, 2016.

[370] Collins G. S., Reitsma J. B., Altman D. G., et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Circulation*, 131(2):211–219, 2015.

[371] Wolff R. F., Moons K. G., Riley R. D., et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019.

[372] Collins G. S., Dhiman P., Navarro C. L. A., et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7):e048008, 2021.

[373] Couckuyt A., Seurinck R., Emmaneel A., et al. Challenges in translational machine learning. *Human Genetics*, 141(9):1451–1466, 2022.

[374] Moons K. G., Kengne A. P., Grobbee D. E., et al. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart*, 98(9):691–698, 2012.

[375] Skinner M. K. Role of epigenetics in developmental biology and transgenerational inheritance. *Birth Defects Research Part C: Embryo Today: Reviews*, 93(1):51–55, 2011.

[376] Moosavi A. and Ardekani A. M. Role of epigenetics in biology and human diseases. *Iranian Biomedical Journal*, 20(5):246, 2016.

[377] Fardi M., Solali S., and Hagh M. F. Epigenetic mechanisms as a new approach in cancer treatment: An updated review. *Genes & Diseases*, 5(4):304–311, 2018.

[378] Mohammad H. P., Barbash O., and Creasy C. L. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. *Nature Medicine*, 25(3):403–418, 2019.

[379] Kronfol M. M., Dozmorov M. G., Huang R., et al. The role of epigenomics in personalized medicine. *Expert review of precision medicine and drug development*, 2(1):33–45, 2017.

[380] Dawson M. A. and Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell*, 150(1):12–27, 2012.

[381] Campbell R. M. and Tummino P. J. Cancer epigenetics drug discovery and development: the challenge of hitting the mark. *The Journal of Clinical Investigation*, 124(1):64–69, 2014.

[382] Nakamura M., Gao Y., Dominguez A. A., et al. CRISPR technologies for precise epigenome editing. *Nature Cell Biology*, 23(1):11–22, 2021.

[383] Nakade S., Yamamoto T., and Sakuma T. Cancer induction and suppression with transcriptional control and epigenome editing technologies. *Journal of Human Genetics*, 63(2):187–194, 2018.

[384] Ansari I., Chaturvedi A., Chitkara D., et al. CRISPR/Cas mediated epigenome editing for cancer therapy. In *Seminars in Cancer Biology*. Elsevier, 2021.

[385] Berson A., Nativio R., Berger S. L., et al. Epigenetic regulation in neurodegenerative diseases. *Trends in Neurosciences*, 41(9):587–598, 2018.

[386] Goyal D., Limesand S. W., and Goyal R. Epigenetic responses and the developmental origins of health and disease. *Journal of Endocrinology*, 242(1):T105–T119, 2019.

[387] Flavahan W. A., Gaskell E., and Bernstein B. E. Epigenetic plasticity and the hallmarks of cancer. *Science*, 357(6348), 2017.

[388] Consortium E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012.

[389] Kundaje A., Meuleman W., Ernst J., et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[390] Pashayan N., Reisel D., and Widschwendter M. Integration of genetic and epigenetic markers for risk stratification: opportunities and challenges. *Personalized Medicine*, 13(2):93–95, 2016.

[391] Baer-Dubowska W., Majchrzak-Celińska A., and Cichocki M. Pharmocoepigenetics: a new approach to predicting individual drug responses and targeting new drugs. *Pharmacological Reports*, 63(2):293–304, 2011.

[392] Tang J., Xiong Y., Zhou H.-H., et al. DNA methylation and personalized medicine. *Journal of Clinical Pharmacy and Therapeutics*, 39(6):621–627, 2014.

[393] Shastry B. S. Role of epigenomics in drug discovery and therapies. *Drug Development Research*, 73(8):513–517, 2012.

[394] Schreiber J., Bilmes J., and Noble W. S. Prioritizing transcriptomic and epigenomic experiments by using an optimization strategy that leverages imputed data. *Bioinformatics*, 09 2020. btaa830.

[395] Ernst J. and Kellis M. Large-scale epigenome imputation improves data quality and disease variant enrichment. *Nature Biotechnology*, 33(4):364, 2015.

[396] Durham T. J., Libbrecht M. W., Howbert J. J., et al. PREDICTD parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature Communications*, 9(1):1–15, 2018.

[397] Schreiber J., Durham T., Bilmes J., et al. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biology*, 21(1):1–18, 2020.

[398] Rozowsky J., Gao J., Borsari B., et al. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell*, 186(7):1493–1511.e40, Mar. 2023.

[399] Zhang Y., Liu T., Meyer C. A., et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):1–9, 2008.

[400] Steinhauser S., Kurzawa N., Eils R., et al. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, 17(6):953–966, 2016.

[401] Schweikert G., Cseke B., Clouaire T., et al. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC Genomics*, 14(1):1–17, 2013.

[402] Stark R. and Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. *R package version*, 100(4.3), 2011.

[403] Schurch N. J., Schofield P., Gierliński M., et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–51, Jun 2016.

[404] Jiang Y.-h., Bressler J., and Beaudet A. L. Epigenetics and human disease. *Annu. Rev. Genomics Hum. Genet.*, 5:479–510, 2004.

[405] Zoghbi H. Y. and Beaudet A. L. Epigenetics and human disease. *Cold Spring Harbor Perspectives in Biology*, 8(2):a019497, 2016.

[406] Feinberg A. P. The key role of epigenetics in human disease prevention and mitigation. *New England Journal of Medicine*, 378(14):1323–1334, 2018.

[407] Coyle K. M., Boudreau J. E., and Marcato P. Genetic mutations and epigenetic modifications: driving cancer and informing precision medicine. *BioMed Research International*, 2017, 2017.

[408] Dumitrescu R. G. Early epigenetic markers for precision medicine. *Cancer Epigenetics for Precision Medicine*, pages 3–17, 2018.

[409] Beltrán-García J., Osca-Verdegal R., Mena-Mollá S., et al. Epigenetic IVD tests for personalized precision medicine in cancer. *Frontiers in Genetics*, 10:621, 2019.

[410] Davis C. A., Hitz B. C., Sloan C. A., et al. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018.

[411] Saito T. and Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[412] Hoffman M. M., Buske O. J., Wang J., et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473, 2012.

[413] Vincent P., Larochelle H., Lajoie I., et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12), 2010.

[414] Srivastava N., Hinton G., Krizhevsky A., et al. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[415] Visonà G. Data for reproducing the training of eDICE model ("getting personal with epigenetics: Towards individual-specific epigenomic imputation with machine learning"), 2023.

[416] Hawkins-Hooker A., Visonà G., and Tanmayee Narendra. alex-hh/eDICE: Publication release, 2023.

[417] Murray C. J., Ikuta K. S., Sharara F., et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022.

[418] Benkova M., Soukup O., and Marek J. Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice. *Journal of Applied Microbiology*, 129(4):806–822, 2020.

[419] Barlam T. F., Cosgrove S. E., Abbo L. M., et al. Implementing an antibiotic stewardship program: guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. *Clinical Infectious Diseases*, 62(10):e51–e77, 2016.

[420] Arena F., Giani T., Pollini S., et al. Molecular antibiogram in diagnostic clinical microbiology: advantages and challenges, 2017.

[421] Feucherolles M., Nennig M., Becker S. L., et al. Combination of MALDI-TOF mass spectrometry and machine learning for rapid antimicrobial resistance screening: The case of campylobacter spp. *Frontiers in Microbiology*, 12:804484, 2022.

[422] Bookstaver P., Nimmich E., Smith III T., et al. Cumulative effect of an antimicrobial stewardship and rapid diagnostic testing bundle on early streamlining of antimicrobial therapy in gram-negative bloodstream infections. *Antimicrobial Agents and Chemotherapy*, 61(9):e00189–17, 2017.

[423] Mangioni D., Viaggi B., Giani T., et al. Diagnostic stewardship for sepsis: the need for risk stratification to triage patients for fast microbiology workflows, 2019.

[424] Han S.-S., Jeong Y.-S., and Choi S.-K. Current scenario and challenges in the direct identification of microorganisms using MALDI TOF MS. *Microorganisms*, 9(9):1917, 2021.

[425] De Carolis E., Vella A., Vaccaro L., et al. Application of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *The Journal of Infection in Developing Countries*, 8(09):1081–1088, 2014.

[426] Weis C. V., Jutzeler C. R., and Borgwardt K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clinical Microbiology and Infection*, 26(10):1310–1317, 2020.

[427] Kim J. I., Maguire F., Tsang K. K., et al. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clinical Microbiology Reviews*, 35(3):e00179–21, 2022.

[428] Goodswen S. J., Barratt J. L., Kennedy P. J., et al. Machine learning and applications in microbiology. *FEMS Microbiology Reviews*, 45(5):fuab015, 2021.

[429] Jia B., Raphenya A. R., Alcock B., et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, page gkw1004, 2016.

[430] Feldgarden M., Brover V., Haft D. H., et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial Agents and Chemotherapy*, 63(11):10–1128, 2019.

[431] Yin X., Jiang X.-T., Chai B., et al. ARGs-OAP v2.0 with an expanded SARG database and hidden markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, 34(13):2263–2270, 2018.

[432] Altschul S. F., Gish W., Miller W., et al. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[433] Arango-Argoty G., Garner E., Pruden A., et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):1–15, 2018.

[434] Li Y., Xu Z., Han W., et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome*, 9(1):1–12, 2021.

[435] Sogawa K., Watanabe M., IshigE T., et al. Rapid discrimination between methicillin-sensitive and methicillin-resistant Staphylococcus aureus using MALDI-TOF mass spectrometry. *Biocontrol Science*, 22(3):163–169, 2017.

[436] Wang H.-Y., Chen C.-H., Lee T.-Y., et al. Rapid detection of heterogeneous vancomycin-intermediate Staphylococcus aureus based on matrix-assisted laser desorption ionization time-of-flight: using a machine learning approach and unbiased validation. *Frontiers in Microbiology*, 9:2393, 2018.

[437] Tang W., Ranganathan N., Shahrezaei V., et al. MALDI-TOF mass spectrometry on intact bacteria combined with a refined analysis framework allows accurate classification of MSSA and MRSA. *PLOS ONE*, 14(6):e0218951, 2019.

[438] Sabença C., de Sousa T., Oliveira S., et al. Next-generation sequencing and MALDI mass spectrometry in the study of multiresistant processed meat vancomycin-resistant enterococci (VRE). *Biology*, 9(5):89, 2020.

[439] Sousa T. d., Viala D., Théron L., et al. Putative protein biomarkers of Escherichia coli antibiotic multiresistance identified by MALDI mass spectrometry. *Biology*, 9(3):56, 2020.

[440] Weis C., Cuénod A., Rieck B., et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nature Medicine*, 28(1):164–174, 2022.

[441] Gönen M. and Margolin A. A. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics*, 30(17):i556–i563, 2014.

[442] Costello J. C., Heiser L. M., Georgii E., et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, 2014.

[443] He X., Folkman L., and Borgwardt K. Kernelized rank learning for personalized drug recommendation. *Bioinformatics*, 34(16):2808–2816, 2018.

[444] Corbin C. K., Sung L., Chattopadhyay A., et al. Personalized antibiograms for machine learning driven antibiotic selection. *Communications Medicine*, 2(1):1–14, 2022.

[445] Chiu Y.-C., Chen H.-I. H., Zhang T., et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, 12(1):143–155, 2019.

[446] Baptista D., Ferreira P. G., and Rocha M. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 22(1):360–379, 2021.

[447] Weis C., Cuénod A., Rieck B., et al. DRIAMS: Database of resistance information on antimicrobials and MALDI-TOF mass spectra, 2021.

[448] Willett P., Barnard J. M., and Downs G. M. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

[449] Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences*, 41(2):233–245, 2001.

[450] David L., Thakkar A., Mercado R., et al. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):1–22, 2020.

[451] Durant J. L., Leland B. A., Henry D. R., et al. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.

[452] Wang Y., Bryant S. H., Cheng T., et al. Pubchem bioassay: 2017 update. *Nucleic Acids Research*, 45(D1):D955–D963, 2017.

[453] Morgan H. L. The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

[454] Landrum G., Tosco P., Kelley B., et al. rdkit/rdkit: 2021_09_4 (Q3 2021) release, 2022.

[455] Swain M. PubChemPy documentation, 2014.

[456] Chicco D. Siamese neural networks: An overview. *Artificial Neural Networks*, pages 73–94, 2021.

[457] Szegedy C., Ioffe S., Vanhoucke V., et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[458] Lundberg S. M. and Lee S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[459] Centers for Disease Control and Prevention (U.S.). Antibiotic resistance threats in the united states, 2019. Technical report, National Center for Emerging Zoonotic and Infectious Diseases (U.S.), Nov. 2019.

[460] Yelin I., Snitser O., Novich G., et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature Medicine*, 25(7):1143–1152, 2019.

[461] Leclercq R. Mechanisms of resistance to macrolides and lincosamides: Nature of the resistance elements and their clinical implications. *Clinical Infectious Diseases*, 34(4):482–492, Feb. 2002.

[462] Garneau-Tsodikova S. and Labby K. J. Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. *MedChemComm*, 7(1):11–27, 2016.

[463] Worthington R. J. and Melander C. Overcoming resistance to β-lactam antibiotics. *The Journal of Organic Chemistry*, 78(9):4207–4213, Mar. 2013.

[464] Ren Y., Chakraborty T., Doijad S., et al. Multi-label classification for multi-drug resistance prediction of Escherichia coli. *Computational and Structural Biotechnology Journal*, 20:1264–1270, 2022.

[465] Yoon E.-J. and Jeong S. H. MALDI-TOF mass spectrometry technology as a tool for the rapid diagnosis of antimicrobial resistance in bacteria. *Antibiotics*, 10(8):982, 2021.

[466] Tsoumakas G., Katakis I., and Vlahavas I. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010.

[467] Rokach L., Schclar A., and Itach E. Ensemble methods for multi-label classification. *Expert Systems with Applications*, 41(16):7507–7523, 2014.

[468] Bongini P., Pancino N., Scarselli F., et al. Biognn: How graph neural networks can solve biological problems. In *Artificial Intelligence and Machine Learning for Healthcare*, pages 211–231. Springer, 2023.

[469] Lee M. and Min K. MGCVAE: Multi-objective inverse design via molecular graph conditional variational autoencoder. *Journal of Chemical Information and Modeling*, 62(12):2943–2950, June 2022.

[470] Price W. N. and Cohen I. G. Privacy in the age of medical big data. *Nature Medicine*, 25(1):37–43, 2019.

[471] Price W. N., Gerke S., and Cohen I. G. Potential liability for physicians using artificial intelligence. *Jama*, 322(18):1765–1766, 2019.

[472] Muehlematter U. J., Daniore P., and Vokinger K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 3(3):e195–e203, 2021.

[473] Chen R. J., Wang J. J., Williamson D. F., et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023.

[474] Reddy S., Allan S., Coghlan S., et al. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3):491–497, 2020.

[475] Bradbury J. Human epigenome project—up and running. *PLOS Biology*, 1(3):e82, 2003.

[476] Ferreira L., Sánchez-Juanes F., Porras-Guerra I., et al. Microorganisms direct identification from blood culture by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clinical Microbiology and Infection*, 17(4):546–551, 2011.

[477] Croxatto A., Prod'hom G., and Greub G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews*, 36(2):380–407, 2012.

[478] Rychert J. Benefits and limitations of MALDI-TOF mass spectrometry for the identification of microorganisms. *Journal of Infectiology and Epidemiology*, 2(4), 2019.

[479] Hrabák J., Chudáčková E., and Walková R. Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry for detection of antibiotic resistance mechanisms: from research to routine diagnosis. *Clinical Microbiology Reviews*, 26(1):103–114, 2013.

[480] Blair J. M. A., Webber M. A., Baylay A. J., et al. Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, 13:42–51, 2014.

[481] Florio W., Baldeschi L., Rizzato C., et al. Detection of antibiotic-resistance by MALDI-TOF mass spectrometry: An expanding area. *Frontiers in Cellular and Infection Microbiology*, 10, 2020.

[482] Maugeri G., Lychko I., Sobral R., et al. Identification and antibiotic-susceptibility profiling of infectious bacterial agents: a review of current and future trends. *Biotechnology Journal*, 14(1):1700750, 2019.

[483] Buchanan R. and Wareham D. Mechanisms of antibiotic resistance. *Tutorial Topics in Infection for the Combined Infection Training Programme*, 2019.

[484] Schreiber J., Bilmes J., and Noble W. S. Prioritizing transcriptomic and epigenomic experiments using an optimization strategy that leverages imputed data. *Bioinformatics*, 37(4):439–447, 2021.

[485] Chitpin J. G., Awdeh A., and Perkins T. J. RECAP reveals the true statistical significance of ChIP-seq peak calls. *Bioinformatics*, 35(19):3592–3598, 2019.

[486] Tilgner H., Nikolaou C., Althammer S., et al. Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9):996, 2009.

[487] Karlić R., Chung H.-R., Lasserre J., et al. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010.

[488] Kundaje A., Kyriazopoulou-Panagiotopoulou S., Libbrecht M., et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747, 2012.

[489] Quinlan A. R. and Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[490] Robinson M. D. and Smyth G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[491] Love M. I., Huber W., and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

[492] Robinson M. D., McCarthy D. J., and Smyth G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[493] Amemiya H. M., Kundaje A., and Boyle A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Scientific Reports*, 9(1):1–5, 2019.

[494] McInnes L., Healy J., Saul N., et al. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, Sept. 2018.

[495] Karmodiya K., Krebs A. R., Oulad-Abdelghani M., et al. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1):1–18, 2012.

[496] Calo E. and Wysocka J. Modification of enhancer chromatin: what, how, and why? *Molecular Cell*, 49(5):825–837, 2013.

[497] Zhang T., Zhang Z., Dong Q., et al. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21(1):1–7, 2020.

[498] Song L. and Crawford G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384, 2010.

[499] Barski A., Cuddapah S., Cui K., et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[500] Ferrari K. J., Scelfo A., Jammula S., et al. Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Molecular Cell*, 53(1):49–62, 2014.

[501] Schwartz S., Meshorer E., and Ast G. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, 16(9):990, 2009.

[502] Li T., Liu Q., Garza N., et al. Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing. *Genome Medicine*, 10(1):1–11, 2018.

[503] Farooq Z., Banday S., Pandita T. K., et al. The many faces of histone H3K79 methylation. *Mutation Research/Reviews in Mutation Research*, 768:46–52, 2016.

[504] Vig J., Madani A., Varshney L. R., et al. {BERT}ology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021.

[505] Raganato A. and Tiedemann J. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics, 2018.

[506] Kovaleva O., Romanov A., Rogers A., et al. Revealing the dark secrets of BERT. In Inui K., Jiang J., Ng V., et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4364–4373. Association for Computational Linguistics, 2019.

[507] Lee J., Shin J.-H., and Kim J.-S. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, 2017.

[508] Serrano S. and Smith N. A. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

[509] Jain S. and Wallace B. C. Attention is not explanation. In Burstein J., Doran C., and Solorio T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, 2019.

[510] Howe K. L., Achuthan P., Allen J., et al. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 11 2020.

[511] Shilatifard A. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annu. Rev. Biochem.*, 75:243–269, 2006.

[512] Boyle A. P., Davis S., Shulha H. P., et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.

[513] Robertson A. G., Bilenky M., Tam A., et al. Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding. *Genome Research*, 18(12):1906–1917, 2008.

[514] Mohn F. and Schübeler D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends in Genetics*, 25(3):129–136, 2009.

[515] Voita E., Talbot D., Moiseev F., et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.

Appendix A

Supplementary material for Chapter 4

A.1 Supplementary figures

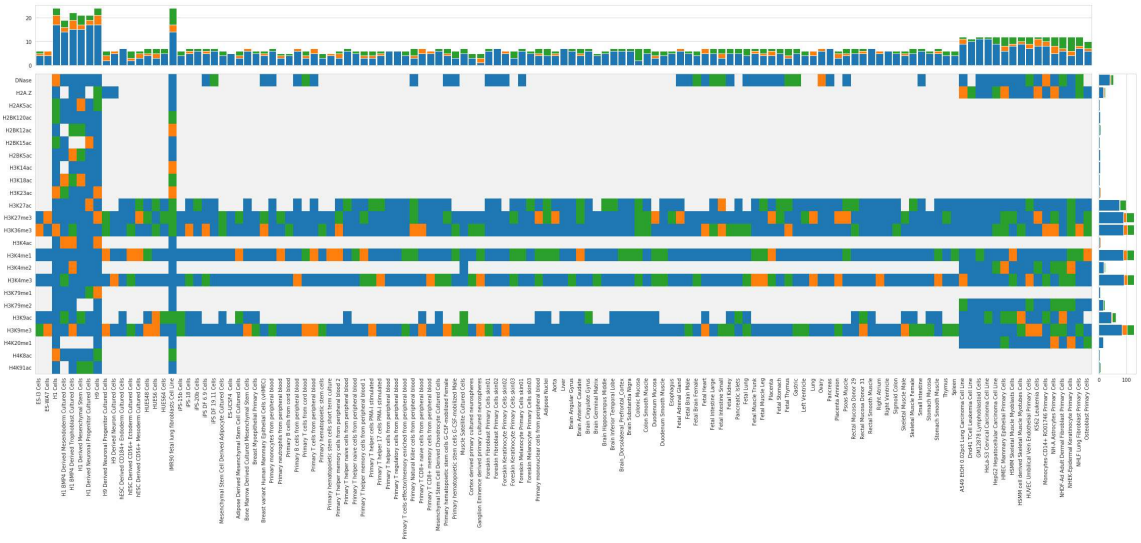


Fig. A.1 **Roadmap data split.** The data matrix shows which tracks from the Roadmap consortium were part of the training (blue), validation (orange) and test (green) sets.

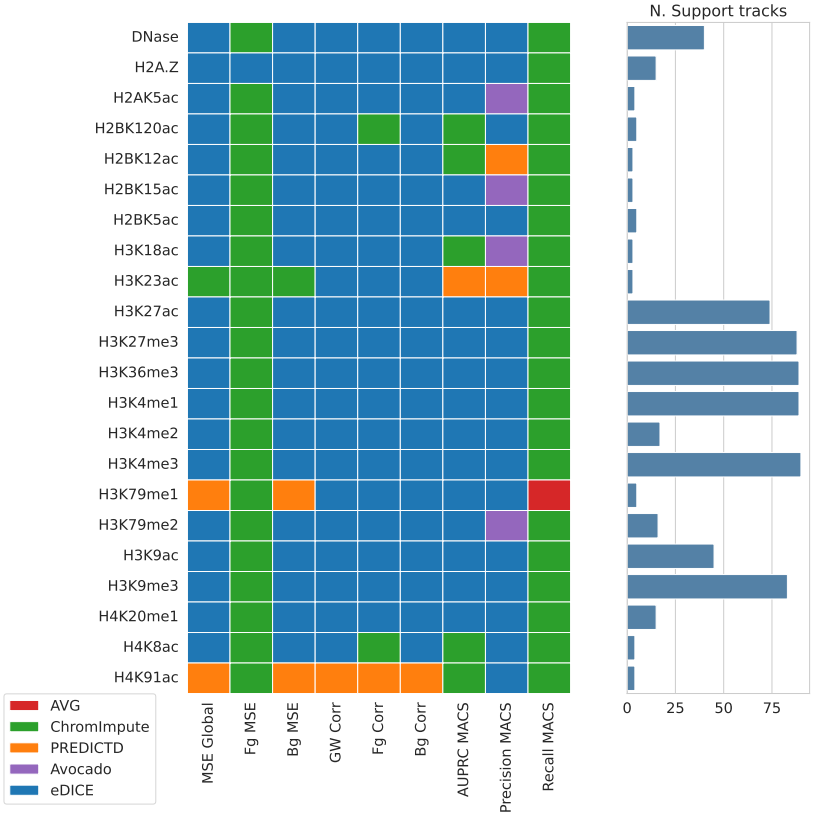


Fig. A.2 **Best model for each assay.** The categorical heatmap showcases which model presents the best average test performance for each assay, divided by metric.

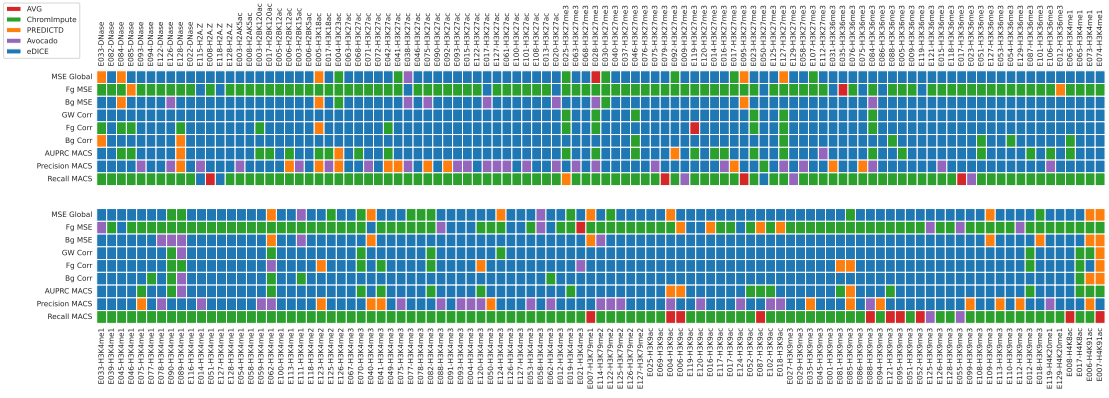


Fig. A.3 **Best model for each test track.** This heatmap displays the model that achieves the best performance for each assay-cell type combination in the test set, split by metric.

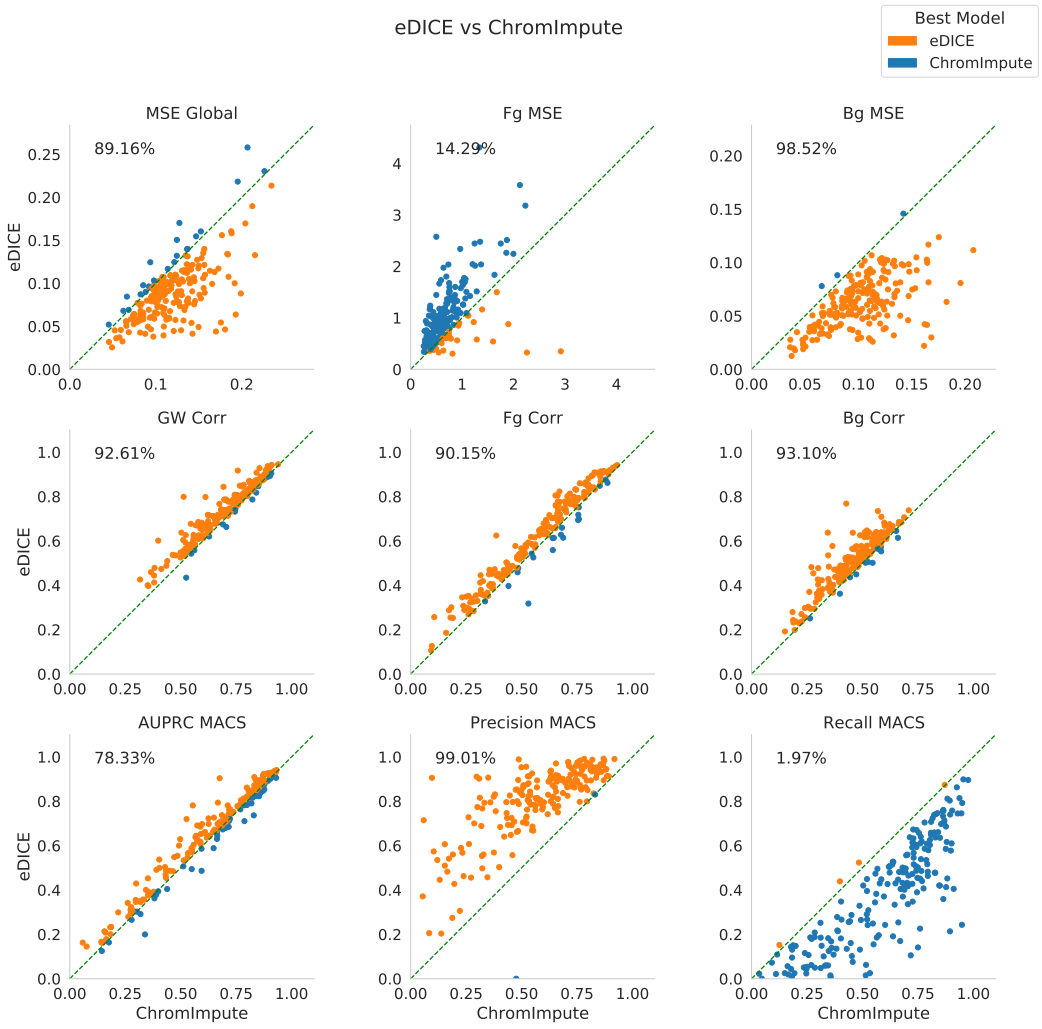


Fig. A.4 **eDICE vs ChromImpute - individual test tracks.** Scatter plots for the comparison of the performance of eDICE against ChromImpute for each individual test track on a subset of the evaluation metrics metric. Each point corresponds to an individual test track, whose coordinates are given by the metric of the plot for the baseline and eDICE. The percentage in the top-left corner of each plot indicates the percentage of tracks on which eDICE outperforms the baseline.

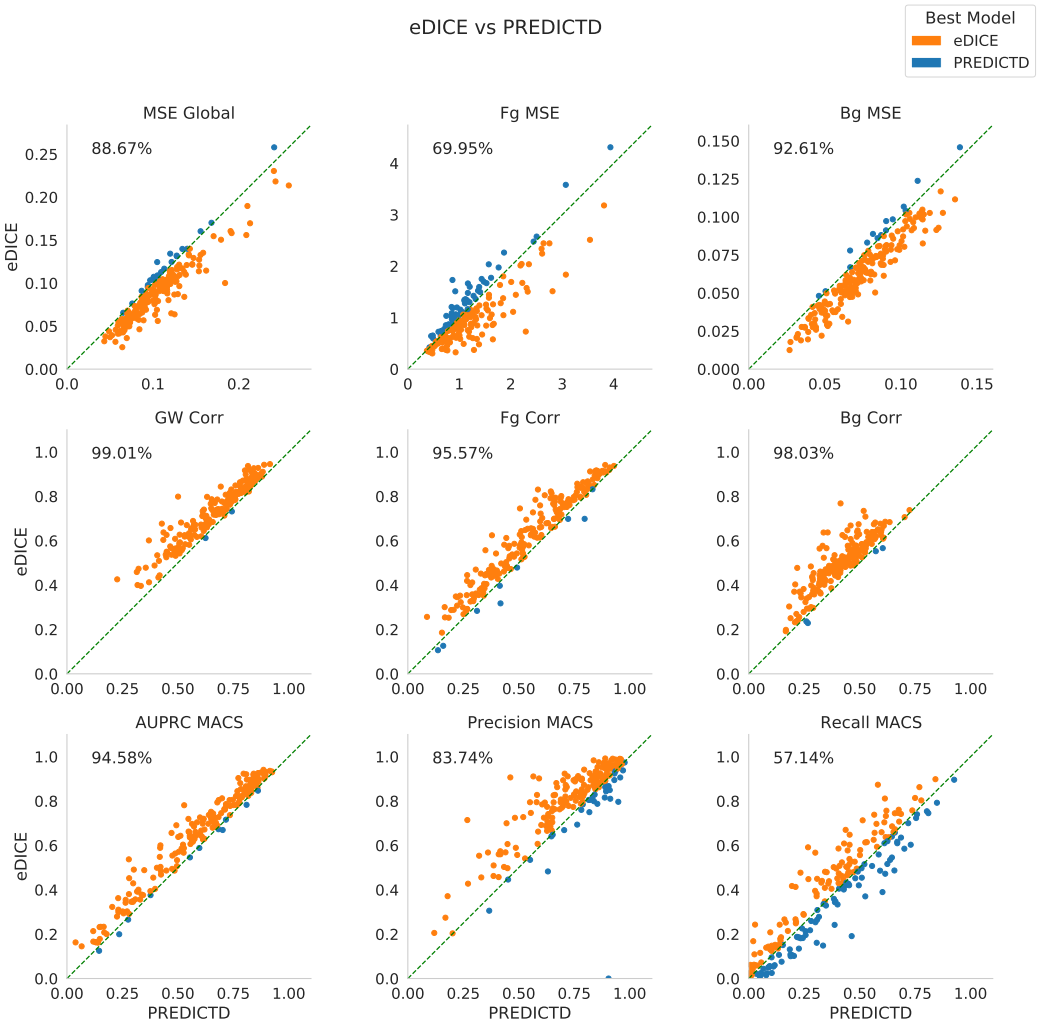


Fig. A.5 **eDICE vs PREDICTD - individual test tracks.** Scatter plots for the comparison of the performance of eDICE against PREDICTD for each individual test track on a subset of the evaluation metrics metric.

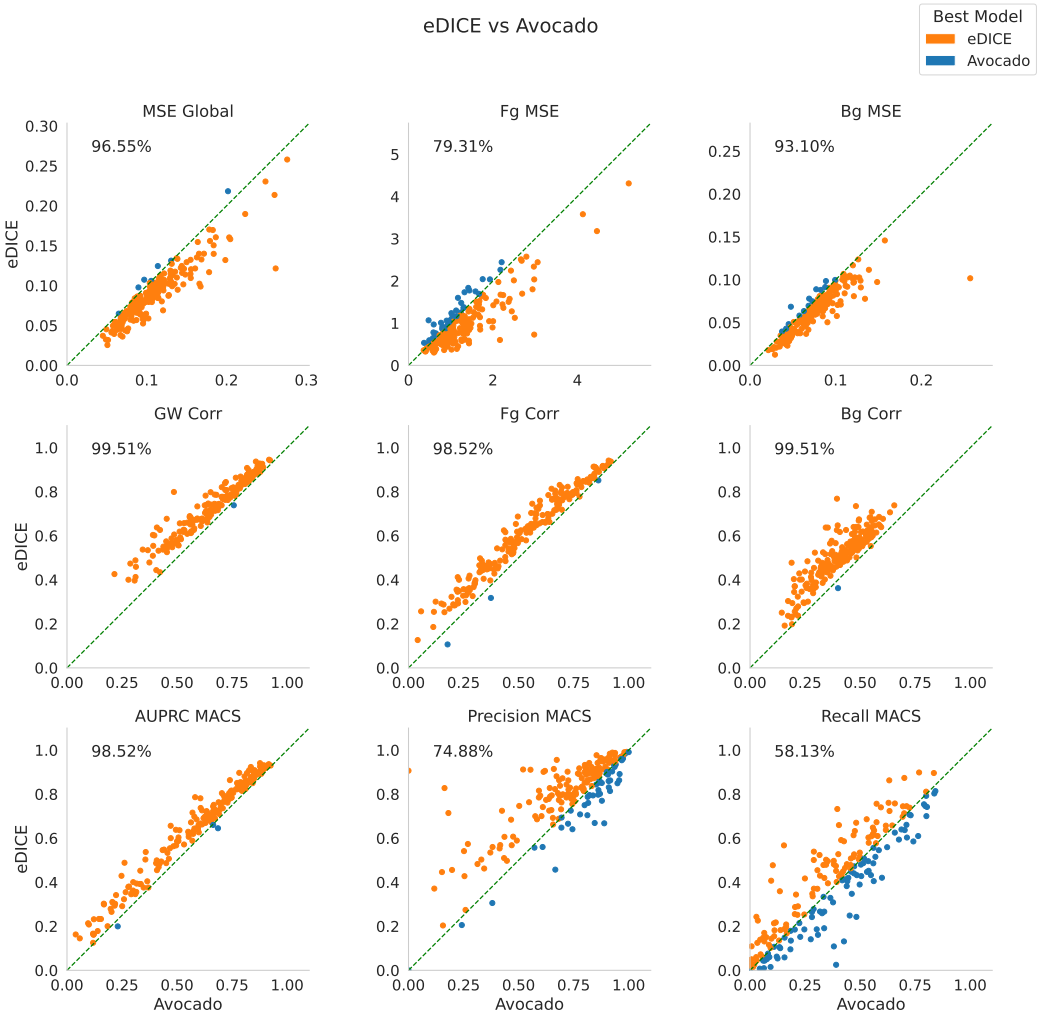


Fig. A.6 **eDICE vs Avocado - individual test tracks.** Scatter plots for the comparison of the performance of eDICE against Avocado for each individual test track on a subset of the evaluation metrics metric.

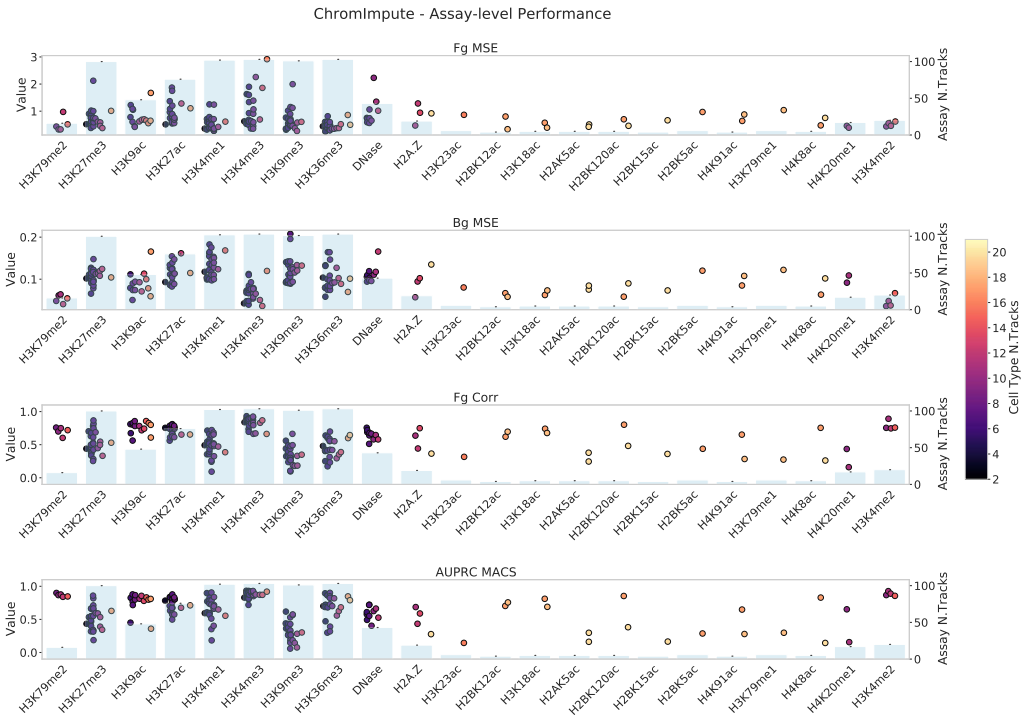


Fig. A.7 **ChromImpute performance on the Roadmap test set.** Track level performance of the imputations divided by assay for ChromImpute on the 203 test tracks from the Roadmap dataset.

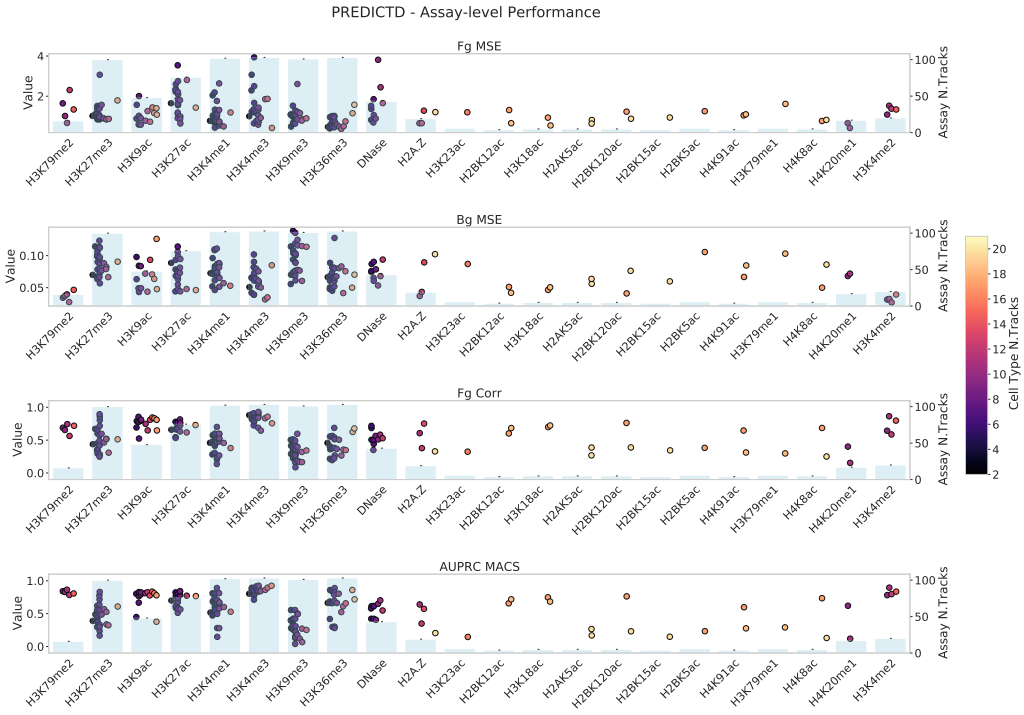


Fig. A.8 **PREDICTD performance on the Roadmap test set.** Track level performance of the imputations divided by assay for PREDICTD on the 203 test tracks from the Roadmap dataset.

Fig. A.9 **Avocado performance on the Roadmap test set.** Track level performance of the imputations divided by assay for Avocado on the 203 test tracks from the Roadmap dataset.

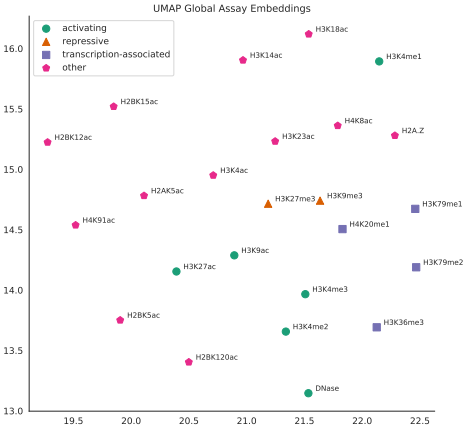


Fig. A.10 **Global assay embeddings.** UMAP projection of assay global embeddings

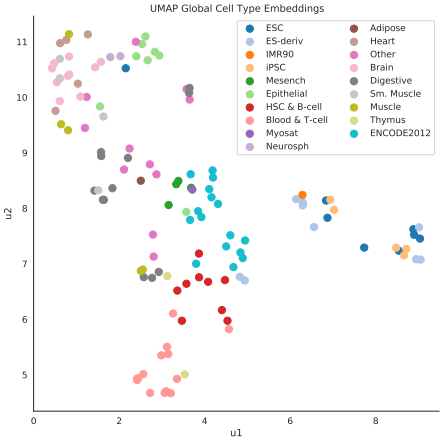


Fig. A.11 **Global cell type embeddings.** UMAP projection of cell type global embeddings

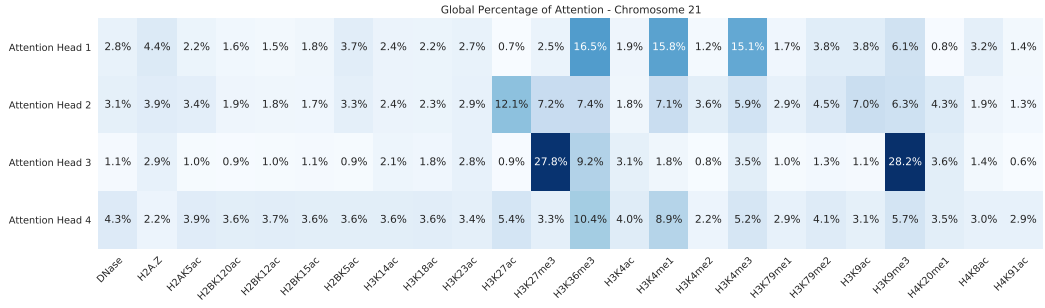


Fig. A.12 **The attention heads focus on a few well mapped modifications.** Percentage of attention that the 4 attention heads dedicate to each assay over the entire chromosome 21.

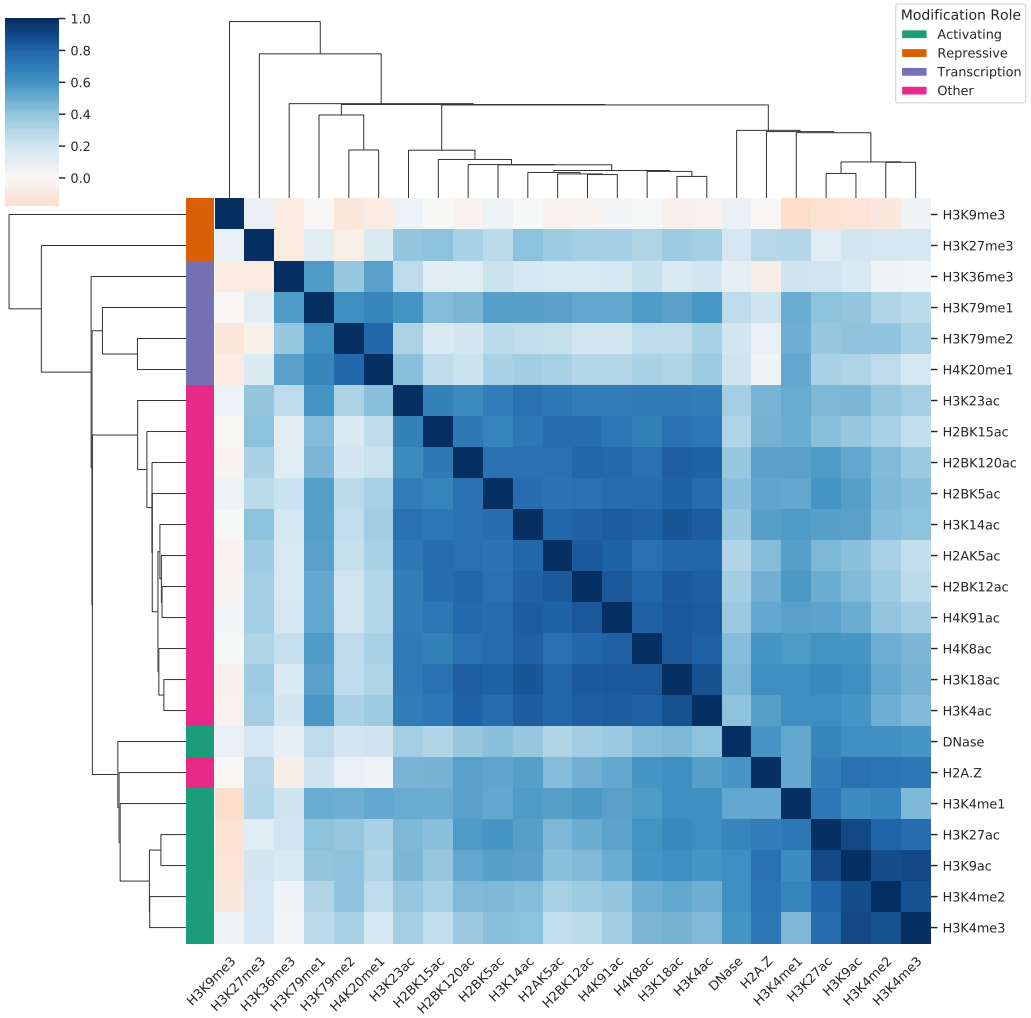


Fig. A.13 **Histone modifications with shared functions show high correlation.** Hierarchical clustering of the correlation matrix between epigenetic marks, calculated using the average tracks over all cell types.

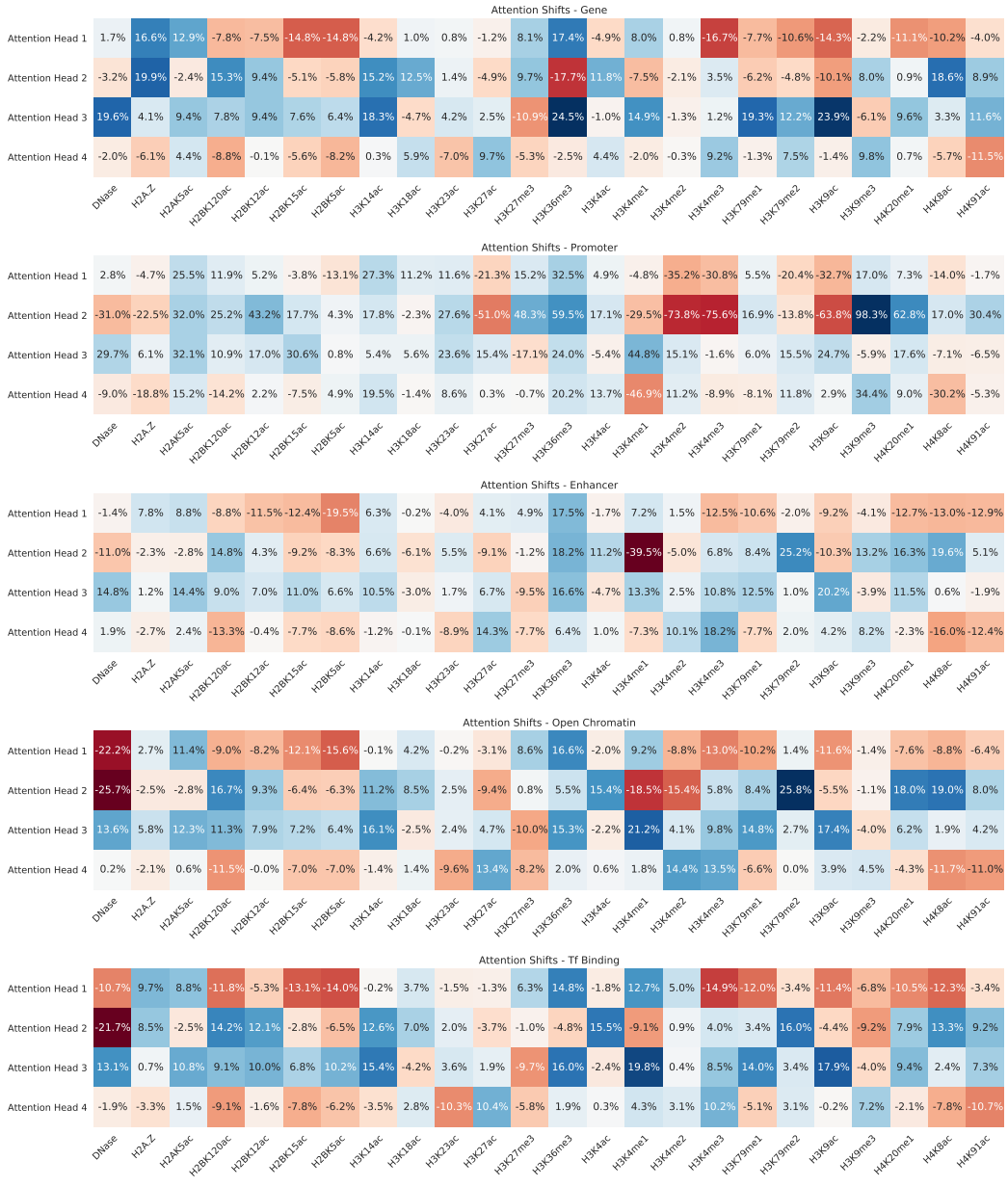


Fig. A.14 **The attention weights shift with different patterns in functional regions of the genome.** Differential percentage of attention that the 4 attention heads dedicate to each assay over the annotated portions of chromosome 21.

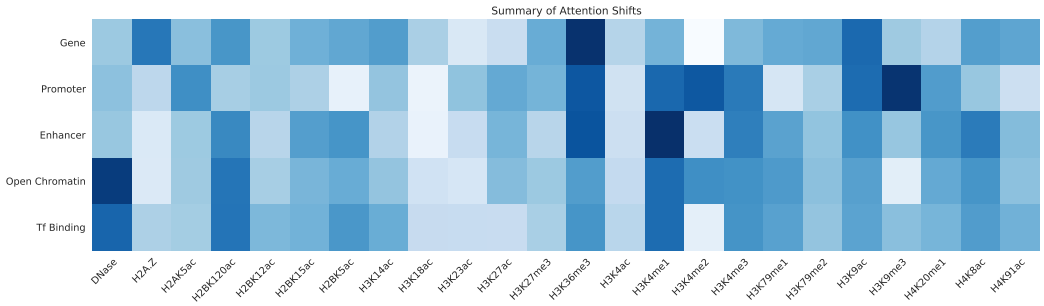


Fig. A.15 **Summary of the attention shifts in the functional regions of the genome.** Visual summary of the attention shifts for each assay in annotated regions of the genome.

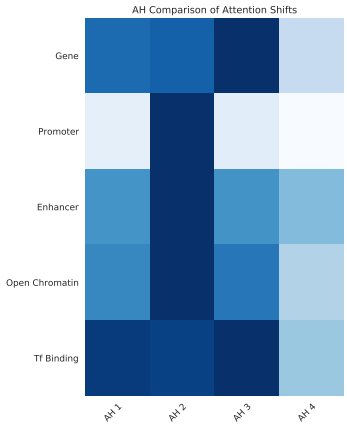


Fig. A.16 **Summarized attention shifts for each transformer head.** Visual summary of the attention shifts for each attention head in annotated regions of the genome.

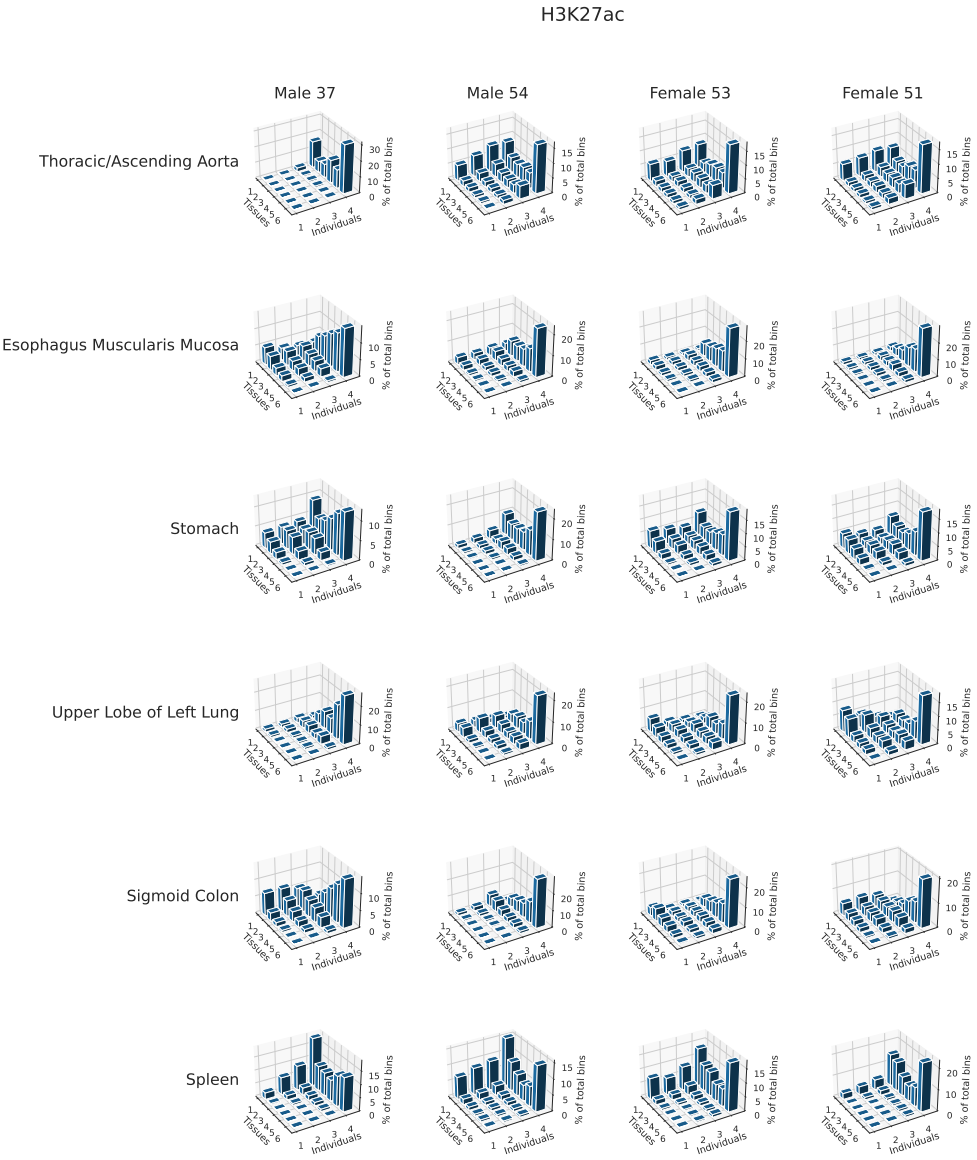


Fig. A.17 **3D occupancy histogram - H3K27ac.** 3D Histograms representing between how many tissues and individuals each enriched bin is shared for H3K27ac tracks in the EN-TEx dataset.

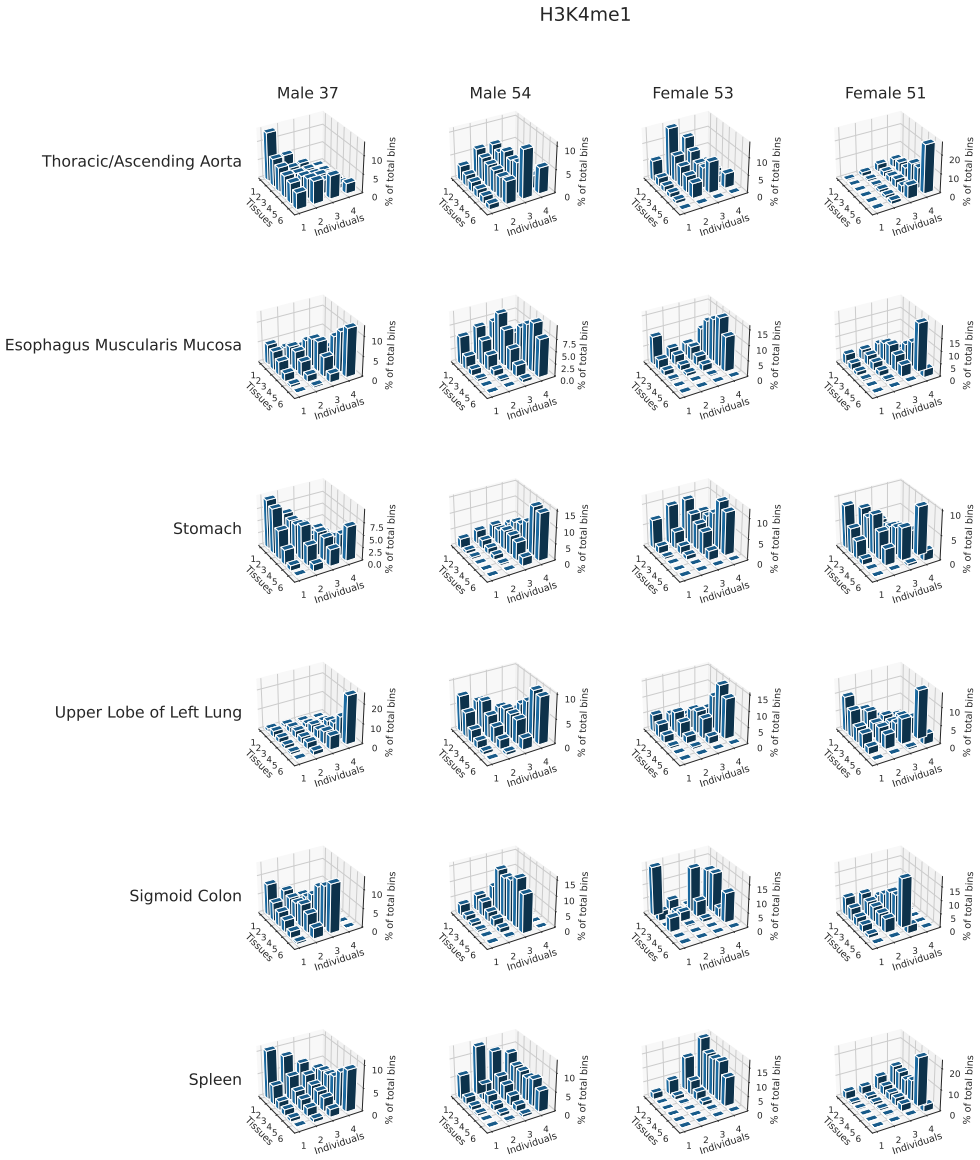


Fig. A.18 **3D occupancy histogram - H3K4me1.** 3D Histograms representing between how many tissues and individuals each enriched bin is shared for H3K4me1 tracks in the EN-TEx dataset.

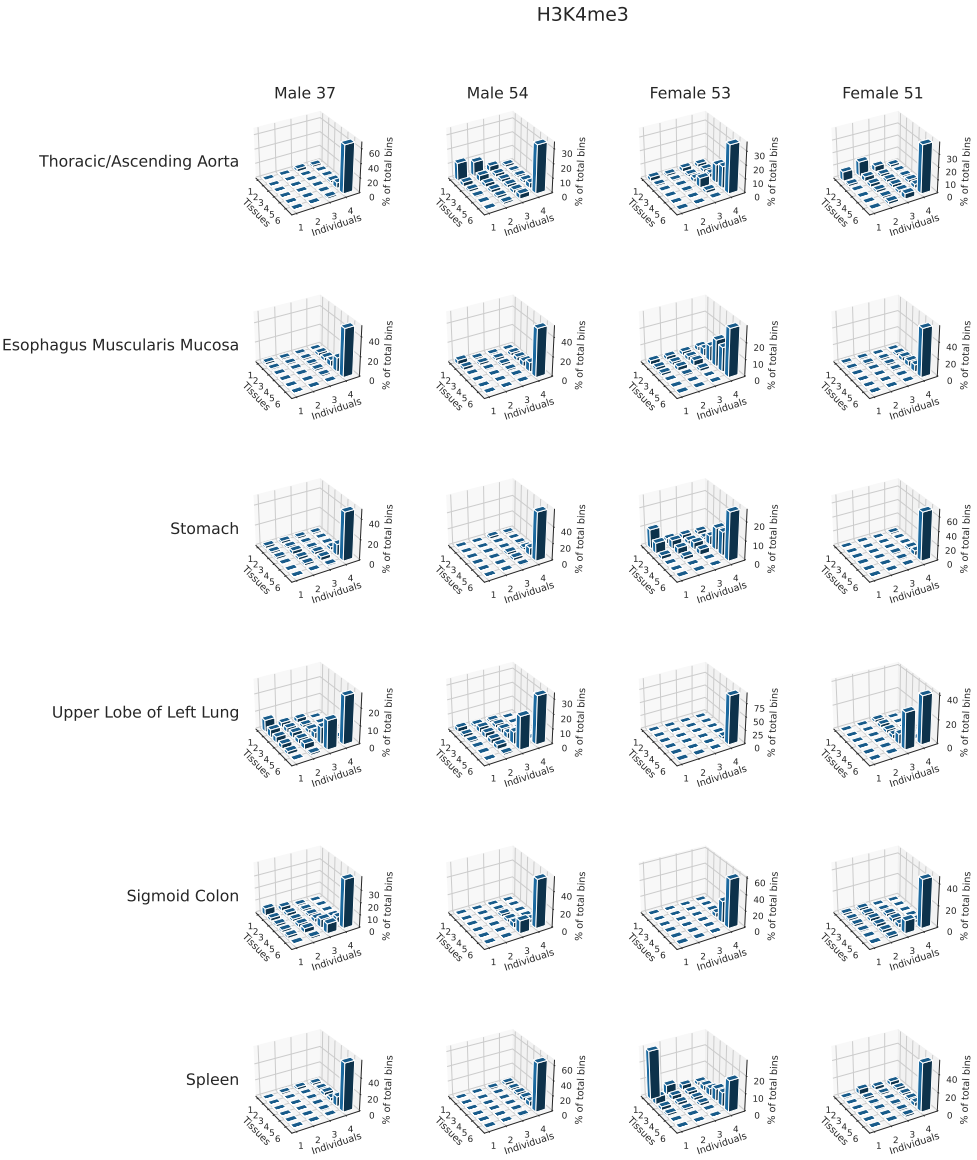


Fig. A.19 **3D occupancy histogram - H3K4me3.** 3D Histograms representing between how many tissues and individuals each enriched bin is shared for H3K4me3 tracks in the EN-TEx dataset.

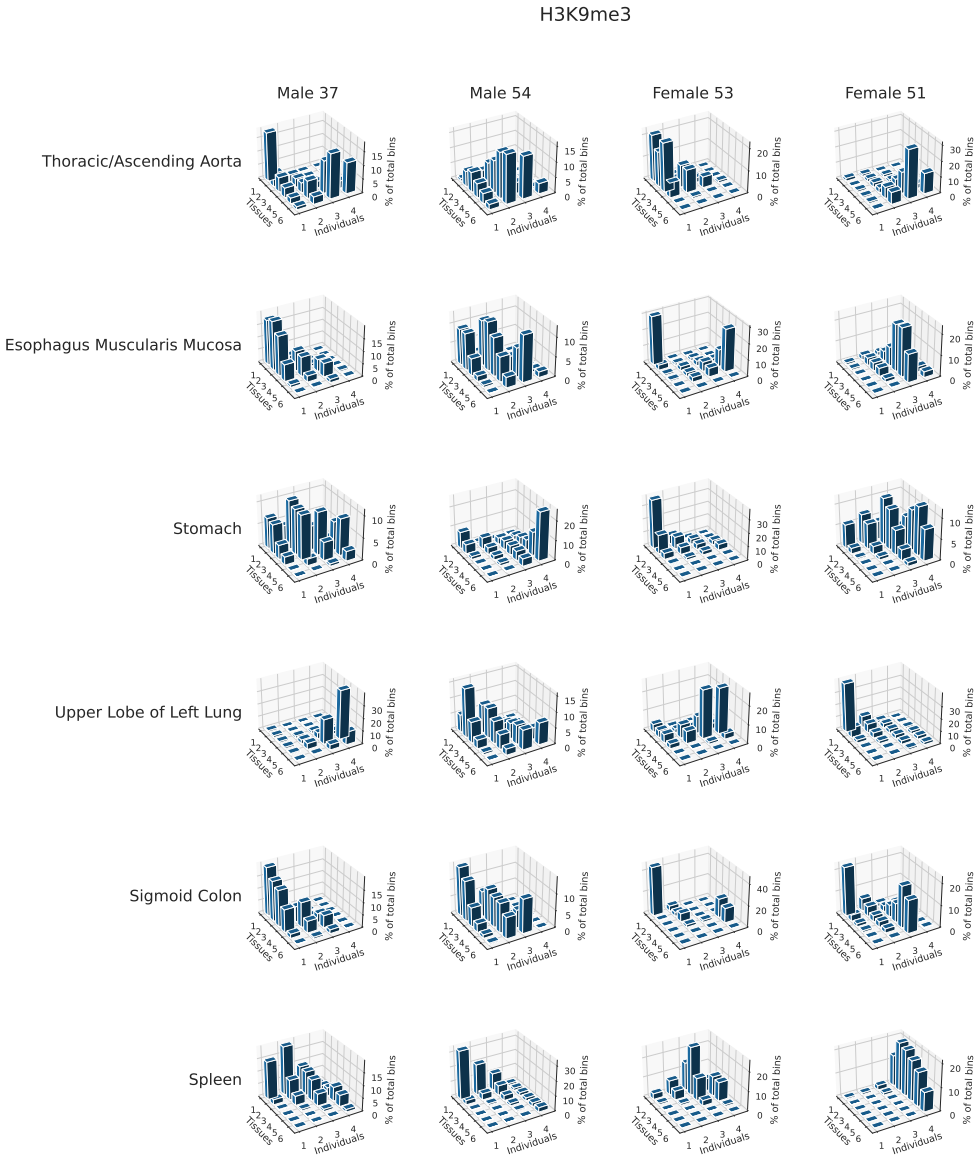


Fig. A.20 **3D occupancy histogram - H3K9me3.** 3D Histograms representing between how many tissues and individuals each enriched bin is shared for H3K9me3 tracks in the EN-TEx dataset.

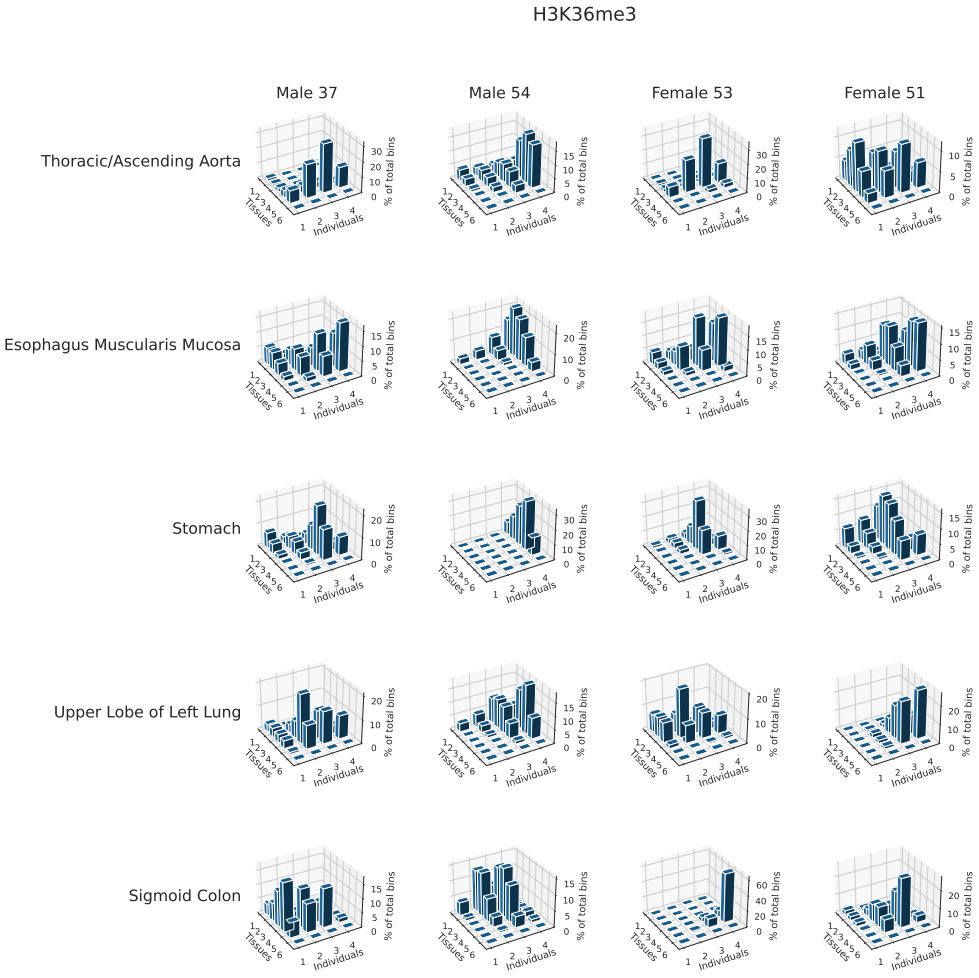


Fig. A.21 **3D occupancy histogram - H3K36me3.** 3D Histograms representing between how many tissues and individuals each enriched bin is shared for H3K36me3 tracks in the EN-TEx dataset.

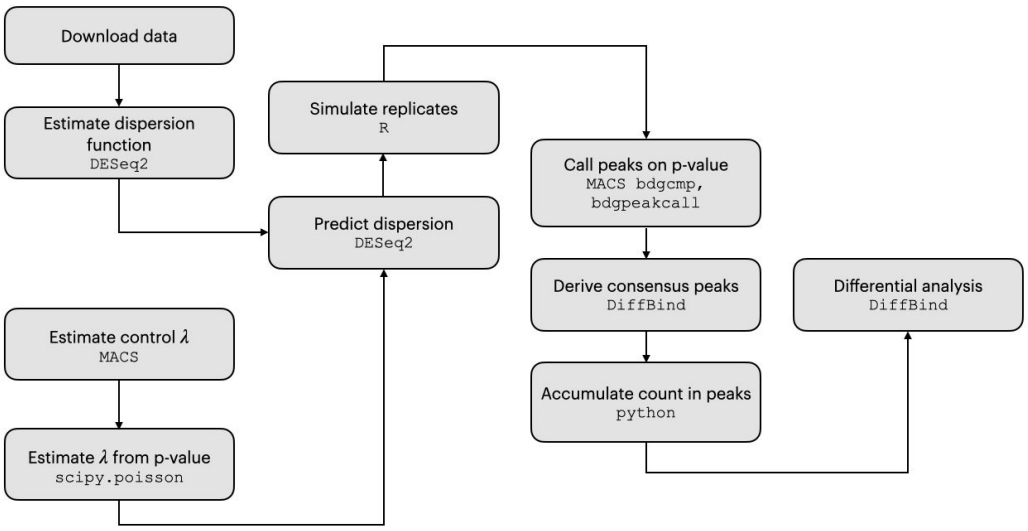


Fig. A.22 **Differential analysis pipeline.** Schematic of the steps performed to compare peaks between two different cell types.

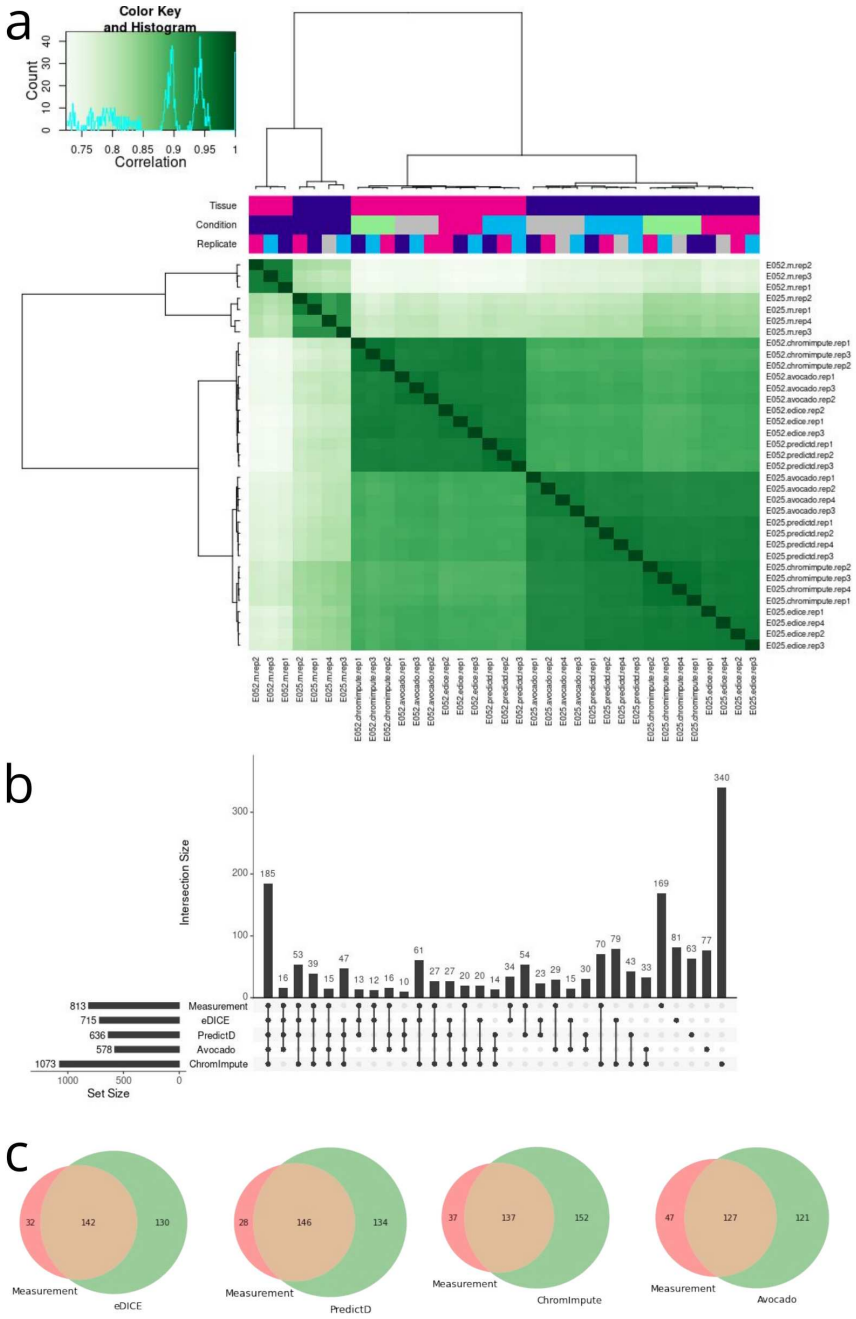


Fig. A.23 **Differential peak analysis pipeline comparing different imputation methods.** Only measurement samples were used to derive the consensus peakset: **(a)** Correlation heatmap of the affinity scores for different methods and replicates **(b)** Upset plot showing the size of set intersections of differentially enriched peaksets among imputation methods **(c)** Venn diagram showing peaks that are detected as differentially enriched between tissues using imputed and measured signals.

A.2 Supplementary tables

Table A.1 **Number of models and parameters per model required to make genome-wide predictions on the selected Roadmap test set.** While PREDICTD and Avocado require several billion parameters for genome-wide prediction, eDICE requires only 6 million to obtain competitive performance.

Model	# models	# parameters per model
ChromImpute	203	/
PREDICTD	8	$\sim 11.5\,B$
Avocado	1	$\sim 3.4\,B$
eDICE	1	$\sim$ **6 M**

Table A.2 **Performance metrics for the imputation of the 203 test tracks on chromosome 21.** In bold, we highlighted the best performance for each metric. For each performance metric, we checked whether any model other than the best one failed to reject the null hypothesis of a shared mean in a two-sided paired t-test with a significance level of 0.01 (N=203). No comparable performance was found.

	Genome-Wide Reconstruction		**MACS vs Imp. Classification**	
	GW Corr	MSE Global	AUPRC MACS	AUROC MACS
AVG	0.563 ± 0.02	0.159 ± 0.008	0.45 ± 0.029	0.931 ± 0.006
Avocado	0.668 ± 0.022	0.108 ± 0.005	0.593 ± 0.033	0.956 ± 0.005
ChromImpute	0.697 ± 0.019	0.119 ± 0.005	0.625 ± 0.032	0.967 ± 0.004
PREDICTD	0.669 ± 0.02	0.106 ± 0.005	0.592 ± 0.032	0.96 ± 0.005
eDICE	$\mathbf{0.735 \pm 0.018}$	$\mathbf{0.091 \pm 0.005}$	$\mathbf{0.651 \pm 0.031}$	$\mathbf{0.969 \pm 0.004}$

	Background Reconstruction		**Foreground Reconstruction**	
	Bg Corr	Bg MSE	Fg Corr	Fg MSE
AVG	0.311 ± 0.015	0.124 ± 0.005	0.467 ± 0.028	1.328 ± 0.097
Avocado	0.401 ± 0.016	0.073 ± 0.004	0.537 ± 0.03	1.283 ± 0.089
ChromImpute	0.455 ± 0.016	0.102 ± 0.004	0.571 ± 0.028	$\mathbf{0.73 \pm 0.057}$
PREDICTD	0.418 ± 0.016	0.074 ± 0.003	0.548 ± 0.029	1.189 ± 0.087
eDICE	$\mathbf{0.503 \pm 0.016}$	$\mathbf{0.063 \pm 0.003}$	$\mathbf{0.614 \pm 0.028}$	1.041 ± 0.082

	Obs. MACS vs Imp. MACS Classification			
	F1 MACS	MCC MACS	Precision MACS	Recall MACS
AVG	0.41 ± 0.024	0.415 ± 0.023	0.451 ± 0.027	0.451 ± 0.03
Avocado	0.45 ± 0.034	0.484 ± 0.031	0.744 ± 0.029	0.373 ± 0.033
ChromImpute	$\mathbf{0.544 \pm 0.027}$	$\mathbf{0.554 \pm 0.025}$	0.574 ± 0.029	$\mathbf{0.603 \pm 0.033}$
PREDICTD	0.453 ± 0.032	0.485 ± 0.029	0.739 ± 0.025	0.371 ± 0.031
eDICE	0.481 ± 0.034	0.52 ± 0.029	$\mathbf{0.803 \pm 0.023}$	0.392 ± 0.033

Tissue	Assay	ENCDO845WKR (male 37 years old)	ENCDO451RUA (male 54 years old)	ENCDO793LXB (female 53 years old)	ENCDO271OUW (female 51 years old)
ascending aorta	H3K27ac			ENCFF472VCY	ENCFF002VCM
	H3K36me3			ENCFF529EEQ	ENCFF602IDJ
	H3K4me1			ENCFF971TZV	ENCFF843GKY
	H3K4me3			ENCFF229NZH	ENCFF383JQQ
	H3K9me3			ENCFF616UAG	ENCFF644UVB
thoracic aorta	H3K27ac	ENCFF652JBU	ENCFF344YDV		
	H3K36me3	ENCFF700ZLW	ENCFF556REM		
	H3K4me1	ENCFF141KEC	ENCFF388BWY		
	H3K4me3	ENCFF327VEH	ENCFF723SJI		
	H3K9me3	ENCFF221YER	ENCFF198OYC		
esophagus muscularis mucosa	H3K27ac	ENCFF969PNR	ENCFF437XOA	ENCFF282WYM	ENCFF707HDC
	H3K36me3	ENCFF476ILL	ENCFF567TNH	ENCFF723KNC	ENCFF861VVB
	H3K4me1	ENCFF295FVH	ENCFF478GEN	ENCFF067DEH	ENCFF877IZN
	H3K4me3	ENCFF280VIW	ENCFF979ERR	ENCFF156BAJ	ENCFF636QVZ
	H3K9me3	ENCFF501FLV	ENCFF380LRK	ENCFF885KSL	ENCFF729RCC
stomach	H3K27ac	ENCFF440VXQ	ENCFF241NFR	ENCFF134NHD	ENCFF140JKH
	H3K36me3	ENCFF302OSD	ENCFF414FQR	ENCFF428LSW	ENCFF441GAP
	H3K4me1	ENCFF140RPG	ENCFF034VHX	ENCFF411ZSD	ENCFF131UFO
	H3K4me3	ENCFF175WCH	ENCFF221JKN	ENCFF309DCR	ENCFF468UJW
	H3K9me3	ENCFF285GPE	ENCFF765FRJ	ENCFF674DMF	ENCFF554RBC
upper lobe of left lung	H3K27ac	ENCFF204PWD	ENCFF305XCG	ENCFF071QUR	ENCFF178DCM
	H3K36me3	ENCFF549FDA	ENCFF572VAG	ENCFF267DPL	ENCFF700BQU
	H3K4me1	ENCFF076BNQ	ENCFF614CUN	ENCFF777AOP	ENCFF344GZI
	H3K4me3	ENCFF898DYH	ENCFF526RLU	ENCFF887IUV	ENCFF578OBF
	H3K9me3	ENCFF672QOH	ENCFF583KBZ	ENCFF076RRH	ENCFF317KCA
sigmoid colon	H3K27ac	ENCFF669TSI	ENCFF507VUO	ENCFF156QWL	ENCFF497GAF
	H3K36me3	ENCFF284VTI	ENCFF312AHC	ENCFF577WWK	ENCFF744WIO
	H3K4me1	ENCFF686SIG	ENCFF427WMO	ENCFF126IMK	ENCFF113YOI
	H3K4me3	ENCFF892GNP	ENCFF445QHF	ENCFF905JEN	ENCFF135XHH
	H3K9me3	ENCFF811EKT	ENCFF925APW	ENCFF782WTP	ENCFF477NHW
spleen	H3K27ac	ENCFF230ZQJ	ENCFF786NLG	ENCFF735GCH	ENCFF702AJP
	H3K4me1	ENCFF873LAX	ENCFF049HBP	ENCFF402YZU	ENCFF599IBO
	H3K4me3	ENCFF422ZXF	ENCFF189KXE	ENCFF656AAS	ENCFF462VQE
	H3K9me3	ENCFF952NOQ	ENCFF418KYC	ENCFF447CEL	ENCFF088KFT

Table A.3 **EN-TEx accession codes.** Accession codes for the 116 selected tracks from the EN-TEx dataset used in the imputation of personalized epigenomes. The measurements for "thoracic aorta" and "ascending aorta" were merged into a single tissue label "aorta".

A.3 Supplementary notes

A.3.1 Evaluation strategy

Measuring the performance of an imputation model is far from trivial. Properly designing an evaluation strategy is crucial to ensure that the results reflect the workings of the models and that the comparison to a selected set of baselines aligns with the use case intended for the model.

Ernst and Kellis [395] validated the performance of ChromImpute with a leave-one-out (LOO) strategy, where all tracks short of the one specific target test track are used during different training instances. This strategy provides a substantial prediction advantage due to the optimal number of training data sets. However, it may not give the best assessment of the method's generalization ability [484]. In contrast, the authors of PREDICTD and Avocado have analysed the imputed values for a hold-out set of test tracks using 5-fold cross-validation. This procedure has also been adopted in a recent ENCODE Imputation Challenge [328], and we will do the same in this work. The split of the data is identical to one of the folds used in [397]. Supplementary Figure A.1 depicts a data matrix showing which tracks from the Roadmap consortium were part of the training, validation and test sets. Quality control of the Roadmap Consortium guarantees a minimum standard for all data sets [389].

In previous work, including the Encode Imputation Challenge, training, validation and test tracks are processed following established computational pipelines[1]. This pipeline includes a pooling step for biological and technical replicates and subsampling to maximum library size. Notably, the MACS2 peak caller is applied to calculate the statistical significance of enrichment at each base pair in the genome. The resulting genome-wide signal tracks containing the statistical significance of enrichment (i.e., the $-log_{10}$ p-values) at each base pair in the genome have been used both as input training data and to validate imputations [397]. Recently, however, it has been demonstrated that MACS2 outputs biased p-values and false discovery rate estimates that can be many orders of magnitude too optimistic [485]. This systematic bias raises several problems; for instance, there is no sound basis for choosing a p-value cut off for reporting results. Additionally, as the introduced bias depends on the ChIP-Seq data from which the p values are computed, there can be "distributional shifts" between training and test sets, as observed in the Encode Imputation Challenge.

[1] https://github.com/ENCODE-DCC/chip-seq-pipeline2

To reduce the impact of possible systematic biases, we employ a varied set of performance metrics, each aiming to capture different facets of the imputation task. For example, comparing models using Area-Under-Curve metrics ensures that choosing a specific p-value cut-off is not the determining factor in ranking the imputation models.

A.3.2 Performance metrics

An appropriate choice of metrics is essential for robust comparisons between imputation models. The most immediate and intuitive approach from a machine learning point of view is to measure the ability of a model to reconstruct genome-wide signals (i.e. corrected $-log_{10}$ p-values). Commonly used measures include genome-wide mean-squared-error (GW MSE) and genome-wide Pearson correlation (GW Corr) [397]. Importantly, however, most histone modifications are only significantly enriched on a small subset of genomic bins. The remaining bins have low background signals, which are strongly determined by experimental conditions, including unspecific binding of the antibody. As these background regions are over-represented on the genome, they easily dominate genome-wide assessment measures.

Durham et al. [396] therefore defined additional performance measures, including MSE1obs, which measures the MSE on the top 1% of the positions according to the ChIP-seq signal; MSE1imp measures the MSE on the top 1% of the positions according to the imputed signal. On the other hand, outside the imputation literature, it is a common strategy to analyse histone modification measurements by focusing on foreground regions which are identified using peak callers such as MACS2 [399]. In this case, significant enriched regions are merged if they are close together and fluctuations below a given threshold are tolerated within a foreground peak if they are small enough. Additionally, regions of significant enrichment have to have a minimum length (at least as big as the fragment length) to be called foreground peaks. Peak callers therefore provide a more robust segmentation of the genome into foreground and background regions. Here, we follow a hierarchical analysis to evaluate the imputations. First, we test the ability of the model to correctly classify genomic regions into foreground and background regions (e.g. [486]). The signal reconstruction potential of the models is then separately analysed for bins categorized as foreground and background. For the foreground regions, we are additionally interested in how well the peak intensities are predicted (e.g., [487]), and whether the shape of the signal in the peak is well captured by the prediction. (e.g., [488, 401]).

A.3.3 Validation on the Roadmap reference epigenome

Table A.2 presents the numerical results for the imputation of the test tracks of the Roadmap dataset. For each metric, we highlighted the best-performing model, as well as the competitors for which a 2-sided t-test failed to reject the null hypothesis with a significance threshold of 0.01 (N=203).

Comparing eDICE to the baselines at the assay level (Figure A.2) and at the individual track level (Figures A.3, A.4, and A.5) highlights that the reported improvements are consistent across the board, with the notable exception of four metrics in which ChromImpute outperforms eDICE (Fg MSE, Precision MACS, F1 MACS, and MCC MACS). For all four of them, the underlying issue is that factorization models such as eDICE and PREDICT consistently underestimate the height of peaks, leading to reduced detection of enriched regions. As a consequence, eDICE is more conservative than ChromImpute in its prediction, a fact that is mirrored in the Recall MACS metrics.

Different assays present different overall imputation performance. For example, in the evaluation of eDICE, H3K9me3 and H3K27me3 displayed consistently suboptimal performance. Figures A.7 and A.8 highlight how this tendency is not exclusive to eDICE, but shared by the baseline models.

A.3.4 Differential peak analysis

Differential peak analysis aims to detect meaningful differences between biological samples and typically involves multiple replicates in each group of interest. The inclusion of multiple samples enables the discrimination of natural variability within cell types from differentially enriched regions caused by the biological differences between tissues.

To analyse the use of imputed signals for detecting differential enrichment, we chose the H3K9ac assays for cell types E025 (Adipose-Derived Mesenchymal Stem Cell Cultured Cells) and E052 (Muscle Satellite Cultured Cells) as samples for a case study. These two epigenomic tracks are part of the hold-out test set, share a comparable number of tracks for the two tissues in the training, validation, and test sets, and have 4 and 3 measured replicates, respectively.

We tested the use of imputations for differential analysis with the DiffBind package [402]. DiffBind requires specifying the peak set and providing the aligned reads for

treatment and control in a sample sheet. For the measurement group, we use the data from the Roadmap consortium, which includes the aligned reads for each replicate in *.tagAlign* format and the peaks called with MACS2 in *.narrowPeak* format. In addition, each replicate is associated with a cell-type-specific control track. The aligned *.tagAlign* files are converted to BAM format with BedTools [489].

For the imputed tracks, we devised a procedure to simulate replicates from the predicted p-value signal. Such a procedure is an approximation that requires several assumptions, yet it suffices to provide a proof of concept for using imputation models in bioinformatics software pipelines. Nevertheless, for such applications, future imputation models would likely need to focus on modelling not only the average value of a signal but also the uncertainty involved.

To simulate replicates using p-value signals, we assume that the read counts at each genomic position follow a negative binomial distribution, a common model used for sequencing data [490–492]. To draw samples from this distribution, we need to estimate two parameters, mean and dispersion.

We estimate the mean parameter by inverting the procedure used to generate the p-value tracks. The p-value signal results from the survival function of a background Poisson model. The genomic-position specific mean parameter of the background distribution is calculated from tissue-specific control samples available in the Roadmap dataset. Using the parameters of the background distribution, the read counts for a given track are obtained through the inverse survival function (ISF).

To infer the dispersion parameters for each cell type, we made use of the software package DESeq2 [491]. Specifically, DESeq2 is capable of fitting a function that, given the mean value as input, can output the dispersion value. We selected tracks from the training set for H3K9ac whose tissues present sufficient similarity to the target tissues (E023 for E025 and E107-E108 for E052). These selected tracks are used to estimate the mean-dispersion functions, which are then used on the previously estimated mean parameters to generate bin-wise dispersion parameters.

Given these bin-specific parameters derived from the imputed signals, we generate three samples per cell type by sampling from the negative binomial distribution.

The MACS [399] command *bdgcmp* and *bdgpeakcall* were used to generate the peak set on the p-value track for each cell type in both measurement and simulated replicates. Next, the consensus peakset for all samples was estimated using DiffBind. ENCODE

blacklisted regions [493] were removed from the consensus peakset. Aggregated counts for these consensus peaks were computed for both measurement and simulated replicates, and were directly read into the DiffBind DBA object with *dba.peakset* function. Using a false discovery rate (FDR) threshold of 0.05, we determined the differentially enriched peaks for both measurements and imputations. An overview of the differential peak analysis pipeline can be found in A.22.

A.4 Interpreting the model

A.4.1 Global embeddings

eDICE learns global embeddings that capture the high-level relationships between assays and cell types, which we can display through a UMAP [494] projection.

Epigenetic modifications that are generally associated with the same functionality tend to cluster together in the learned representations (Figure A.10). Assigning global roles to epigenetic modifications neglects much of the nuance of the regulatory mechanisms of the genome. Nonetheless, we can consider some general patterns.

H3K9ac, H3K4me2, H3K4me3 are modifications typically associated with active promoters [495, 496], while H3K4me1 and H3K27ac (and partially H3K4me2) are linked with active enhancers [496, 497]. We classify these marks, together with the DNase-seq assay [498], under the broad label of activating modifications.

H3K27me3 is an antagonist to H3K27ac and is broadly associated to gene repression together with H3K9me3 [499, 500].

H3K36me3, H4K20me1, H3K79me1 are linked to exons [501], and we consider these modifications under the label of transcription-associated together with H3K79me2, which is associated with active gene bodies and alternative splicing [502, 503].

The partitioning of the epigenetic marks reflects the correlation patterns that emerge from the observed data. We calculated the average tracks for each assay across all cell types and performed agglomerative hierarchical clustering on these tracks. The results, summarised in Supplementary Figure A.13, show that modifications with shared global functions tend to cluster together.

Likewise, eDICE is capable of learning the broad similarities between tissue types through the global cell embeddings represented in Figure A.11.

A.4.2 Measures of attention

Inspired by the analysis of attention mechanisms in protein language models [504], we define the portion of attention that each attention head h dedicates to a specific assay a in the set of observed assays A over a portion g of the genome as:

$$p_h(\alpha) = \sum_{g \in G} \sum_{a \in A} w_{a\alpha}^{(g,h)} / \sum_{g \in G} \sum_{a \in A} \sum_{a' \in A} w_{aa'}^{(g,h)} \tag{A.1}$$

where $w_{a\alpha}^{(g,h)}$ denotes the attention weight for the key assay α given the query assay a in the genomic bin g for attention head h. Within the formalism of key-value-query attention in Transformer models, this definition of portion of attention corresponds to the average attention given to each key. Due to the softmax function used in the calculation of the attention weights in the scaled dot-product attention, the portion of attention adds up to one over the space of all assays for a single attention head, which makes it suitable to be expressed as a percentage of the total attention of the attention head.

To evaluate how the state of the attention blocks changes in correspondence with the functional regions of the genome, we calculated of the attention percentage while restricting the genomic bins to the aforementioned annotated regions and evaluated the relative difference in the percentage of attention dedicated to the assays compared to the global pattern in the chromosome. We define this attention shift for a given region R of the genome as:

$$d_h^{(R)}(\alpha) = \frac{p_h(\alpha)|_{G \equiv R} - p_h(\alpha)}{p_h(\alpha)} \tag{A.2}$$

where R is any type of annotated region of the genome, e.g. promoters or enhancers.

A.4.3 Interpreting the attention weights

Attention models assign a weight to the relationship between elements in an input set or sequence, which is often considered a proxy to interpret the underlying interactions between said elements. This area of research is mainly active within the application of Transformer models to natural language processing (NLP) tasks [505, 506], with a strong focus on the visualization of the attention weights (e.g. [507]).

Despite this excitement, recent works have begun to highlight the limitations of attention as an explanation method, showing that in specific contexts, attention weights are at best noisy predictors of the importance of input components for a model [508] or outright misleading and weak to adversarial attacks [509]. Although these studies present some critical limitations, such as the focus on NLP models or the comparison to alternative measures of feature importance which are not necessarily valid as "ground truth," we believe caution in the analysis of attention weights to be warranted. Adding these considerations to the non-deterministic result of training attention models, we consider the analysis presented here as a diagnostic exploration rather than a detailed claim on the biology examined by the model.

The results presented are derived from the same instance of eDICE used for the reference epigenome validation. Although the details vary due to the stochastic training of the attention block, we observed that the conclusions drawn tend to hold across training instances.

We specifically focus on the attention block used for the contextualization of the assay embeddings. The attention measures employed (detailed in Supplementary Section A.4.2) are based on averages of attention weights, which paint a higher-level picture than comparisons of individual coefficients.

To obtain a general overview of the eDICE model, we calculated the percentage of attention given by each head to each assay for the entire chromosome 21. These results, displayed in Figure A.12, highlight how at least two of the attention heads consistently dedicate a significant portion of attention to just a few histone modifications. These marks are among the core modifications extensively mapped across many cell types in the Roadmap dataset, i.e. H3K27ac, H3K36me3 H3K9me3, H3K4me1, and H3K4me3.

Interestingly, this phenomenon proved quite consistent over different experiments, with minor variations such as focusing on H3K27me3 rather than H3K27ac. Overall, this result suggests that the model relies extensively on marks for which a large amount of information is present during training.

To further explore the workings of the attention layer in eDICE, we gathered genomic annotations for chromosome 21 from the Ensembl [510] version 104 GRCh37 assembly for protein-coding gene bodies, enhancers, promoters, open chromatin, and transcription factors binding sites. We calculated how the percentage of attention shifts within these functional regions of the genome and displayed the results in Figure A.14.

Overall, these attention shifts reveal known patterns from the literature yet present unexplained peculiarities, such as the significant shift within enhancer regions for H3K4me1, a known chromatin hallmark for these functional regions. Intuitively, one would expect the attention weight for H3K4me1 to grow within enhancers, given its biological importance, yet we observe the opposite, especially for attention head 2. Nevertheless, this pattern is inconsistent for such a setting: H3K36me3, a histone modification enriched on the gene body region associated with active gene transcription [511], shows a positive average attention shift within protein-coding genes.

Much remains to be understood about the workings of attention models. Therefore, we limit our analysis to the qualitative consideration that the larger attention shifts within functional regions correspond to known epigenetic marks that characterize said regions. Examples of these patterns include the DNase shifts for open chromatin regions [512], the shifts in DNase and H3K4me1 for transcription factor binding sites [513], and the shifts within promoters for H3K4 methylation [514], H3K9ac, H3K27ac, and their antagonists H3K9me3 and H3K27me3 [495].

We include two high-level visual summaries of the attention shifts in Figures A.15 and A.16, in which the absolute value of the attention shifts was summed up over the attention heads and the assays, respectively, and then 0-1 normalized for each annotated category. The resulting heatmaps provide at a glance an overview of which assays present the most considerable relative changes and which attention heads are significantly affected within functional regions of the genome.

Interestingly, it appears that the attention heads specialize at least partially, with attention head 2 being the most affected in promoters, enhancers, and open chromatin regions. In contrast, attention head 3 presents the most significant shifts for protein-coding gene bodies and transcription factor binding sites. The scarce shifts for attention head 4 also indicate that this head is not contributing significantly to the final prediction, a known phenomenon in the use of multi-headed attention models, and could be a candidate for pruning [515].

A.5 ENCODE imputation challenge model

An early version of our model was used as the basis of our third-placed entry in the 2019 ENCODE Imputation Challenge [328]. This model follows Avocado in adopting the factorized structure of Equation 4.1, but instead of directly parametrizing genomic

bin representations $\mathbf{b}_k$, computes non-linear, low dimensional transformations of the input signal at each bin,

$$\hat{Y}_{ijk} = g_\theta(\mathbf{c}_i, \mathbf{a}_j, f_\phi(\text{vec}(Y^k))) \quad , \tag{A.3}$$

where $\text{vec}(Y^k)$ is a fixed-size vector whose entries are the signal values for all observed tracks at the kth bin, together with zeros for all unobserved tracks. In practice, the functions g_θ and f_ϕ are both implemented with two-layer MLPs, and $\mathbf{c}_i$ and $\mathbf{a}_i$ are learned cell and assay embeddings, equivalent to the 'global embeddings' $\mathbf{u}_{c_i}$ and $\mathbf{u}_{a_i}$ described in the Online Methods (Equations 4.4 and 4.6). This model can be seen as a kind of denoising autoencoder in which the decoder has a factorized structure, inherited from Avocado, allowing it to impute signal values for previously unobserved cell-assay combinations. In contrast, our model is fully factorized, replacing a single global transformation of the input signal with cell- and assay-dependent transformations of the input signal, learned via self-attention (Equation 4.2).

The challenge model is used as a strong baseline in our experiments, since it has demonstrated competitive performance with state-of-the-art methods including Avocado in the blind-test setting of the ENCODE Imputation Challenge.

Appendix B

Supplementary material for Chapter 5

B.1 Supplementary Tables

Table B.1 **Overview of the DRIAMS datasets.** Location, collection period, and number of samples gathered for the four DRIAMS datasets.

Dataset	Laboratory	Collection period	MALDI-TOF MS	AMR labels	AM drugs
DRIAMS-A	University Hospital Basel	34 months (11/2015–08/2018)	145,341	3 101,660	71
DRIAMS-B	Canton Hospital Basel-Land	6 months (01/2018–06/2018)	6,416	37,453	44
DRIAMS-C	Canton Hospital Aarau	8 months (01/2018–08/2018)	22,500	50,114	56
DRIAMS-D	Viollier	6 months (01/2018–06/2018)	75,813	98,708	52

Table B.2 **Direct AMR prediction results with multi-drug models.** The best average metric is highlighted for each dataset and split.

Dataset	Split type	Model	Cross-validation performance score - mean (SD)		
			AUPRC	**Balanced accuracy**	**MCC**
DRIAMS-A	Random	PCA + LR	0.644 (0.003)	0.704 (0.002)	0.491 (0.003)
		Species-ResMLP	0.30 (0.04)	0.53 (0.02)	0.08 (0.05)
		ResMLP	**0.92 (0.01)**	**0.90 (0.01)**	**0.81 (0.01)**
	Species-drug zero-shot	PCA + LR	0.28 (0.02)	0.57 (0.01)	0.16 (0.02)
		Species-ResMLP	0.31 (0.04)	0.54 (0.02)	0.09 (0.03)
		ResMLP	**0.42 (0.03)**	**0.64 (0.02)**	**0.28 (0.03)**
	Drug zero-shot	PCA + LR	0.35 (0.27)	0.55 (0.14)	0.11 (0.25)
		Species-ResMLP	0.34 (0.32)	0.51 (0.07)	0.02 (0.14)
		ResMLP	**0.47 (0.29)**	**0.65 (0.13)**	**0.28 (0.24)**
DRIAMS-B	Random	PCA + LR	0.64 (0.02)	0.705 (0.007)	0.51 (0.02)
		Siamese + LR	0.49 (0.01)	0.76 (0.01)	0.53 (0.02)
		Species-ResMLP	0.35 (0.04)	0.59 (0.03)	0.21 (0.05)
		ResMLP	**0.87 (0.02)**	**0.90 (0.01)**	**0.79 (0.02)**
	Species-drug zero-shot	PCA + LR	0.44 (0.04)	0.63 (0.02)	0.30 (0.04)
		Siamese + LR	0.42 (0.01)	0.664 (0.004)	**0.40 (0.01)**
		Species-ResMLP	0.52 (0.04)	0.62 (0.02)	0.30 (0.04)
		ResMLP	**0.54 (0.04)**	**0.70 (0.02)**	0.39 (0.03)
	Drug zero-shot	PCA + LR	0.33 (0.25)	0.57 (0.12)	0.12 (0.16)
		Siamese + LR	0.18 (0.16)	0.52 (0.05)	0.08 (0.14)
		Species-ResMLP	0.17 (0.16)	0.50 (0.12)	0.01 (0.17)
		ResMLP	**0.47 (0.31)**	**0.71 (0.15)**	**0.35 (0.28)**
DRIAMS-C	Random	PCA + LR	0.62 (0.01)	0.72 (0.01)	0.49 (0.01)
		Species-ResMLP	0.41 (0.11)	0.59 (0.06)	0.17 (0.13)
		ResMLP	**0.92 (0.01)**	**0.89 (0.01)**	**0.81 (0.02)**
	Species-drug zero-shot	PCA + LR	0.39 (0.04)	0.60 (0.02)	0.23 (0.05)
		Species-ResMLP	0.48 (0.05)	0.62 (0.04)	0.28 (0.08)
		ResMLP	**0.55 (0.03)**	**0.69 (0.02)**	**0.39 (0.00)**
	Drug zero-shot	PCA + LR	0.24 (0.29)	0.63 (0.23)	0.03 (0.13)
		Species-ResMLP	0.22 (0.27)	0.47 (0.23)	0.05 (0.26)
		ResMLP	**0.34 (0.35)**	**0.66 (0.25)**	**0.17 (0.28)**
DRIAMS-D	Random	PCA + LR	0.67 (0.01)	0.76 (0.01)	0.60 (0.01)
		Species-ResMLP	0.57 (0.07)	0.71 (0.03)	0.47 (0.08)
		ResMLP	**0.76 (0.01)**	**0.82 (0.01)**	**0.64 (0.01)**
	Species-drug zero-shot	PCA + LR	0.49 (0.06)	0.63 (0.02)	0.36 (0.05)
		Species-ResMLP	**0.67 (0.04)**	**0.72 (0.02)**	**0.51 (0.04)**
		ResMLP	0.63 (0.05)	0.72 (0.03)	0.47 (0.05)
	Drug zero-shot	PCA + LR	0.23 (0.27)	0.52 (0.15)	0.02 (0.04)
		Species-ResMLP	0.18 (0.28)	0.52 (0.16)	0.01 (0.05)
		ResMLP	**0.36 (0.31)**	**0.66 (0.16)**	**0.20 (0.22)**

Table B.3 **Results of one-way Kruskal-Wallis analysis of variance for comparing the performance of the ResMLP model using the different categories of molecular fingerprints (MACCS, 1024-dimensional Morgan, Pubchem).** Using the random split setting, 10 train/test splits were performed on each of the 4 DRIAMS collection sites. A ResMLP model was trained using the same configuration as the main section of the classification experiments. The resulting metrics were used to calculate the Kruskal-Wallis H statistic. The resulting p-values indicate that no fingerprint choice leads to significantly different results, using a significance threshold of 0.05.

Metric	Kruskal-Wallis H Statistic	P-Value
MCC	4.296865	0.116667
AUPRC	1.373165	0.503293
Balanced Accuracy	1.367955	0.504606

B.2 Supplementary Figures

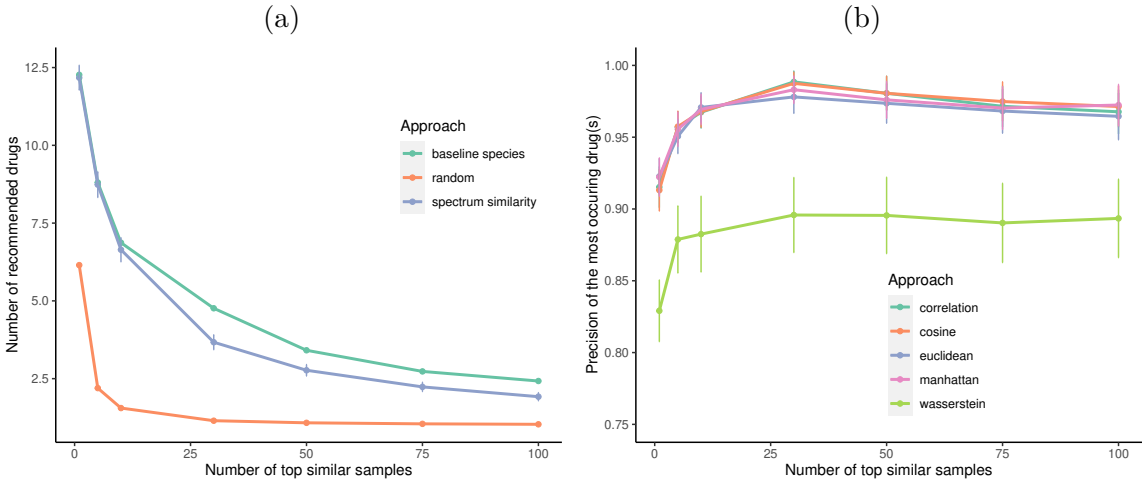


Fig. B.1 **Properties of recommendations generated with the baselines that depend on the selection of samples.** **(a)** Comparison of the number of drugs recommended with different variations of the approach based on the similarity between spectra. The error bars represent the 95% confidence interval. **(b)** Comparison of performance of different variants of the spectrum similarity set-up based on the top k similarity. The y-axis shows the average precision across all individuals in the test set. The error bars represent the 95% confidence interval.

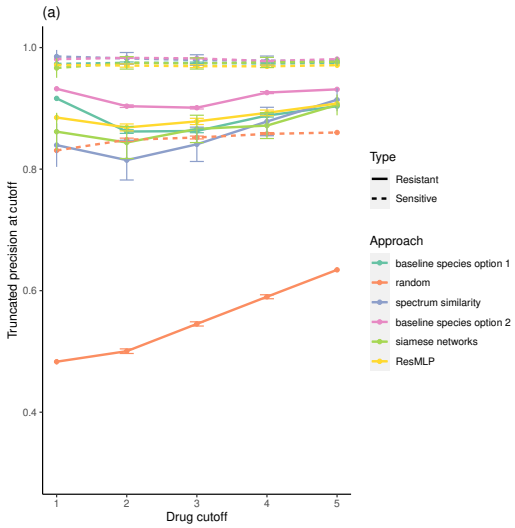
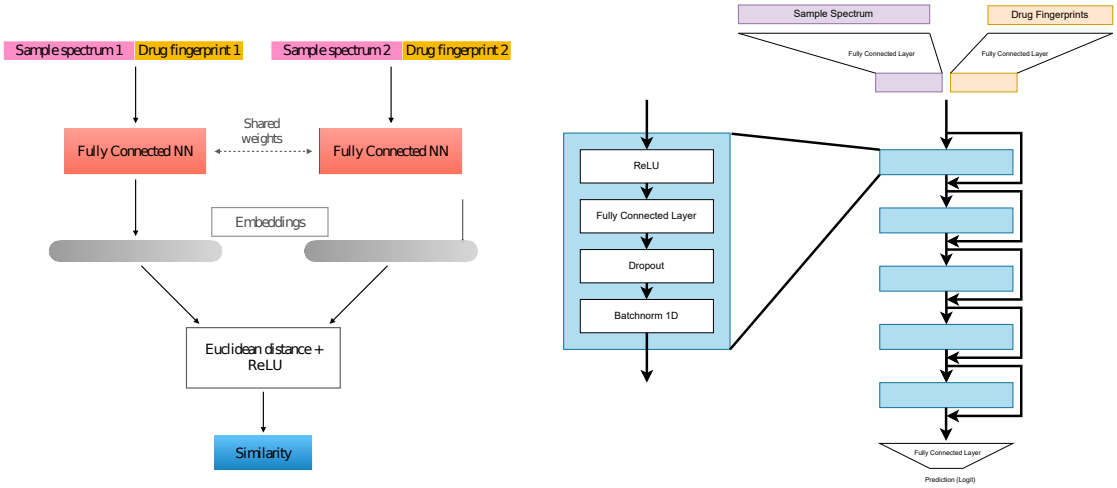


Fig. B.2 **Truncated precision for the recommendation of drugs to which the sample is sensitive.** Truncated precision at cut-offs 1, 2, 3, 4, and 5 for the different recommendation set-ups. For the random baseline, baseline species, and spectrum similarity approaches, the number of top neighbours k is set to 30.



(a) Siamese network (b) Residual Multi-Layer Perceptron network

Fig. B.3 **Architecture of the Siamese networks and ResMLP model.** Schematic representations of the Siamese network **(a)** and ResMLP **(b)** used for the experiments performed.

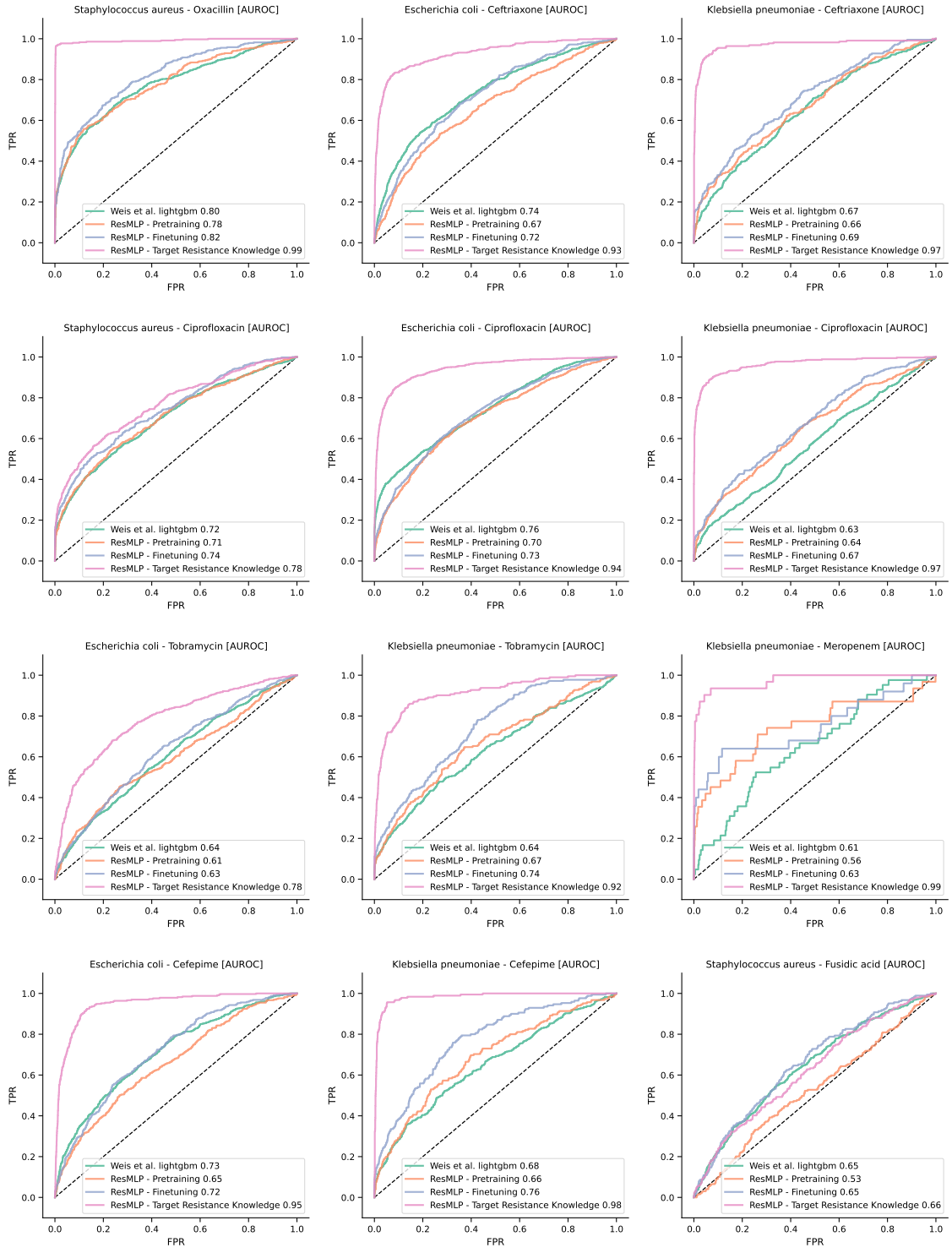


Fig. B.4 **ROC curves for 12 selected pathogen-drug combinations.** Full set of Receiver Operating Characteristic curves for the combinations presented in Figure 2 from [440]. Numerical values are reported from the average performance over 5 train-test splits for the target drug-pathogen combination.

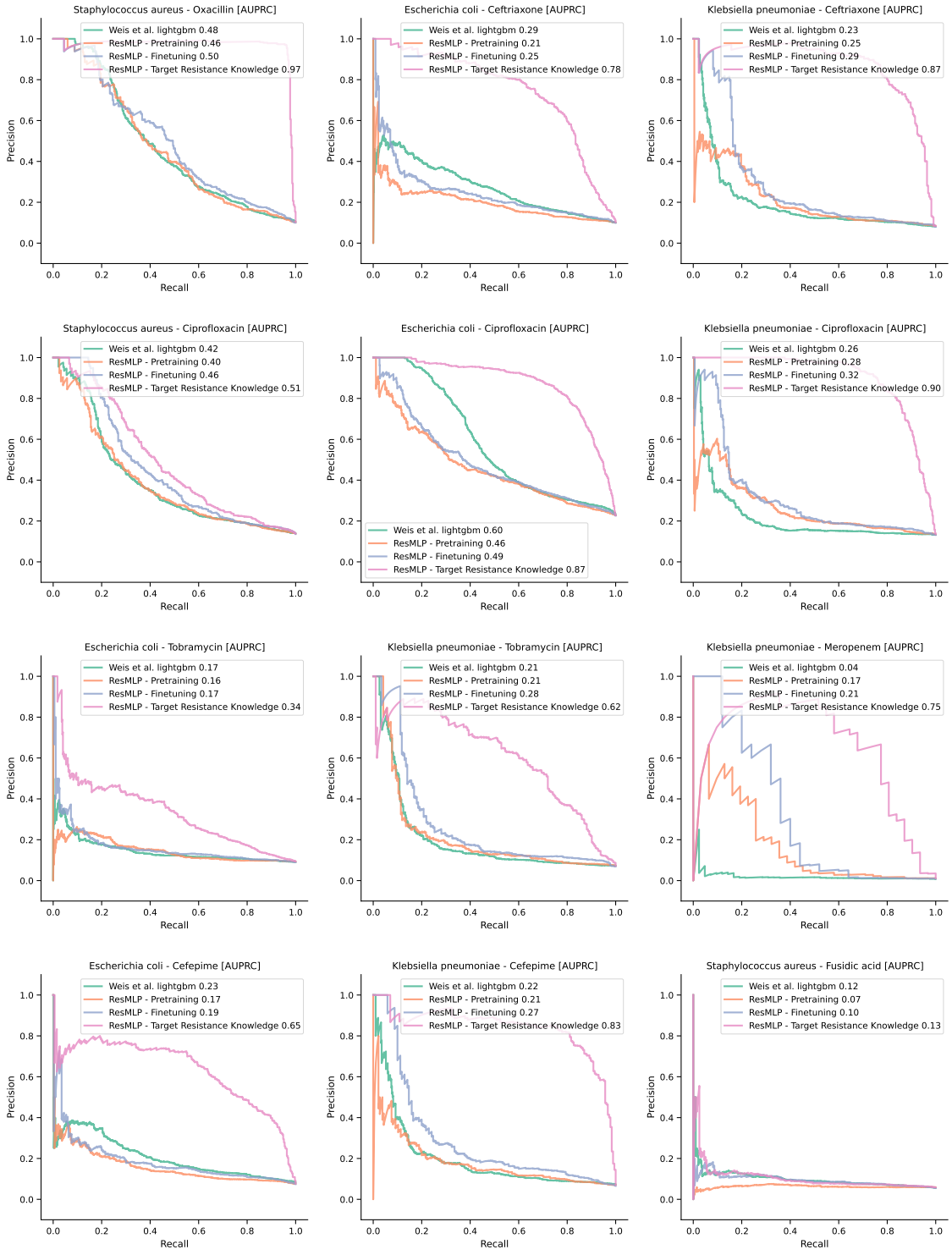


Fig. B.5 **PR curves for 12 selected pathogen-drug combinations.** Full set of Precision-Recall curves for the combinations presented in Figure 2 from [440]. Numerical values are reported from the average performance over 5 train-test splits for the target drug-pathogen combination.

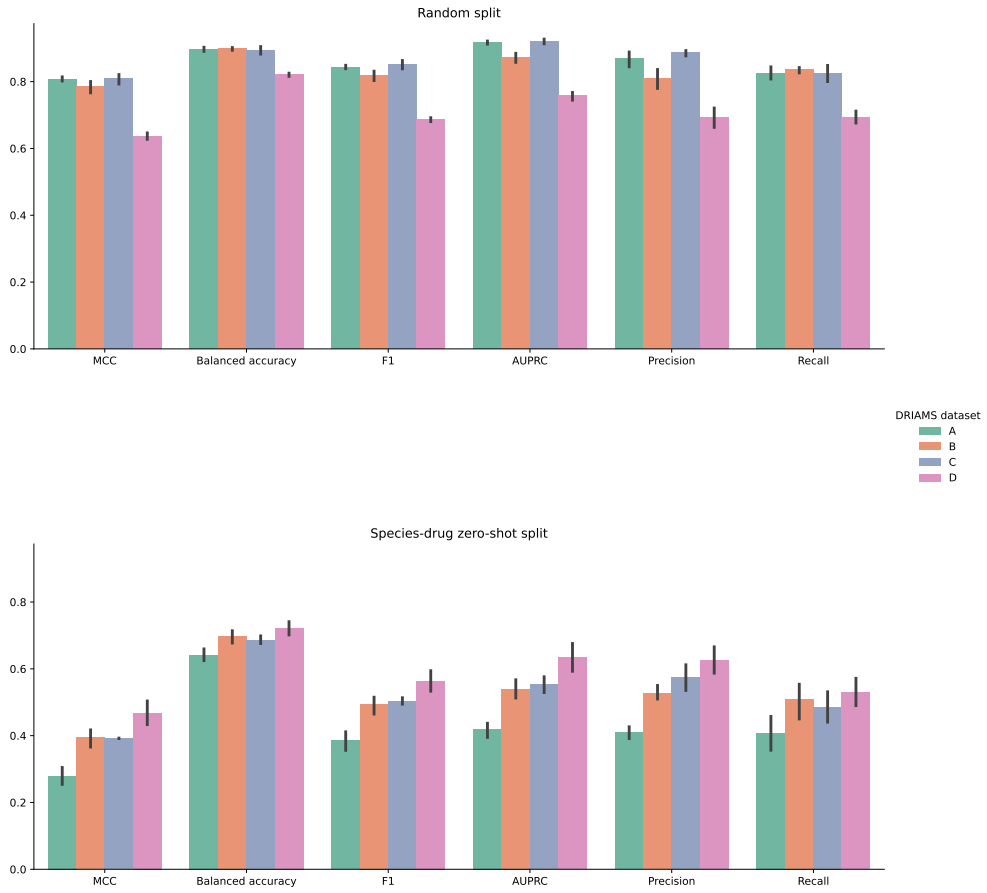


Fig. B.6 **Visualization of the performance metrics for two data-generating splits.** Comparison of ResMLP predictions on the four DRIAMS datasets for random and species-drug zero-shot data splits. The error bars reported represent a 95% confidence interval.

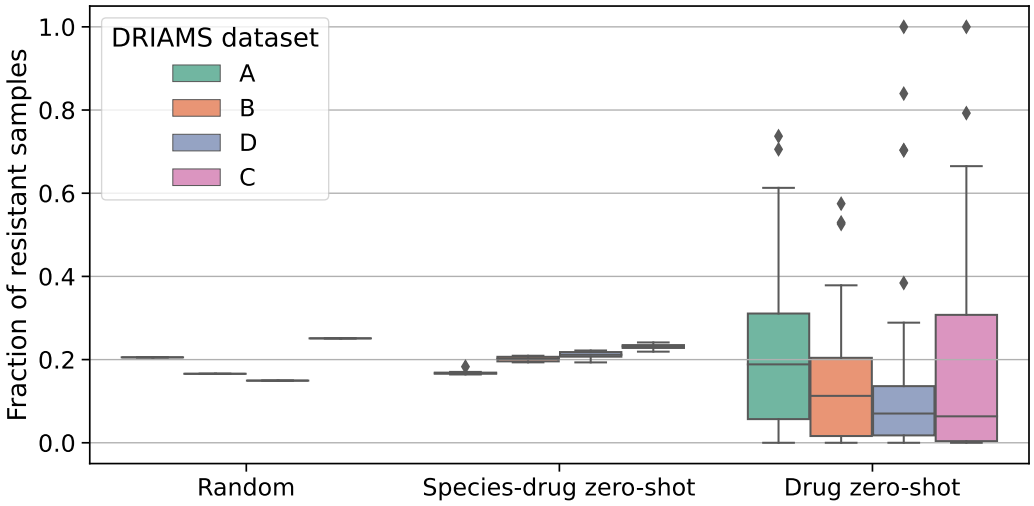


Fig. B.7 **Test set imbalance.** Fraction of the test samples with a positive label (i.e. displaying resistance to the drug). The random split allows us to sample with stratification to obtain consistent test splits. The species-drug zero-shot split relies on a heuristic to construct test sets that approximately constitute a specific percentage of the total available samples, leading to small fluctuations in the fraction of positive samples in each test split. The drug zero-shot setting, on the other hand, does not allow for any control on the imbalance of the test split. The resulting variability is a challenge for the machine learning models tested to overcome.

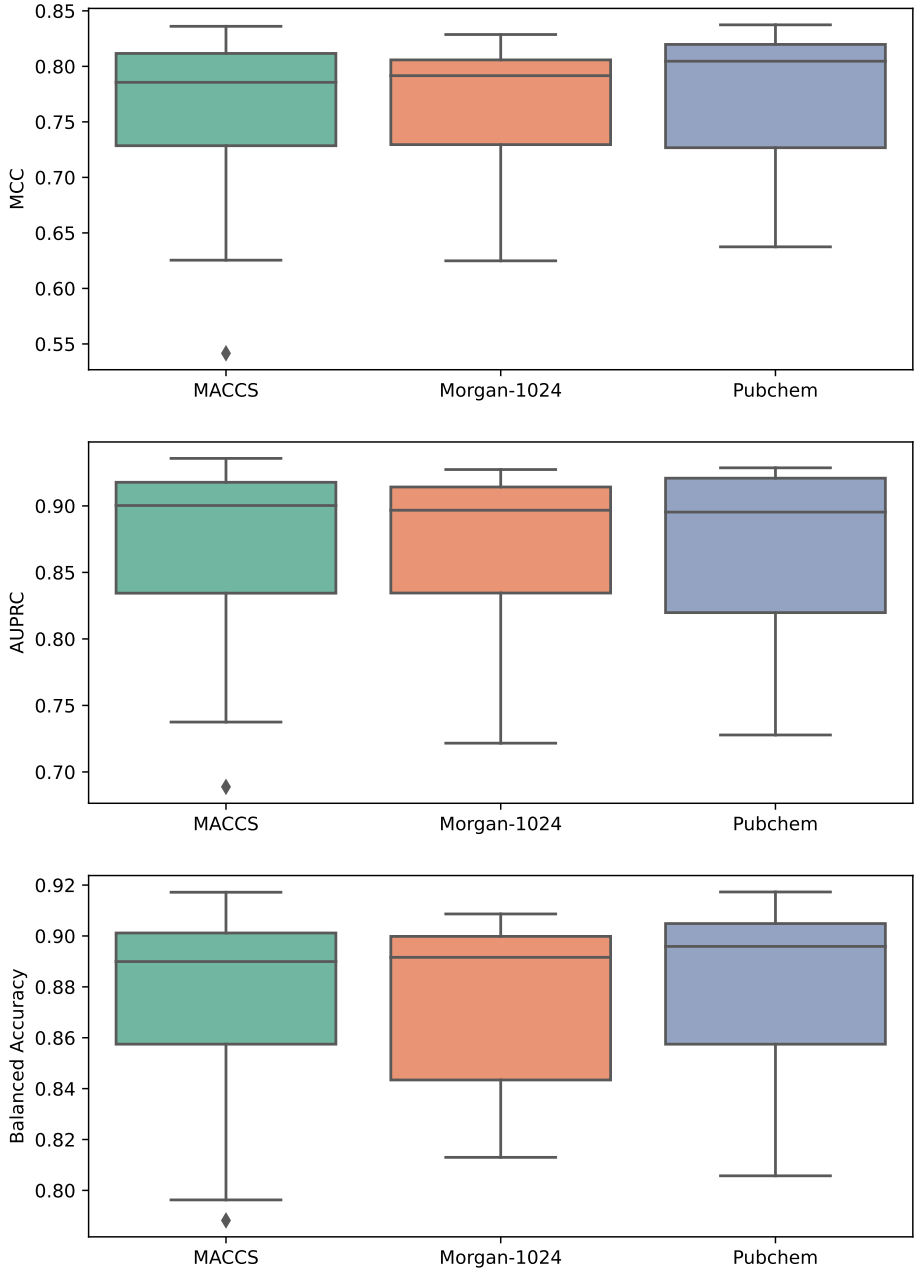


Fig. B.8 **Comparison of the performance of a ResMLP model trained using the different types of molecular fingerprints.** For each DRIAMS collection site, 10 train/test splits are randomly selected, and a ResMLP model is trained using the same configuration presented for the results in the classification performance. The resulting metrics display a small level of variation. An analysis of variance test, however, confirmed that the choice of molecular fingerprint is not statistically significant.

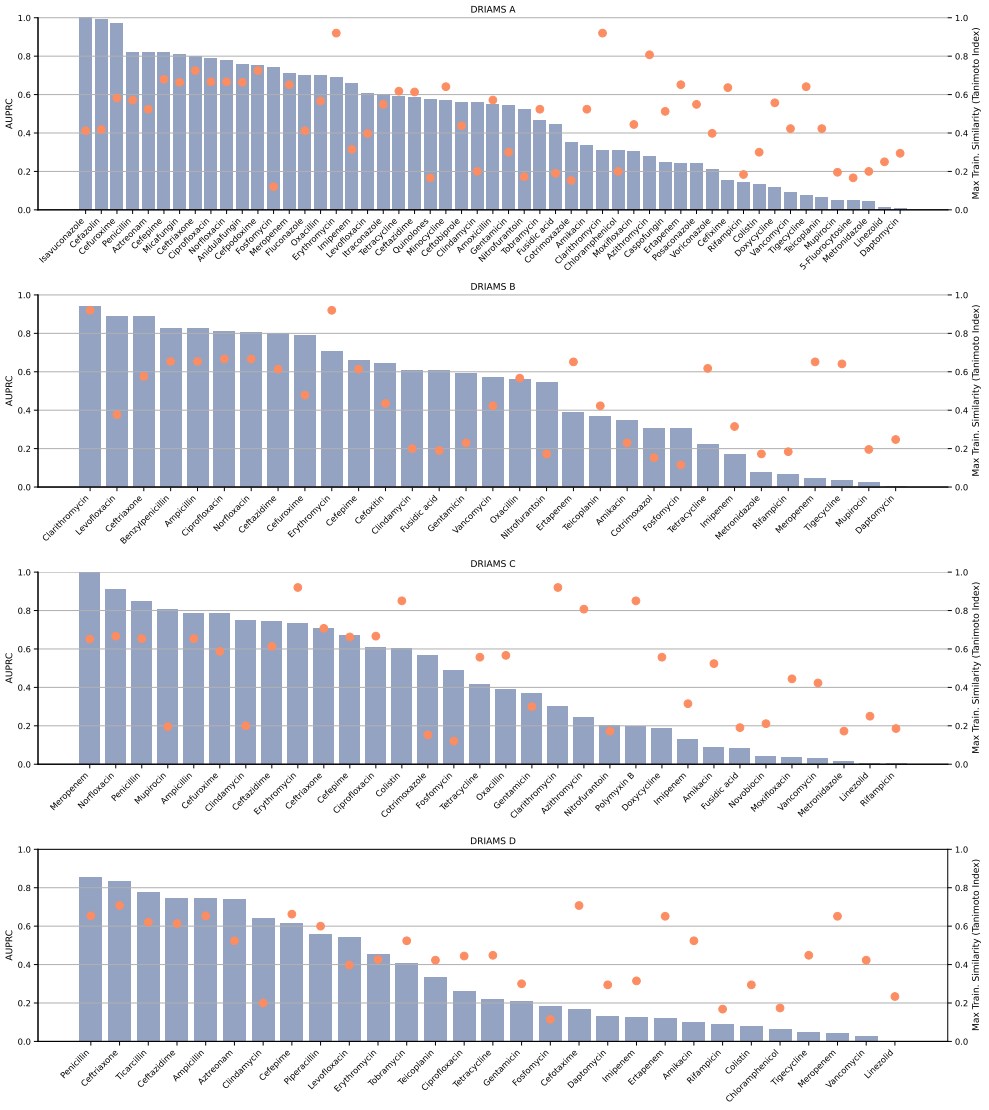


Fig. B.9 **AUPRC for the predictions in the drug zero-shot task on the four DRIAMS datasets.** Drugs for which only one class of response was available have been excluded from the analysis. In orange, we represented as dots the highest Tanimoto index calculated by comparing the MACCS fingerprints of the target drug with the rest of the compounds in the dataset. This measure of similarity does not appear to share a pattern with the generalization performance measured by the AUPRC.

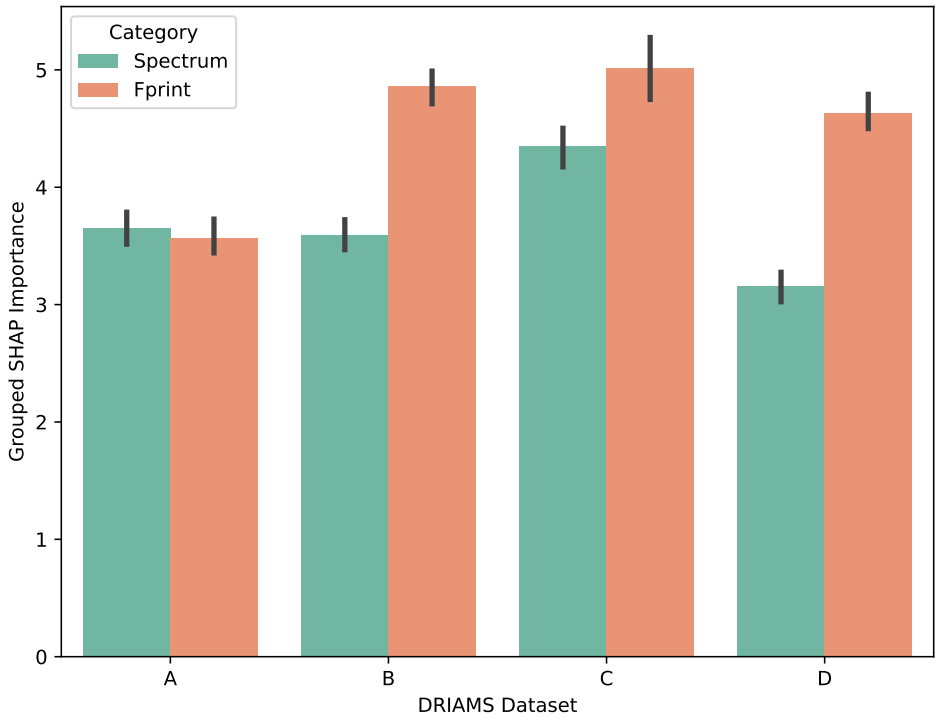


Fig. B.10 **Grouped SHAP importance for the two sets of features.** Across multiple training runs with different randomization, the ResMLP model assigns almost equal overall importance to the spectrum features and the chemical features, as measured by SHAP

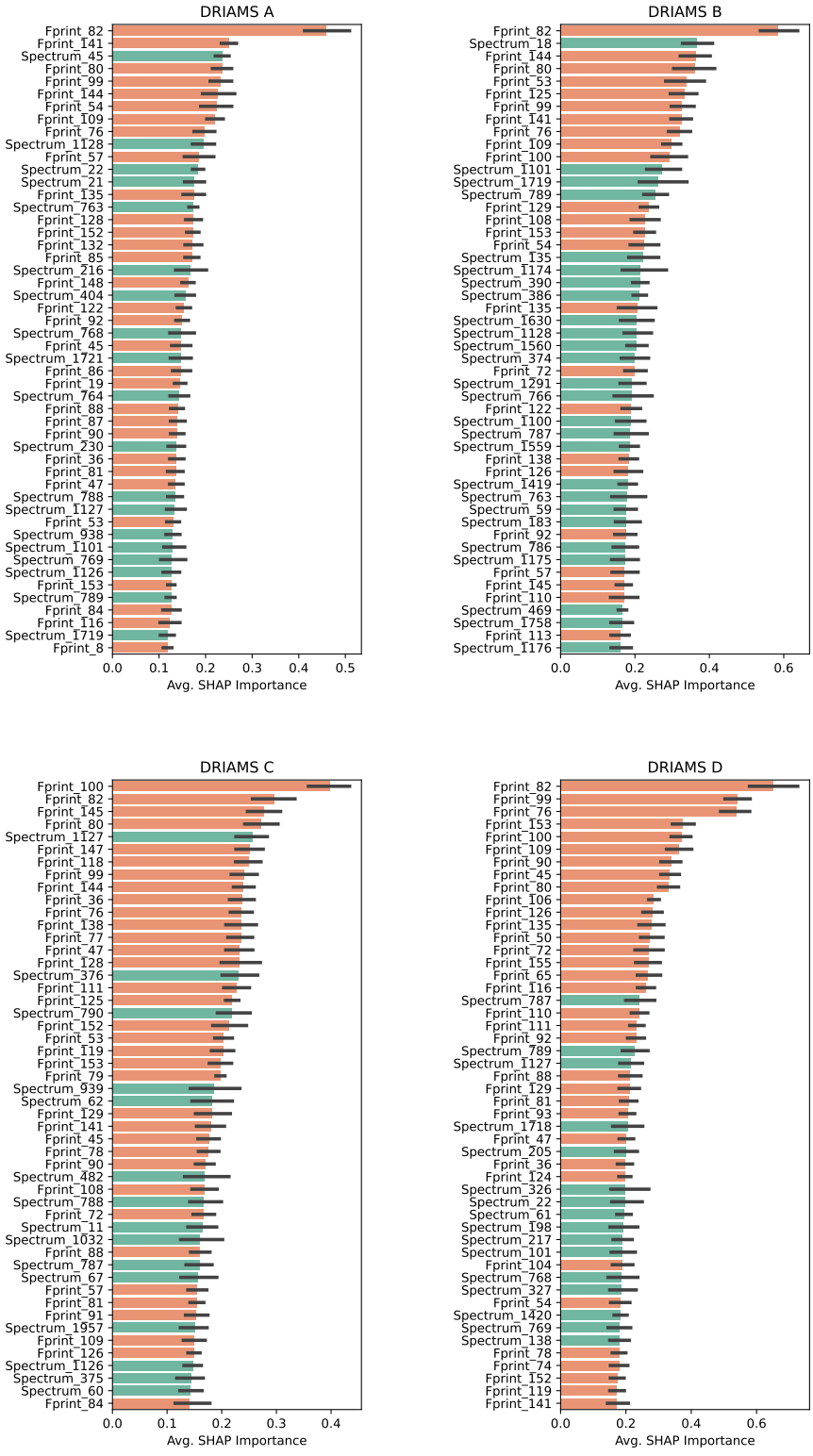


Fig. B.11 **Most important features according to SHAP.** The 50 most important features for the AMR prediction task for the ResMLP model trained on the MACCS fingerprints, split by DRIAMS dataset. In each case, we clearly see contributions from both the spectrum and the chemical fingerprint.

B.3 Hyperparameters and training configurations

Siamese parameters in all analyses: 512 generated features, models trained for 200 epochs with a batch size of 256, and input of 100,000 pairs.

The PCA baseline used a number of components sufficient to capture 95% of the variance in the training set for each data source (i.e. MALDI-TOF spectra and chemical fingerprints). The resulting projections are concatenated before being used to train logistic regression models. The results presented in the paper are obtained using the MACCS keys as chemical fingerprints.

The ResMLP configuration and training are kept constant in all experiments. The input projections encode the 6000-dimensional MALDI-TOF spectra and the 1024-dimensional Morgan fingerprints into 512-dimensional vectors, which are then concatenated. The model consists of 5 residual blocks, including ReLU activation, a linear layer of dimensionality 1024, a dropout layer with probability 0.2, and a BatchNorm layer. The ResMLP is trained with early stopping with a patience parameter of 50 epochs using an Adam optimiser with a learning rate of $3 * 10^{-4}$ and a weight decay of 10^{-5}. For the predictions on the single drug and species combinations, to compare to previous work, the first four blocks of the model were frozen, and the last block was additionally tuned with a reduced learning rate of $3 * 10^{-5}$ on the subset of samples for the target drug-species combinations that are not part of the test split using early stopping.

B.4 Drug recommendation - Evaluation

To evaluate the recommendations produced, we use as metric the precision P, defined as the number of correct drugs recommended divided by the number of drugs in the intersection between the test data and the recommendation set. If no drug exists in the intersection, precision is set to 0. To analyse the effect of choosing different sizes for the recommendation set, we use the precision at cutoff n ($P@n$), which corresponds to the precision calculated for the top n recommendations. The truncated version of $P@n$ adjusts the calculation when fewer than n items are retrieved to avoid penalizing the system for returning fewer items. Finally, we computed the mean Average precision at cutoff n ($mAP@n$), an informative measure that considers not only the number of correct predictions but also the order of the recommended drugs. $mAP@n$ is the

mean of the Average Precision at cutoff n, $AP@n$, calculated over all available queries. $AP@n$ is calculated as $AP@n = \frac{1}{n}\sum_{k=1}^{n} P@k$.

In the literature, an alternative formulation of $AP@n$ is also frequently found, which is calculated as $AP@n = \frac{1}{\min(n,TP(total))}\sum_{i=1}^{n}\frac{TP(i)\cdot \mathrm{rel}(i)}{i}$, where TP stands for True Positives, and rel(i) is a binary indicator with value 1 if the i^{th} item is relevant, 0 otherwise. This alternate formulation focuses more on the relevant items, which may lead to overly optimistic estimates in case of sparsity of relevant items in the top-n recommendations. The previously introduced definition, on the other hand, offers a more comprehensive summary of the recommendation results overall.

B.5 SHAP feature importance

To analyse the feature importance of the ResMLP model and verify the impact of the chemical fingerprints, we employed SHAP analysis [458]. We used the DeepExplainer implementation of the Python SHAP package.

To interpret the importance assigned by the ResMLP to the chemical features, the model was trained with the same configuration as for the quantitative evaluation of the classification performance over 10 randomized train/test splits for each DRIAMS dataset. The overall importance of a feature is computed by averaging the absolute value of the SHAP value for that feature over the samples in the test set. The importance of a group of features is computed by first summing the SHAP values corresponding to the features for each sample, and then calculating the overall importance of this sum as before.