THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

New AI-based methods for studying antibiotic-resistant bacteria

JUAN SALVADOR INDA DÍAZ

UNIVERSITY OF GOTHENBURG

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
Gothenburg, Sweden 2023

New AI-based methods for studying antibiotic-resistant bacteria
Juan Salvador Inda Díaz
Gothenburg 2023
ISBN: 978-91-8069-503-9 (PRINT)
ISBN: 978-91-8069-504-6 (PDF)
Available at: http://hdl.handle.net/2077/78675

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
SE-412 96 Gothenburg
Sweden

Typeset with LaTeX
Printed by Stema Specialtryck AB, Borås, Sweden, 2023

New AI-based methods for studying antibiotic-resistant bacteria

Juan Salvador Inda Díaz

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Abstract

Antibiotic resistance is a growing challenge for human health, causing millions of deaths worldwide annually. Antibiotic resistance genes (ARGs), acquired through mutations in existing genes or horizontal gene transfer, are the primary cause of bacterial resistance. In clinical settings, the increased prevalence of multidrug-resistant bacteria has severely compromised the effectiveness of antibiotic treatments. The rapid development of artificial intelligence (AI) has facilitated the analysis and interpretation of complex data and provided new possibilities to face this problem. This is demonstrated in this thesis, where new AI methods for the surveillance and diagnostics of antibiotic-resistant bacteria are presented in the form of three scientific papers.

Paper I presents a comprehensive characterization of the resistome in various microbial communities, covering both well–studied *established* ARGs and *latent* ARGs not currently found in existing repositories. A widespread presence of latent ARGs was observed in all examined environments, signifying a potential reservoir for recruitment to pathogens. Moreover, some latent ARGs exhibited high mobile potential and were located in human pathogens. Hence, they could constitute emerging threats to human health. Paper II introduces a new AI-based method for identifying novel ARGs from metagenomic data. This method demonstrated high performance in identifying short fragments associated with 20 distinct ARG classes with an average accuracy of 96%. The method, based on transformers, significantly surpassed established alignment-based techniques. Paper III presents a novel AI-based method to predict complete antibiotic susceptibility profiles using patient data and incomplete diagnostic information. The method incorporates conformal prediction and accurately predicts, while controlling the error rates, susceptibility profiles for the 16 included antibiotics even when diagnostic information was limited.

The results presented in this thesis conclude that recent AI methodologies have the potential to improve the diagnostics and surveillance of antibiotic-resistant bacteria.

Keywords: Transformers, Antibiotic Resistance, Infectious Diseases, Metagenomics, Data-driven Diagnostics

Sammanfattning på svenska

Antibiotikaresistens är ett växande folkhälsoproblem som årligen orsakar miljontals dödsfall globalt. Antibiotikaresistensgener (ARGs) är den primära orsaken till antibiotikaresistens hos bakterier. Dessa resistenta gener uppkommer antingen via mutationer på existerande gener eller via horisontell genöverföring mellan bakterier. Sjukvården är beroende av effektiva antibiotika och förekomsten av multi-resistenta bakterier är ett alvarligt problem som försvårar eller omöjliggör behandling av infektioner, vilket i sin tur orsakar lidande och höga kostnader. Den snabba utvecklingen inom artificiell intelligens (AI) har dock möjliggjort analys och tolkning av komplexa data på ett sätt som öppnar upp för potentiella möjligheter att tackla problemet med antibiotikaresistens. I denna uppsats presenteras tre artiklar som utforskar nya AI-metoder för att hitta och kartlägga spridningen av antibiotikaresistenta bakterier samt testa bakterier för antibiotikaresistens.

I den första artikeln "Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes" presenteras en övergripande karaktärisering av resistensgener hos olika grupper av mikrober som täcker både välstuderade och *etablerade* ARGs samt *latenta* gener om inte återfinns i existerande genbanker. En påtaglig närvaro av latenta ARGs observerades i alla studerade prover, vilket representerar potentiella källor för rekrytering för patogener. En del latenta ARGs uppvisade dessutom en potentiellt hög mobilitet och återfanns i mänskliga patogener. Följaktligen skulle de därmed kunna utgöra hot mot folkhälsan. I den andra artikeln "Alignment-free identification of antibiotic resistance genes" presenteras en nyutvecklad AI metod för identifiering av tidigare okända ARGs från metagenomiska data. Metoden uppvisar goda resultat, den ger rätt svar i 96% av fallen när det gäller att identifiera korta fragment associerade med 20 distinkta klasser för antibiotikaresistenta gener. Metoden, som är transformer-modell-baserad, överträffar etablerade sekvensalignment-metoder. Den tredje artikeln "Confidence-based Prediction of Antibiotic Resistance at the Patient-level Using Transformers" introducerar en AI-metod för att prediktera fullständiga känslighetsprofiler för antibiotika från patientdata och inkomplett information gällande bakteriernas antibiotikaresistens. Felfrekvensen kontrolleras med hjälp av conformal prediction och i en studie med 16 antibiotika predikterade metoden rätt känslighetsprofil i alla fall, även när informationen gällande bakterierna var begränsad.

Dessa resultat indikerar att metoden skulle kunna underlätta i sökandet efter och kartläggningen av antibiotikaresistenta bakterier samt testandet av bakterier för antibiotikaresistens.

Nyckelord: Transformer-modell, Antibiotikaresistens, Infektionssjukdomar, Metagenomik, Data-driven diagnostik

Resumen en español

La resistencia de las bacterias a los antibióticos es un desafío creciente para la atención médica y la salud pública, misma que causa millones de muertes al año en todo el mundo. Los genes de resistencia a antibióticos (ARGs por sus siglas en inglés), que pueden ser adquiridos a través de mutaciones o a través de la transferencia horizontal, son la principal causa de la resistencia bacteriana. En entornos clínicos, la prevalencia de bacterias resistentes a múltiples fármacos ha comprometido gravemente la efectividad de los tratamientos con antibióticos. Por otro lado, el rápido desarrollo de la inteligencia artificial (AI por sus siglas en inglés) ha facilitado el análisis e interpretación de datos complejos, y ha brindado nuevas posibilidades para abordar éste problema. En esta tesis se desarrollaron tres nuevos métodos de AI para la vigilancia y el diagnóstico de bacterias resistentes a los antibióticos, y se presentan en forma de tres artículos científicos.

El Artículo I "Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes" presenta una caracterización exhaustiva del resistoma en diversas comunidades microbianas, incluyendo tanto ARGs "establecidos" (o bien estudiados) como ARGs "latentes" (o que no se encuentran en repositorios existentes). Se observó una presencia generalizada de ARGs latentes en todos los entornos examinados, lo que indica una reserva potencial para el reclutamiento de dichos genes por organismos patógenos. Además, algunos ARGs latentes mostraron un alto potencial de movilidad y se encontraron en patógenos humanos, por lo que podrían constituir amenazas emergentes para la salud humana. El artículo II "Alignment-free identification of antibiotic resistance genes" presenta un nuevo método basado en IA desarrollado para la identificación de nuevos ARGs a partir de datos metagenómicos. Este método exhibió un alto rendimiento para identificar fragmentos cortos de genes asociados a 20 clases distintas de ARGs con una precisión media del 96%. El método, basado en transformers, superó significativamente las técnicas basadas en *alineación de secuencias* ya establecidas. El Artículo III "Confidence-based Prediction of Antibiotic Resistance at the Patient-level Using Transformers" presenta un método nuevo basado en AI para predecir perfiles completos de susceptibilidad a antibióticos utilizando datos del paciente e información de diagnóstico incompleta. El método incorpora conformal prediction y predijo con precisión perfiles de susceptibilidad para los 16 antibióticos incluidos, controlando las tasas de error, aún cuando la información de diagnóstico fue limitada.

Esta tesis propone tres métodos novedosos de AI para abordar el desafío de la resistencia a los antibióticos habiendo determinado la latencia de genes resistentes, para poder identificar nuevos genes resistentes, y predecir perfiles de susceptibilidad, aún con información incompleta. Los resultados presentados

en este trabajo prodrían contribuir a mejorar el diagnóstico y la vigilancia de organismos como las bacterias resistentes a los antibióticos.

Palabras clave: Transformers, Resistencia a los Antibióticos, Enfermedades Infecciosas, Metagenómica, Diagnóstico basado en Datos

List of publications

This thesis is based on the work represented by the following papers:

I. **Inda-Díaz, J.S.**, Lund, D., Parras-Moltó, M., Johnning, A., Bengtsson-Palme, J., and Kristiansson, E. (2023). Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes. *Microbiome* 11(44).

II. **Inda-Díaz, J.S.**, Örtenberg-Toftås, M., Salomonsson, E., Berglund, F., Johnning, A., and Kristiansson, E. (2023). Identification of short fragments of antibiotic-resistance genes using transformers. *Manuscript*.

III. **Inda-Díaz, J.S.**, Johnning, A., Hessel, M., Sjöberg, A., Lokrantz, A., Helldal, L., Jirstrand, M., Svensson, L., and Kristiansson, E. (2023). Confidence-based Prediction of Antibiotic Resistance at the Patient-level Using Transformers. bioRxiv 2023.05.09.539832.

Additional papers not included in this thesis:

IV. Diamanti, K., **Inda Díaz, J.S.**, Raine, A., Pan, G., Wadelius, C., and Cavalli, M. (2021) Single nucleus transcriptomics data integration recapitulates the major cell types in human liver. *Hepatol Res*, 51: 233– 238.

V. Gustafsson, J., Robinson, J., **Inda-Díaz, J.S.**, Björnson, E., Jörnsten, R, and Nielsen, J. (2020) DSAVE: Detection of misclassified cells in single-cell RNA-Seq data. *PLOS ONE*, 15(12): e0243360.

VI. Gustavsson, M., Käll, S., Svedberg, P., **Inda-Diaz, J.S.**, Molander, S., Coria, J., Backhaus, T., and Kristiansson, E. (2023) Transformers enable accurate prediction of acute and chronic chemical toxicity in aquatic organisms. *Submitted pre–print*, bioRxiv 2023.04.17.537138.

VII. Lund, D., Parras-Moltó, M., Boström, M., **Inda-Díaz**, J.S, Benson. L., Ebmeyer, S., Larsson, D.G.J, Johnning, A., and Kristiansson, E. (2022) Factors influencing the horizontal gene transfer potential of antibiotic resistance genes. *Manuscript*.

Author contributions

I. Participated in the study design and implemented the analysis pipeline. Retrieved and filtered the metagenomic data from MGnify and ENA. Performed the clustering of antibiotic resistance genes, the alignments between fARGene and ResFinder genes, and between genes and metagenomes. Estimated the abundance and diversity of antibiotic resistance genes in all metagenomes, as well as the pan-resistome and core-resistomes for each environment. Carried out the PCA of the abundance and diversity of ARGs. Generated all the figures. Drafted and edited the manuscript.

II. Participated in study design. Constructed the dataset of antibiotic resistance genes. Designed and implemented the data expansion pipeline by retrieving the protein super-families from Interpro, building the hidden Markov models from the data set of antibiotic resistance genes, and running them against the Interpro super-families. Conceptualized the model architecture, and implemented, trained, and tested the transformer model. Generated all the figures. Drafted and edited the manuscript.

III. Participated in the study design, collected and parsed the data from The European Surveillance System (TESSy), and participated in conceptualizing the model architecture. Also implemented, trained, and tested the transformer model. Designed and implemented the uncertainty algorithm. Generated all the figures. Drafted and edited the manuscript.

Acknowledgments

I am deeply grateful to everyone who has made it possible for me to be here today. For their teaching, their company, and trust. I would like to express special thanks to:

Erik Kristiansson, thank you for the opportunity to work together. Your optimism and excitement for science are contagious. I have been fortunate to be mentored by you. Anna Johnning, thank you for being my co-supervisor. It has been wonderful to have your support these years beyond science.

To the present and former members of Erik Kristiansson's group, especially Fanny, Anna R., Astrid, Marcos, David, Mikael, Patrik, and Martin, it has been a real pleasure to share with you this time, the AWs and summer BBQs.

To all the people I have collaborated with at GU, Chalmers, FCC, Sahlgrenska, and Uppsala. Thank you Lennart for your fantastic teaching and amazing YouTube channel. Thank you Klev, for being an inspiration in the last years.

To all the people at Mathematical Sciences for creating such a great working environment, and always making me feel welcome. To Aila S. and Marija C., thank you for all the encouragement I have received from you. To Rebecka J. for welcoming me to Gothenburg. To my present and former colleagues Felix, Gabrijela, Barbara, Helga, Linnea Ö., Malin M., David, Oskar, Olof Z., Olle, and Malin P. To Marie, Lotta, Pernilla, Helene, Ai-Linh, Ulf, and Jovan, for your friendly support with everything related to administration.

To the people who helped me cope with the curricular activities by doing extracurricular things. Chema y Robert, gracias por su cariño y amistad durante estos años. Tharshiny, Alesia T., and Anna L. for your support and friendship. Rafael, Héctor y Philip, gracias por las aventuras en la roca. Mauri, per essere stato il mio mentore in mare come in porto. Tack grabbarna i *Livskvalitet och mera* gruppen.

Endless admiration and love to my family.

Contents

1 Introduction

One of the most important advances in human health was the discovery of antibiotics and their introduction as a treatment for infectious diseases. During the first half of the 20th century, known as the golden age of antibiotics, several chemical classes of antibiotics, with different targets and mechanisms, were discovered and massively produced. These first antibiotics were effectively employed to prevent and cure diseases of epidemic proportion (Mohr, 2016), having a great impact on life expectancy (Hutchings et al., 2019). Many antibiotics are chemical compounds that are naturally secreted by fungi or bacteria, such as penicillin produced by the fungus *Penicillium*. Some antibiotics are synthetically produced (Hutchings et al., 2019), such as nalidixic acid – the first quinolone which was introduced in the 1960s (Andersson and MacGowan, 2003). The widespread use of antibiotics since their introduction has been accompanied by the rapid adaptation of bacteria, developing and acquiring a diverse set of molecular mechanisms to survive in the presence of antibiotics. In 2019 alone, an alarming 1.9 million deaths were attributed to infections caused by antibiotic-resistant bacteria (Murray et al., 2022). The work presented in this thesis represents an effort to address the rising challenge of antibiotic resistance, providing new methods to surveil the upsurge of new mechanisms of antibiotic resistance as well as to provide diagnostics tools for the effective treatment of bacterial infections.

Antibiotic resistance is a bacterial trait primarily attributed to the acquisition of antibiotic resistance genes (ARGs) or mutations in existing genes. While ARGs naturally occur in microbial communities, the excessive use of antibiotics by humans exerts selective pressure that promote ARGs in bacteria. This phenomenon occurs in both external environmental and host-associated bacterial communities, e.g. in human and animal gut. The transfer of ARGs from commensal and environmental bacteria to pathogens is facilitated by horizontal gene transfer (HGT), which involves the uptake of genetic material from the surrounding environment, including the direct exchange of DNA molecules

between bacterial cells. Typically, ARGs are identified after they have become widespread and are detected in pathogens, by which point they have already become a clinical concern. Hence, it is imperative to develop methods enabling the identification of both established and novel ARGs in bacterial communities and to establish transmission pathways between different environments and into pathogens.

In a clinical setting, the presence of antibiotic-resistant bacteria has introduced significant challenges in the treatment and prevention of infections. From routine medical procedures to major surgical interventions, healthcare professionals and patients are now facing a pressing issue: how to prevent and effectively treat bacterial infections when antibiotics have lost their efficacy. Additionally, the successful recovery from a bacterial infection depends on the timely prescription of an antibiotic that is effective against the infecting bacteria. Any delay in treatment can have detrimental consequences on the morbidity and mortality associated with the infection (Friedman et al., 2016). Various practices have been implemented to efficiently treat bacterial infections. Among these practices is the assessment of the resistance profile of the isolated bacteria using antibiotic susceptibility testing (AST). Bacteria susceptibility to a specific drug is assessed with AST by determining whether the drug can inhibit bacterial growth, indicating susceptibility, or if the bacteria can continue to grow in its presence, indicating resistance. However, current conventional AST techniques are time-consuming and since they rely on the growth rate of the bacteria. Consequently, in cases where a suitable antibiotic cannot be promptly identified, the initiation of an effective treatment may be significantly delayed.

In life-threatening situations, the choice of antibiotic treatment often relies on educated guesses based on limited diagnostic information (Bassetti et al., 2020), which can lead to elevated risks for patients and the overprescription of antibiotics (Battle et al., 2016; Kumar et al., 2009; van den Bosch et al., 2017). Therefore, to meet the increasing prevalence of antibiotic-resistant bacteria, it is critical to develop rapid, personalized, and precise diagnostic tools that integrate patient and bacterial data, providing physicians with comprehensive diagnostic information at an earlier stage of the treatment of the infection.

Artificial Intelligence (AI) has recently captured widespread attention across society. The concept of machines equipped with cognitive functions, enabling them to learn and address challenges through experience and new information, has materialized in a diverse range of applications. From robots and cyborgs to autonomous vehicles and chatbots, AI technologies are increasingly integrated into our everyday lives. The possibilities for AI applications in the healthcare sector are vast, including surgical robots, image analysis for the early detection of cancer, and the utilization of electronic health records to enhance clinical

decision-making (Davenport and Kalakota, 2019). AI undeniably holds the promise to transform life sciences and personalized medicine, particularly in the context of highly complex problems such as antibiotic resistance, where extensive datasets of patients and bacteria are at our disposal. In this thesis, an extensive data-driven analysis of ARGs in bacterial communities as well as new AI methods for the identification of novel ARGs and to facilitate decision-making in antibiotic treatments are presented.

Aims

The overall aim of this thesis is to advance the field of AI by developing methods that enhance the surveillance and diagnostics of antibiotic-resistant bacteria. To prevent the proliferation of antibiotic resistance, it is essential to identify novel ARGs before they spread globally and become a clinical problem. This process could also help us understand the underlying selection pressures that drive the transfer of these genes between different bacterial communities and into pathogens. It is also crucial to have diagnostic methods that can accurately detect antibiotic resistance in infecting bacteria. This would not only aid in reducing patient morbidity and mortality but would also contribute to mitigating the significant societal costs associated with antibiotic resistance. The specific aims of this thesis are:

1. to create a data-driven approach for identifying and estimating the abundance and diversity of ARGs in bacterial communities, including ARGs that are presently absent from specialized antibiotic resistance databases (papers I and II),
   (a) to provide a more comprehensive view of the resistome of bacterial communities, comprising both established and latent ARGs (paper I).
   (b) to develop and evaluate an AI method for the detection of novel and uncharacterized ARGs (paper II).
2. to develop and evaluate an AI method for making personalized predictions of antibiotic susceptibility test results based on incomplete diagnostic data (paper III).

Outline of the thesis

Chapter 2 delves into an in-depth exploration of antibiotic resistance, covering pertinent biological concepts, specialized databases, and methodologies for the identification of ARGs in bacterial communities. Chapter 3 offers an introduction to transformers, the backbone behind the AI-based models outlined in this work, and explores conformal prediction, the algorithm utilized in this thesis for managing uncertainty and generating confidence-based predictions. Chapter 4 offers a summary of the findings presented in each paper. It underlines the importance of investigating both established and latent ARGs and demonstrates how AI techniques effectively detect new resistance genes enhancing diagnostic and clinical decision-making capabilities. Chapter 5 contextualizes the results of the thesis within the broader landscape of antibiotic resistance.

2 Genomic identification of antibiotic resistance

The development of artificial intelligence methods and their evaluation depends greatly on the availability of data. The first aims of this thesis relate to the identification and analysis of antibiotic resistance genes (ARGs) in bacterial communities. In this chapter, the existing tools for ARG identification are introduced, underscoring the necessity for innovative methods.

2.1 Antibiotic resistance genes

Bacteria typically acquire resistance through either mutations in their chromosomal DNA or the uptake of external DNA molecules. In some cases, even a single nucleotide alteration in specific chromosomal genes can confer resistance. For instance, in *Escherichia coli*, a single-point mutation in the *GyrA* gene can modify the target of quinolones within the DNA gyrase enzyme, resulting in quinolone resistance (Jaktaji and Mohiti, 2010). Bacteria can acquire external ARGs through the process of horizontal gene transfer (HGT) (Blair et al., 2015). Horizontal gene transfer refers to the lateral transfer of genetic material between organisms not related to direct reproduction, i.e. vertical gene transfer (Burmeister, 2015), Figure 2.1. Antibiotic resistance genes provide the bacteria with the mechanisms to break down, eject, or modify antibiotics, or alter the drug's target within a cell, making the antibiotics less effective against the host (Peterson and Kaur, 2018).

While ARGs naturally exist in microbial communities, the excessive use of antibiotics by humans has imposed a selection pressure that promotes their transfer into pathogens. This transfer occurs in both external and host-associated bacterial communities. In pathogenic bacteria, ARGs are commonly located

on mobile genetic elements (MGEs), including transposons and conjugative elements such as plasmids (Rodríguez-Beltrán et al., 2021). Mobile genetic elements not only facilitate horizontal gene transfer but can also carry multiple ARGs. The effective accumulation of ARGs within a single MGE was first documented in Japan in 1958, where a plasmid in *Shigella* was found to carry ARGs against streptomycin, tetracycline, chloramphenicol, and sulfonamides (Farrar and Eidson, 1971). Consequently, bacteria can develop resistance to multiple antibiotics either by accumulating various ARGs or by having multidrug ARGs (Nikaido, 2009). This accumulation can occur gradually or through the acquisition of MGEs carrying multiple co-located ARGs. Therefore, obtaining a single MGE can trigger multidrug resistance in bacteria (Paulsen et al., 1996; Botts et al., 2017).

The origins of most well-characterized and clinically relevant ARGs identified to date remain unknown, with only a small fraction of them being linked to a pathogenic or commensal host (Ebmeyer et al., 2021). This suggests that many ARGs are likely to have originated from environmental bacteria since they are underrepresented in sequence data repositories (Allen et al., 2010; Bengtsson-Palme et al., 2017a). Consequently, the lack of knowledge about the origins of ARGs poses a challenge for implementing strategies to prevent the transfer of these genes from environmental bacteria to pathogens.

To evaluate the threat that ARGs pose to humans, it is essential to investigate where these genes are located and how they are transferred to pathogens. It's important to acknowledge the difficulty, if not impossibility, of compiling a comprehensive list of all ARGs found in nature. This is due to the vast biodiversity, with over 10^{12} estimated microbial species on earth (Locey and Lennon, 2016), compared to the approximately 416,924 that have undergone whole-genome sequencing (Mukherjee et al., 2023). Furthermore, conducting functional studies on all known bacterial sequences is currently infeasible, and new ARGs can emerge at any time. The majority of ARGs are primarily found in environmental bacteria, displaying high abundance and diversity (Forsberg et al., 2012). However, the massive production, consumption, and disposal of human-made antibiotics have led to an increased presence of ARGs in various bacterial communities exposed to these antibiotics, including human and livestock microbiomes and wastewater (Bengtsson-Palme et al., 2014, 2016; Pal et al., 2016). To prevent the flow of ARGs to pathogens, it is imperative to understand the mechanisms facilitating ARG mobilization and the consequences of human antibiotic usage. Therefore, it is crucial to characterize the composition, encompassing diversity, abundance, and prevalence of ARGs within bacterial communities, also known as the resistome. Several characteristics of microbial communities – such as the extensive latent ARG repertoire, strong selection pressures exerted on them, the emergence of novel ARGs, and the potential

mobilization and dissemination of ARGs through HGT – pose significant risks that can lead to the acquisition of new ARGs by pathogens, complicating the treatment of infectious diseases (Pal et al., 2016). The transfer of ARGs from environmental bacteria to pathogens has already been documented (Forsberg et al., 2012). However, our knowledge regarding the presence of novel ARGs within MGEs and their spread across various microbial communities, including the human microbiome and wastewater, is currently limited. Therefore, efficient surveillance programs and tools are crucial to enable the prompt and accurate identification of both established and novel ARGs, ideally before they become a clinical problem.

In paper I, an extensive analysis of the abundance and diversity of both established and previously uncharacterized ARGs within host-associated microbiomes and bacterial communities in external environments is presented. The terms pan–resistome and core–resistome are introduced, and the potential selection pressures that may explain their patterns in nature are discussed. In paper II, an artificial intelligence method developed for the identification of short fragments of ARGs is introduced and evaluated.

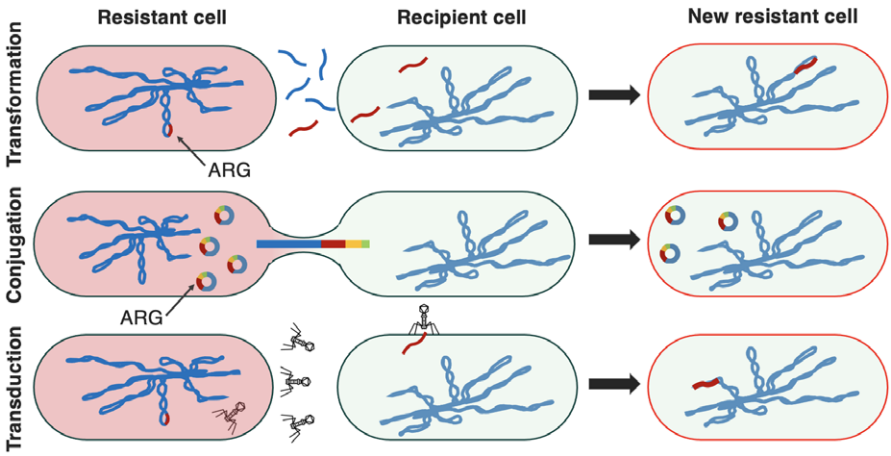

Figure 2.1: Three possible forms of horizontal gene transfer. In transformation, the free DNA material from a resistant cell is taken up by a recipient cell, and the inclusion of foreign genetic material containing ARGs into the genome of the recipient cell can confer resistance to the new host. During conjugation, the donor and recipient cells come into physical contact, facilitating the transfer of a plasmid, which may contain one or several antibiotic resistance genes, from the donor to the recipient. In transduction, a bacteriophage moves genetic material from the donor cell to the recipient cell.

2.2 Diagnostics

Multidrug-resistant bacteria often carry MGEs that harbor several ARGs (Dolejska and Papagiannitsis, 2018; Johnning et al., 2018). As a result, resistance mechanisms can be collectively transferred between bacteria via HGT, leading to highly correlated patterns in resistance profiles across bacterial populations. Moreover, the resistance profiles of infecting bacteria are influenced by patient demographics, including age and sex, as well as geographic locations and travel patterns. These factors can expose patients to specific infections, a dimension that current antibiotic susceptibility testing (AST) techniques often overlook (Dias et al., 2022; Murray et al., 2015; Yelin et al., 2019).

Therefore, innovative solutions that incorporate both AST results and patient information hold the potential to provide accurate predictions of susceptibility profiles, offering physicians valuable diagnostic information during early stages of infections. In paper III, an artificial intelligence method designed to predict antibiotic susceptibility test results for 16 different antibiotics is presented. The method exploits strong resistance patterns across bacteria for these antibiotics, and utilizes incomplete diagnostics data and patient information to predict missing susceptibility tests results.

2.3 Genomics

Life on earth depends on fundamental molecules that store and carry the information needed by all living organisms to grow and function. Deoxyribonucleic acid (DNA), and ribonucleic acid (RNA) are molecules formed by sequences of nucleotides, a sugar molecule attached to a phosphate group and a nitrogen-containing base. DNA is formed by the deoxyribose sugar, and the four bases adenine (A), cytosine (C), guanine (G), and thymine (T). In RNA, the sugar is ribose, and the bases are similar to the ones in DNA except that uracil (U) takes the place of thymine. RNA can be formed by a single sequence of nucleotides, while DNA is composed of two complementary sequences or strands of nucleotides connected by chemical bonds between the bases: adenines bond with thymines, and cytosines bond with guanines. All the DNA material found in an organism forms its genome, and unicellular organisms as well as each cell in multicellular organisms store copies of their genome. Genes, in turn, are precise and ordered sequences of nucleotides within genomes that determine traits in individuals and encode the information needed to produce functional RNA molecules and proteins.

It is fascinating that life and organisms can be determined and represented, to some extent, by sequences of four letters: A, C, G, and T. Sequences consisting of as few as 159,662 of these letters can form a living organism such as the bacterium *Carsonella ruddii* (Nakabachi et al., 2006), while the bacterium *Escherichia coli*, is defined by approximately 5.6 million nucleotides. Humans in turn, have a genome size of approximately 3,298 million nucleotides encoding for about 20,000 proteins, while other organisms can have extreme genome sizes such as the amphibian axolotl and the sugar pine with genomes 8 times larger than humans (NCBI, 2023). The determination of the exact order, *the sequence*, of nucleotides in an organism is done through DNA sequencing. The methodology and technology for DNA sequencing have evolved from the so-called *first-generation sequencing* introduced in the 1950s which produced precise sequences of less than 1,000 nucleotides in length, to high-throughput and parallelized *second–generation* technologies, also known as *next generation sequencing* (NGS), which was introduced in the 2000s. Third–generation technologies became available in the 2010s and enabled the sequencing of full single molecules of DNA witout any pre-amplification (Heather and Chain, 2016; Barba et al., 2014). The dropping cost and higher accessibility to sequencing technologies have allowed us to collect approximately 487,000 fully sequenced organisms, from which 85% represent bacterial genomes (Mukherjee et al., 2023). It is through the sequencing of genes in antibiotic-resistant bacteria, that many ARGs have been initially identified. Furthermore, the amount and speed of sequencing data being produced nowadays represent an invaluable source of knowledge for the study and prevention of antibiotic resistance.

2.4 Metagenomics

Metagenomics is the characterization of all genetic material present in a bacterial community and it is used to identify the taxonomic and functional composition of microbial communities (Fricke et al., 2011). Metagenomics enables the direct study of bacterial populations at their source, eliminating the need to isolate individual bacteria or to culture. It allows for the simultaneous sampling and sequencing from potentially all the genetic material present in the environment (Schloss and Handelsman, 2005).

Early metagenomic techniques consisted of gene cloning. In gene cloning, DNA fragments from environmental microbial communities are extracted and inserted into host bacteria. Subsequently, the clones are subjected to screening to identify specific traits. The selected clones are then sequenced to elucidate the genetic content responsible for the particular trait of interest (Zhang et al., 2021). With the introduction of next generation sequencing technologies, meta-

genomics has transitioned from clone screening and sequencing of target traits to shotgun metagenomics. This approach entails high-throughput sequencing of fragments sampled from the entire genetic material of bacterial communities, Figure 2.2.

Figure 2.2: In shot-gun metagenomics, all DNA is extracted directly from a microbial sample taken from an environment, e.g. human gut. Next generation sequencing (NGS) can then be used to sequence the genetic material. The result of sequencing is short metagenomic reads that, through *de novo* assembly, can be used, for example, to reconstruct full genomes of the bacteria present in the microbial sample.

In recent years, the accessibility and cost-effectiveness of NGS technology have led to a substantial increase in data generation (Keegan et al., 2016). Metagenomic data produced through NGS methods typically comprises short DNA sequences, referred to as reads, which represent highly fragmented DNA molecules. Consequently, individual genes are often represented by multiple reads, including ARGs. As a result, the primary challenge has shifted from data generation to the development of efficient data storage and access mechanisms, along with the creation of computational and analytical tools for data exploration. For reference, the initial dataset for paper I in this thesis consisted of a vast dataset comprising 150 terabytes of data, which included 22,272 metagenomes and 4×10^{11} short sequence reads. The complexity of NGS metagenomic data is not limited to storage requirements, but also applies to its high-dimensionality, sparsity, and significant technical and biological variability. Nonetheless, a diverse range of tools and techniques have been developed to handle genomic and metagenomic data effectively, including sequence alignment software, normalization methodologies, and a combination of unsupervised and supervised methods that together provide reliable and statistically valid results (Bengtsson-Palme et al., 2017b).

2.5 Identification of antibiotic resistance genes in metagenomes

Data-driven identification of ARGs in metagenomic data primarily relies on alignment methods. In sequence alignment, two DNA or protein sequences are compared to each other and measures of similarity between these sequences are obtained. Significant similarity between sequences can imply homology, indicating shared ancestry and a common function. Two widely used alignment tools are BLAST (Altschul et al., 1990) and DIAMOND (Buchfink et al., 2015). Certain alignment methods utilize Hidden Markov Models (HMMs), which incorporate position-specific weights, rendering them more suitable for identifying distant homologs to known genes (Eddy, 2011).

To identify ARGs in metagenomic data through alignment tools, a comparison is made between sequences in metagenomes and known ARG sequences. Public ARG databases, including ResFinder (Zankari et al., 2012) and CARD (McArthur et al., 2013), contain manually curated collections of ARGs, often providing information about their associated phenotypes. These various databases have distinct criteria for the inclusion of ARGs, and their contents mostly, but not completely, coincide.

The alignment of metagenomes against ARG databases can be implemented either directly on the short NGS sequences or on longer, complete genes or genomes formed through a process known as *de novo* assembly. Creating full-sized genes and longer contiguous genomic regions from short reads through *de novo* assembly is a computationally complex process. Although *de novo* assembly offers the possibility to compare full-sized genes to reference ARG databases, it can be especially challenging for resistant bacteria due to the often highly repetitive nature of the genetic context of ARGs (Brown et al., 2021; Bengtsson-Palme et al., 2015).

To circumvent the need for assembly, several tools have been developed for the direct detection of ARGs from short reads. These tools include ARM++ (Bonin et al., 2023), ARGs-OAP (Yin et al., 2022), deepARG (Arango-Argoty et al., 2018), and fARGene (Berglund et al., 2019). Among these, deepARG and fARGene are specifically designed to identify novel ARGs. deepARG combines sequence alignment with neural networks. The alignment step serves as a filter, with only highly similar reads to the reference ARGs being subjected to the neural network analysis. The task of the neural network is to distinguish between reads originating from ARGs and those from other bacterial genes. On the other hand, fARGene employs a probabilistic approach based on HMMs, with each HMM optimized for a specific class of ARGs. Additionally, fARGene

can reconstruct full-length genes using the identified sequence reads.

Existing approaches for identifying ARGs from short-read sequences rely on sequence alignment, including deepARG and fARGene. These methods are limited to identifying new ARGs that exhibit sufficient sequence similarity to known genes while overlooking structural similarities between them, similarities that could be used to identify novel ARGs (Berglund et al., 2021; Ruppé et al., 2019). Consequently, there is a need for tools working directly on short metagenomic reads that incorporate analysis beyond mere sequence alignment to enhance ARG identification.

3 Transformers and Conformal Prediction

Machine learning (ML) and Artificial Intelligence (AI) are disciplines within computer science devoted to developing mathematical models to help a computer, based on experience rather than with direct instruction, continuously learn on its own. The recent technological advances in computational power, the explosion of data collection, and the development of new methods have made ML and AI two extremely popular tools with a wide range of applications. Advances in infrastructure and hardware together with the availability of large and dedicated datasets, have made it possible to implement and train sophisticated AI-based models (Lai et al., 2018). The use of labeled data during the learning process is referred to as supervised learning, while the process of learning hidden patterns in data without any labeled data is known as unsupervised learning. Within supervised learning, classification is the process of assigning observed events to a finite set of categories. The binary or multi-class categorization of the output of an ML classification model is most often presented as a single-label prediction, which is rarely accompanied by measures of predictive uncertainty. The background for the AI-based models presented in this thesis, transformers, and uncertainty control, are presented in this chapter.

3.1 Transformers

In the last years, Natural Language Processing (NLP), a field in linguistics and computer science, has, in combination with AI, reached significant milestones in human language tasks, such as text and speech translation and context-dependent text generation (Hirschberg and Manning, 2015). In 2017, transformers, a NLP architecture, was introduced and, since then, they have revolutionized AI. Transformers are based on self-attention, a mechanism

that exploits the association between all different input positions of a single sequence to compute a new representation of the sequence (Vaswani et al., 2017). The applicability of transformers is large within NLP, and they are used for, e.g. language translation (Vaswani et al., 2017), question answering and language inference (Devlin et al., 2018), and text generation (Brown et al., 2020). Transformers have also been applied in computer vision, including image recognition (Dosovitskiy et al., 2020), object detection (Carion et al., 2020), image segmentation (Ye et al., 2019), and video understanding (Girdhar et al., 2019). In biology, AlphaFold (Jumper et al., 2021), a breakthrough transformer–based model predicting 3D structures of proteins from amino acid sequences, has become the reference point in structural bioinformatics.

Transformers operate on sequential data, often structured into sentences formed by words or "tokens" which are subjected to self–attention. Transformers convert a sentence through a context-aware weighted average that exploits dependencies between words in the input sentence (Devlin et al., 2018). Originally, transformer models were intended for language translation and the first model was, therefore, composed of a transformer encoder, for the input of the original language, and a transformer decoder, for the output in the target language. Today, however, many applications of transformers depend mainly on the encoder part, commonly embedded in larger models developed for, among others, classification, Figure 3.1.

The transformer encoder

A transformer encoder takes as input a sequence of length n of words (or "tokens"), $X = \{x_1, x_2, \ldots, x_n\}$, which can have a free placing or a positional order. Next, each of the n elements of the input sequence is represented in a one-hot encoding matrix defined on $\mathbb{N}^{N\times n}$, n column vectors of length N, where only one entry in each vector is 1 and the rest are 0. Each column represents a word in the sentence and N represents the number of words in the vocabulary, i.e. the number of all accepted values of x_i. The position of the value 1 in the vector i represents the position of the word x_i in the vocabulary. The one-hot encoding matrix for the input sequence, X^*, can be expressed as follows,

$$X^* = \begin{bmatrix} 1 & 0 & & 0 \\ 0 & \vdots & & 1 \\ 0 & 0 & \ldots & 0 \\ \vdots & 1 & & \vdots \\ 0 & 0 & & 0 \end{bmatrix}_{N\times n} .$$

In the transformer architecture, the encoding matrix X^* undergoes a word embedding, i.e. a linear transformation $h_E(W_E, X^*) = W_E X^*$ is applied to the one-hot encoding vectors using a matrix of weights $W^E \in \mathbb{R}^{d \times N}$. The output of the word embedding is the matrix $X_E \in \mathbb{R}^{d \times n}$, where each of the n columns represents a word in the input sequence X, undergoing a dimension reduction (or expansion) from N elements in the vocabulary to the pre-defined size d.

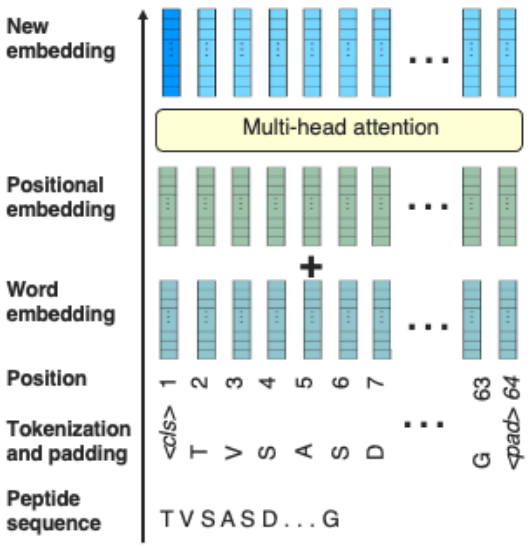

Figure 3.1: A transformer encoder. The input peptide sequence is split into tokens (amino acids), complemented with a *cls* token at the start, and padded to a specific length with *pad* tokens. Thereafter, the sequence goes through a word embedding and a positional embedding. The embeddings are summed together and passed through a multi-head attention layer. The output of the multi-head attention layer is a context-aware weighted average representation of the peptide sequence.

The output of the word embedding passes through a self–attention mechanism, defined by the function $h_A = (X_E, W_Q, W_K, W_V)$. The output of the self-attention is a new representation of the input sequence calculated by context-aware weighted averages, where the weights depend on the relationship between words in the language as a whole. The weights for the weighted average in the self-attention mechanism are calculated based on three linear transformations on the word embedding X_E. The first two linear transformations are the matrices $Q = W_Q X_E$ and $K = W_K X_E$, followed by a softmax transformation and re-scaling of the product between Q and K. The matrix Q contains n word-specific (query) vectors of size D. The matrix K contains n (key) vectors and all of them are used to take the inner product with each of

the query vectors so that the (weight) vectors in $K^T Q$ are also word-specific. The output of the self–attention mechanism, X_A, is the weighted average of the third linear transformation on X_E, the matrix $V = W_V X_E$. The output of the attention layer, X_A, is, thus, obtained as followed,

$$X_A = W_V X_E \, h_S \left(\frac{1}{\sqrt{d}} \left(W_K X_E \right)^T W_Q X_E \right),$$

where the h_S is the softmax transformation, taken over the column vectors of a matrix $Z = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{d \times n}$, this is, for element w_i, $w_i = \frac{e^{z_i}}{\sum_{i=1}^{d} e^{z_i}}$, $i = 1, 2, \ldots n$.

The attention mechanism can be expressed in the commonly used alternative form:

$$X_A = VW,$$

where,

$$\begin{aligned} Q &= W_Q X_E, \\ K &= W_K X_E, \\ V &= W_V X_E, \\ Z &= \frac{1}{\sqrt{d}} K^T Q \text{ and,} \\ W &= \text{softmax}\left(Z\right). \end{aligned}$$

The output of the self–attention mechanism, $X_A = [x_{A,1}, x_{A,2}, \ldots, x_{A,n}]$, conserves the same dimension as the word embedding matrix X_E ($d \times n$). Thus, each of the column vectors $x_{A,i}$ in X_A is a new representation of the corresponding word embedding x_i in X_E that considers all the words in the input sequence and dependencies between all the words in the vocabulary. The parameter matrices W_Q and W_K are both defined on $\mathbb{R}^{D \times d}$ and will be squared matrices if $D = d$, while the matrix W_V is defined on $\mathbb{R}^{d \times d}$. The scaling factor $\sqrt{d}$ is introduced to prevent vanishing gradients (Lin et al., 2022).

The output of the attention layer, X_A, undergoes an add & normalize layer (A&N), defined by the function $h_{A\&N}\left(X, Y\right)$, followed by a position-wise feed-forward network (PFFN), defined by the function $h_{FN}\left(X\right)$, and, then, a second A&N layer. The A&N function $h_{A\&N}$ takes as input matrices X and Y, both

defined on $\mathbb{R}^{d\times n}$, and normalizes the sum of the two matrices as follows,

$$h_{A\&N}(X,Y) = \frac{(X+Y)-\mu}{\sqrt{\sigma^2}},$$

where

$$\mu = \frac{1}{dn}\sum_{i=1}^{d}\sum_{j=1}^{n}(X+Y)_{ij}\text{, and}$$

$$\sigma^2 = \frac{1}{dn}\sum_{i=1}^{d}\sum_{j=1}^{n}\left((X+Y)_{ij}-\mu\right)^2.$$

The PFFN consists of applying one neural network to each of the columns of the input matrix, e.g. two linear transformations with the Rectified Linear Unit (ReLU) activation function in between. The layers of the neural network are linear transformations and together with the ReLU activation function, thus, become,

$$h_{FN}(X, W_{f1}, W_{f2}, b_{f1}, b_{f2}) = W_{f2}h_R(W_{f1}X + b_{f1}) + b_{f2},$$

where h_R is the ReLU activation, $h = \max(x, 0)$, W_{f1} and W_{f2} are weight matrices defined on $\mathbb{R}^{d^*\times d}$ and $\mathbb{R}^{d\times d^*}$, respectively, and b_{f1} and b_{f2} are bias vectors defined on $\mathbb{R}^d$.

Thus, the output of the attention layer X_A and the word embeddings X_E are fed to the (A&N) layers and the PFFN as follows,

$$X_O = h_{A\&N}\left(h_{FN}\left(h_{A\&N}(X_A, X_E), W_{f1}, W_{f2}, b_{f1}, b_{f2}\right), X_A, X_E\right),$$

resulting in the final output of the transformer, the matrix $X_O \in \mathbb{R}^{d\times n}$, having n columns, one per each word in the input of the sequence X. Each of the column vectors in X_O is a context-and-language-aware representation $\left(\in \mathbb{R}^d\right)$ of the corresponding word in the input sequence X. The added part of the (A&N) layers represents a residual connection (Lin et al., 2022). The normalization reduces the covariance shift and centers the embeddings around zero. The PFFN allows the network to generalize to sequences of varying lengths.

Positional-variant encoder

The attention mechanism is positional-invariant, which means that the order of the elements in the input sequence does not affect the output. In situations

where the order of the elements contains important information, a positional embedding can be implemented as a linear transformation of the one-hot encoding system. The positional embedding X_P of the ordered sequence $P = 1, 2, 3, \ldots, n$, represents the positions of the words in the input sequence X. This is, $X_P = W_P P^*$, where P^* is the one-hot encoding of the ordered sequence P. Having both word and positional embeddings, the input to the self–attention mechanism simply is the sum of both embeddings, $X_I = X_E + X_P$.

Multi-head and multi-layer attention

The transformer encoder can also consist of a multi-head attention layer consisting of m parallel self–attention layers, called heads, which are then merged into a single output. If $X_E \in \mathbb{R}^{d\times n}$ is the input to the multi-head attention layer having m heads, each with an output $X_{A_j} \in \mathbb{R}^{d\times n}$, $j = 1, 2, \ldots, m$, the output of the heads are stacked into the matrix X_H row-wise, so the matrix $X_{A_{j-1}}$ is expanded with all the rows of the matrix X_{A_j}. Then, a linear transformation with the weight matrix $W_H \in \mathbb{R}^{d\times dm}$ is applied to X_H. The result is the matrix $X_A \in \mathbb{R}^{d\times n}$ of equal dimension as the input X_I, which is subsequently fed to the (A&N), PFFN and (A&N) layers,

$$X_A = W_H X_H = W_H \begin{bmatrix} X_{A_1} \\ X_{A_2} \\ \vdots \\ X_{A_m} \end{bmatrix},$$

and each of the self–attention heads has its own set of weight matrices W_{Q_j}, W_{K_j}, and W_{V_j}.

Similarly, since the input and the output of a self–attention (or multi-head attention) layer, with the correspondent (A&N), PFFN, and (A&N) layers, share the same dimension, several l– attention mechanisms can be stacked one after the other forming a multi-layer attention. In a multi-layer attention mechanism, the output of the first attention layer serves as input to the second one, the output of the second layer serves as input to the third one, and so forth, until the output of the $l - 1^{th}$ layer serves as input to the l^{th} one. Each layer has its own weights for the attention part and the PFFN. Several heads and layers in the attention mechanism provide the characteristic of deep learning to transformers. Each of the heads and layers allows the models to learn different characteristics and dependencies of the data, allegedly.

Pre-training of transformers

Most often, transformers undergo a pre-training step before being trained for classification. If the transformer is used in tasks requiring supervised learning, such as classification, pre-training enables an initial parameter estimation using unlabeled data, a process sometimes referred to as self-supervised learning. During pre-training, the model learns the grammar and the semantics of the data. Masked language models (MLM) and next sentence predictions (Devlin et al., 2018) are examples of strategies employed for pre-training the transformers. In an MLM, a proportion of the tokens in the input sequence are masked, i.e. the token is replaced by *msk*, or changed to a random token. The other tokens are left unchanged. The objective of the transformer is to learn to predict the correct value of the tokens that were changed to *msk*. For that, the output of the transformer X_O is passed through yet another feed-forward neural network $h_{FN}\left(X_O, W_{M1}, W_{M2}, b_{M1}, b_{M2}\right)$, with an output sharing the same dimensions as the one–hot encoding X^*.

Let $K \subset 1, 2, \ldots, n$ represent the subset of indexes from the input sequence X that have been randomly masked with the token *msk* with their corresponding one-hot encoding vectors x^*_κ, where $\kappa \in K$. The pre-training using an MLM consists of minimizing the cross-entropy loss between the output of the feed-forward network coming after the transformer, X^κ_M, and the one-hot encoding vectors x^*_κ, where κ in $X_{M,\kappa}$ represents the index of a column vector in X_M and $\kappa \in K$. For a transformer with a single head and attention layer, this is,

$$\underset{\substack{W_E, W_Q, W_K, W_V, \\ W_{f1}, W_{f2}, b_{f1}, b_{f2}, \\ W_{M1}, W_{M2}, b_{M1}, b_{M2}}}{\arg\min} \sum_{\kappa \in K} L_M\left(x^*_\kappa, x^\kappa_M\right),$$

where, as before,

$$\begin{aligned}
&X_M = h_{FN}\left(X_O, W_{M1}, W_{M2}, b_{M1}, b_{M2}\right)\\
&X_O = h_{A\&N}\left(h_{A\&N}\left(X_A, X_E\right), h_{FN}\left(h_{A\&N}\left(X_A, X_E\right), W_{f1}, W_{f2}, b_{f1}, b_{f2}\right)\right),\\
&X_A = h_A\left(X_E, W_Q, W_K, W_V\right),\\
&X_E = h_E\left(W_E, X^*\right).
\end{aligned}$$

L_M is the cross-entropy loss, defined here as,

$$L_M\left(x_\kappa, X^\kappa_M\right) = -\log\left(\frac{e^{x^\kappa_M(i)}}{\sum_{i=1}^N e^{x^\kappa_M(i)}}\right) x^*_k\left(i\right),$$

where $x^\kappa_M\left(i\right)$ and $x^*_k\left(i\right)$ are the i–th elements of the κ–th column vectors of the

matrices X_M and X^*, respectively.

Transformers for classification

The output of transformers can serve as input to classification tasks, which is the main application of transformers in papers II and III of this thesis. Often, the input sequence to a transformer is complemented in the first position by a classification token *cls* before being fed to the encoder. The first column vector $X_{O_{i,1}}$, $i = 1, \ldots, d$, where X_O is the transformer encoder output, is usually used as input in classification models, but other functions of the output can also be applied. In either case, the classification model and the transformer encoder become a single unit that needs to be trained to solve a specific task, a process known as fine-tuning. During fine-tuning, the encoder, through self–attention, learns the semantics, context, and meaning of the sequence concerning the classification task and summarizes it in the first vector of the output.

In classification, the aim is to correctly assign an input sequence into C classes. Let the feed-forward neural network be the function $h_{FN}\left(X_O\left(1\right), W_{C1}, W_{C2}, b_{C1}, b_{C2}\right)$ used as the classification model and assume that the input sequences belongs to the class y. During fine-tuning, we aim to minimize the cross entropy loss between the output $x_C \in \mathbb{R}^C$, of the feed-forward network and y. For a transformer with a single head and attention layer, this is,

$$\underset{\substack{W_E, W_Q, W_K, W_V, \\ W_{f1}, W_{f2}, b_{f1}, b_{f2}, \\ W_{C1}, W_{C2}, b_{C1}, b_{C2}}}{\arg\min} \quad L_C\left(y, x_C\right),$$

where, as before,

$$\begin{aligned}
x_C &= h_{FN}\left(X_O, W_{C1}, W_{C2}, b_{C1}, b_{C2}\right) \\
X_O &= h_{A\&N}\left(h_{A\&N}\left(X_A, X_E\right), h_{FN}\left(h_{A\&N}\left(X_A, X_E\right), W_{f1}, W_{f2}, b_{f1}, b_{f2}\right)\right), \\
X_A &= h_A\left(X_E, W_Q, W_K, W_V\right), \\
X_E &= h_E\left(W_E, X^*\right).
\end{aligned}$$

L_C is the cross-entropy loss, defined here as,

$$L_C\left(y, x_C\right) = -log\left(\frac{e^{x_{C,y}}}{\sum_{i=1}^{C} e^{x_{C,i}}}\right),$$

where $x_{C,i}$ is the i-th element of the vector x_C and $x_{C,y}$ is the element of the vector corresponding to the label y. In this case, all parameters in the transformer and the classification model are trained simultaneously. The transformer, thus, will be adjusted, and fine-tuned, to minimize the loss function that is purely defined based on its classification performance.

3.2 Confidence–based predictions

In ML models, usually, predefined output values from models serve as thresholds for categorizing input data into any of the classes the model can choose from. With this setup, we do not know the certainty of an individual prediction, since we only estimate an average level of confidence for validation datasets using cross-validation. Thus, we only control false positive rates by varying the threshold, which does not provide information on certainty for single predictions. The possibility to assess the confidence of predictions is vital in critical decision-making, especially in healthcare settings where incorrect treatment decisions can directly impact the well-being of the patients. Current AI methods for diagnostics lack uncertainty estimation.

Conformal Prediction (CP) (Vovk et al., 2005) was developed to provide a measure of certainty to predictions done by ML classifiers. Contrary to single-label predictions, conformal prediction returns a prediction set that can contain zero, one, or multiple labels, depending on the certainty. The main characteristic of CP is that the prediction sets are valid, i.e. the true label will, on average, be in the prediction set without exceeding a pre-specified error rate.

Conformity measures

Assume that there is an example space $Z = X \times Y$ in a classification task, defined by the Cartesian product between the feature space X and the label space Y, where any $y \in Y$ can take c different values ($c = |Y|$). For a sequence of observations $Z^l = \{z_1, \ldots, z_l\}$, we define a conformity measure as a function $A : Z^l \times Z \to \mathbb{R}$, such that $A(\zeta, z)$ is independent of the order of the elements in $\zeta \in Z^l$. The conformity measure A is designed to quantify the degree of concordance between an example z and the sequence ζ. Conformity measures can also be defined in terms of the output of ML models. If h is a prediction model and F measures its prediction error, then the conformity measure for the observation x_i to the label $y^* \in Y$ based on the sequence ζ can be defined as $A(\zeta, (x_i, y^*)) = 1 - F(h(x_i, y^*))$ (Linusson et al., 2017).

Neural networks trained for classification, most commonly have as final output a vector v_o of size c. Thus, the softmax transformation on v_o can be used as conformity measure of the feature input x_i to any $y_j \in Y$. This is, $A(\zeta, (x_i, y_j)) = \frac{e^{v_{o,j}}}{\sum_{i=1}^{n} e^{v_{o,j}}}$.

Having a classification model trained with $Z^l = \{z_1, \ldots, z_l\}$ observations and the conformity measure A, the prediction set Γ^ϵ for a testing observation x' with respect to Z^l and A can be defined as,

$$\Gamma_l^\epsilon(z_1, \ldots, z_l, x') = \{y \mid p^y > \epsilon\},$$

where,

$$p^y = \frac{|i \in \{1, \ldots, l\} : \alpha_i \leq \alpha^y| + 1}{l+1}, \quad y \in Y,$$

and,

$$\begin{aligned} \alpha_i &= A((z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_l), z_i), i \in \{1, \ldots, l\} \\ \alpha^y &= A((z_1, \ldots, z_l), (x', y)), \end{aligned}$$

Under the assumption that the examples $\{z_1, \ldots, z_l\}$ are exchangeable, the probability that true label y' of x' is not in Γ^ϵ will not exceed ϵ for any $\epsilon \in [0, 1]$ (Vovk, 2012).

Inductive conformal predictor

Conformal Prediction can be a highly computer-intensive method. Under the setup presented, the ML classification model needs to be re-trained for each element z_i in the training dataset to calculate the conformity measure α_i. The computation time can be reduced by implementing Inductive Conformal Prediction (ICP), (Papadopoulos, 2008). ICP involves splitting the l training observations in Z^l, into two subsets Z^t and Z^c ($l = t + c$, $c > t$). The classification model is then trained using the training subset Z^t, and c conformity measures are calculated for each observation in Z^c based on Z^t,

$$\alpha_i = A((z_1, \ldots, z_{t-1}, z_t), z_i), i \in \{t+1, \ldots, t+c\}.$$

The conformity measures are only computed for the true label y_i for each observation $i \in Z^c$, and not for every possible outcome. Then, for a new

observation x', the prediction set Γ^ϵ is constructed as follow,

$$\Gamma^\epsilon\left(x'\right) = \{y| \quad p^y > \epsilon\},$$

where,

$$p^y = \frac{|i \in \{t+1, \ldots, t+c\} : \alpha_i \leq \alpha^y| + 1}{c+1},$$

and

$$\alpha^y = A\left(\left(z_1, \ldots, z_t\right), \left(x', y\right)\right),$$

Splitting the data into calibration and training entails a reduction in the information efficiency of the prediction, using only a fraction of the data to train the classifier model and another fraction to calculate the non-conformity measures (Linusson et al., 2017). Therefore, this is only possible in situations where there is a large number of observations.

Evaluation of conformal predictors

The assumption that the random examples $Z_1, \ldots, Z_l$ are exchangeable, i.e. any permutation of the examples is equally likely to occur, makes predictors in CP and ICP unconditionally valid (Vovk, 2012). Miscalibration curves on test datasets can be built to empirically evaluate whether prediction regions return error rates that are equal to or less than the expected error rate (ϵ). Although ICPs are unconditionally valid, the variability of their performance is affected by the size of the calibration and testing sets, and the ML algorithm employed (Vovk, 2015).

Conformal predictors can also be evaluated in terms of their efficiency, for example, the size of the prediction set. Ideally, prediction sets should contain one label, nevertheless, to keep the expected error rate ϵ, some prediction sets can be empty or contain two or more labels. The average set size for the test data set is a good estimate of how informative a conformal predictor is. As prediction sets containing one single label are more informative and thus preferred, it is also necessary to evaluate the efficiency of these single sets, i.e. how often the predictors contain one single and correct label. In paper 3 (Inda-Diaz et al., 2023), used the metrics region size and correct singletons defined as follows,

$$\text{region size} = \frac{1}{k} \sum_{i=l+1}^{l+k} |\Gamma_i^\epsilon|$$

$$\text{correct singletons} = \frac{1}{k} \sum_{i=l+1}^{l+k} 1_{\{\Gamma_i^\epsilon = \{y_i\}\}}.$$

where Γ^ϵ is a conformal prediction set. I refer to Vovk (Vovk, 2015) for other performance metrics on conformal predictors.

4 Summary of results

In this chapter, I provide a concise overview of the key discoveries presented in the three papers comprising this thesis. This summary aims to enhance comprehension of their collective impact on the research domain. Paper I focuses on the identification and quantification of established and latent antibiotic resistance genes, their prevalence, and diversity across various environments. Paper II introduces a novel artificial intelligence model designed for the identification of resistance genes within metagenomic data. Lastly, paper III presents a new approach to enhancing clinical diagnostics by predicting the resistance profiles of bacterial isolates, ultimately aimed at more efficient and timely treatment of bacterial infections.

4.1 Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes (paper I)

In this paper, we present an estimation of the abundance and diversity of antibiotic resistance genes in external and host-associated microbial environments. We categorized ARGs into two primary groups: "established" genes, which are well-documented and present in resistance gene databases, and "latent" genes, which are computationally predicted and lack a comprehensive characterization, Figure 4.1 A. These two groups comprised a total of 572 established genes and 23,502 latent genes, spanning across 17 gene classes, including six aminoglycoside, five β-lactam, two macrolides, one quinolone, and three tetracycline resistance gene classes, Figure 4.1 B. Our analysis encompassed a comprehensive examination of 10,744 metagenomic samples across 20 different environmental types.

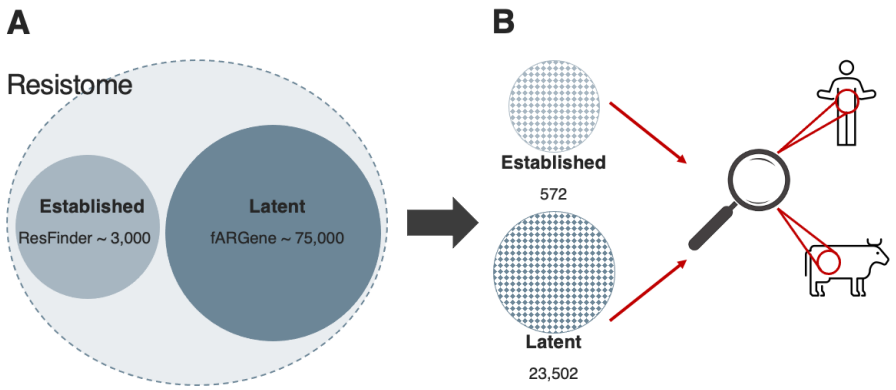

Figure 4.1: A) The total number of ARGs forming the resistome is unknown and presumed to be constantly changing. To gain insights into the resistome, we constructed two distinct groups: the well-defined "established" genes, which already have clinical relevance, and the "latent" genes, which are computationally predicted. Each group was carefully curated to encompass genetically dissimilar ARGs. B) Subsequently, we searched for established and latent ARGs within an extensive metagenomic database spanning diverse environments.

We defined the pan–resistome of an environment as the collection of all ARGs present in any of the metagenomic samples within that environment. In contrast, the core-resistome was defined as the subset of ARGs that are consistently found across a specific environment. Our analysis involved estimating both the pan-resistome and core-resistomes for various environments and outlining the underlying factors that contribute to their composition, Figure 4.2 A. Additionally, we identified numerous latent ARGs that were widespread across different environments and pathogens and underscored their potential adverse implications in clinical settings. This study expanded upon prior research, which predominantly focused on established ARGs, by providing novel insights into the prevalence of latent ARGs in natural settings. Notably, our findings revealed that latent ARGs constitute a substantial portion of the overall resistome. Specifically, the pan-resistomes of both external and host-associated environments primarily consisted of latent genes, accounting for 91% to 98% and 71% to 90%, respectively. On average, the pan-resistomes of external environments exhibited greater diversity and had a larger pool of ARGs compared to host-associated environments, Figure 4.3. Moreover, latent genes comprised a significant proportion (40% to 73%) of the core-resistomes within host-associated environments.

Our results showed that the core–resistomes within the digestive systems of humans and animals, as well as wastewater, were extensive, and shared a

significant number of genes. This similarity in both size and gene composition suggests that these environments experience similar selection pressures that influence individual gene variants, Figure 4.2 B. In contrast, the core–resistome of external environments, including soil, rhizosphere, marine water, lentic water, and freshwater, was significantly smaller than host-associated environments and wastewater. These findings imply that external environments may not exert sufficiently strong selection pressures to favor the fixation of specific gene variants within their bacterial communities.

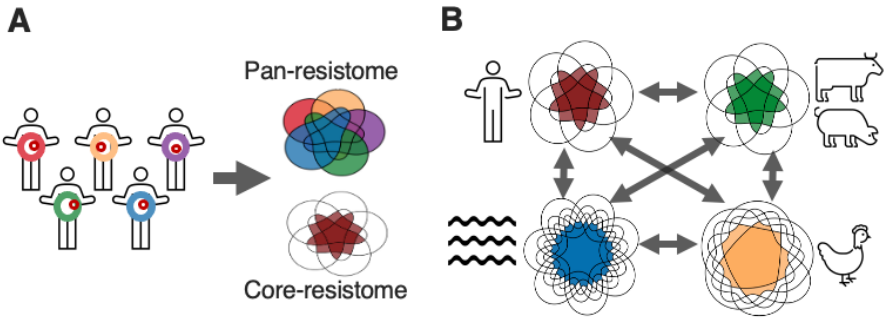

Figure 4.2: A) We defined the pan–resistome of an environment as the ARGs present in any of the metagenomic samples from that environment, and the core–resistome as the ARGs commonly found in that environment (a degree of intersection between the samples in the environment). B) The core–resistomes in human and animal digestive systems and wastewater had a large degree of similarity in size and the genes found within, indicating that these environments have common selection pressures that are acting on individual gene variants.

There were variations in the abundance and diversity of each ARG class across different environments. Additionally, the composition of the pan-resistome significantly differed from that of the core-resistome, suggesting that selection pressures act differently on each ARG class. Interestingly, our analysis indicated that wastewater bacterial communities may act as hotspots for the mobilization of latent ARGs. This assessment is based on two criteria: the presence of a vast and diverse pan-resistome of latent ARGs and a high abundance of established mobile ARGs, which implies the existence of mobile genetic elements necessary for mobilization.

In summary, our results emphasize the presence of latent ARGs in the resistome of both external and host-associated environments. These latent ARGs constituted a majority of the genes in the pan-resistomes, with several of them also making up the core-resistomes of human and animal-associated metagenomes. Consequently, we argue that latent ARGs should be given greater considera-

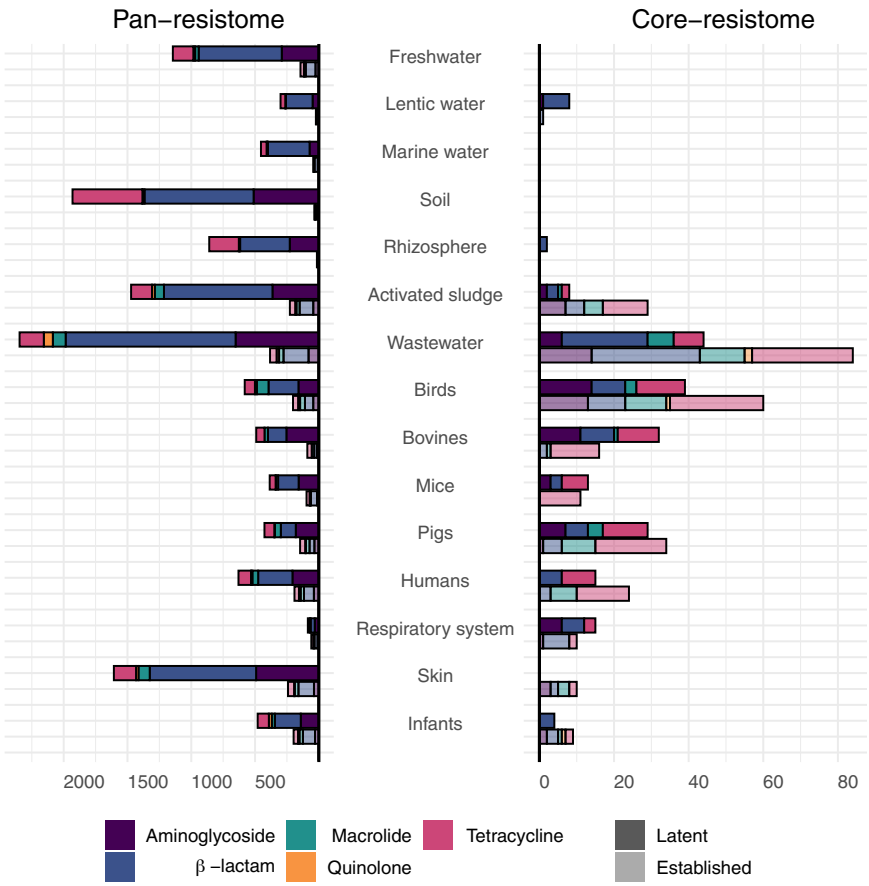

Figure 4.3: Pan-resistomes and core-resistomes. The length of the left and right bars describe the size of the pan- and core-resistome, respectively. The pan-resistome includes all genes encountered in at least one sample from the environment, and the core-resistome includes all genes that were commonly encountered (at least 50% of the samples). The colors indicate antibiotic type with higher opacity for latent ARGs and higher transparency for established ARGs. The computations were done based on rarefied metagenomes that, for each environment, were repeatedly subsampled down to 100 samples. The figure shows the average number of genes over all 100 samples. The labels Birds, Bovines, Mice, Pigs, Humans, and Infants denote metagenomes from the corresponding digestive system. The respiratory system and skin only include human samples.

tion in future resistome studies. Restricting the study of ARGs in microbial communities solely to established genes is incomplete. Future studies should, therefore, also include latent ARGs to gain a more comprehensive understanding of antibiotic resistance and its potential implications on human health.

Main contributions

1. The characterization of the overlooked latent ARGs, revealing their abundance, diversity, and mobility in various bacterial communities as well as their presence in pathogens.
2. The characterization of the pan- and core-resistomes of bacterial communities in host-associated and external environments, composed of established and latent ARGs.
3. The description of wastewater microbiomes as a potentially high-risk environment for the mobilization and promotion of latent ARGs.

4.2 Identification of short fragments of antibiotic-resistance genes using transformers (paper II)

In a world where antibiotic-resistant bacteria continue to pose a significant threat, and given the rich diversity of ARGs within microbial communities, there is a pressing need for methods that can identify novel ARGs in both environmental and pathogenic bacteria. These methods could play a crucial role in surveillance efforts and help mitigate the spread of ARGs to clinical settings. Traditionally, the search for ARGs in bacterial communities relies on alignment-based methods. In this process, the short DNA sequences from metagenomes are compared to reference databases, and only those reference genes with sufficiently high similarity to any metagenomic read are said to be present in the bacterial community from which the metagenomic sample was derived. Alternatively, metagenomes can be assembled into complete genes and genomes, a time and resource-demanding process. The assembled genes and genomes can then be compared to reference datasets with greater confidence. In this paper, we introduce TISARG, a **t**ransformer for the **i**dentification of **s**hort **a**ntibiotic **r**esistance **g**enes fragments. Unlike alignment-based tools, such as hidden Markov models, TISARG leverages deep learning and operates

in an alignment-free manner. TISARG can identify antibiotic resistance genes across 20 distinct gene classes with a global accuracy of 96%.

TISARG operates by taking a short peptide sequence as input, which is then processed through a transformer encoder connected to a deep neural network. This network classifies the peptide into one of 20 common and clinically relevant ARG classes or as "not an ARG" class, Figure 4.4 C. The ARG classes that TISARG has been trained to identify include class A, B1-B2, B3, C, D1, and D2 beta-lactamases; *aph(2")*, *aph(3')*, *aph(6)*, *aac(2')*, two classes of *aac(3)*, and three classes of *aac(6')* aminoglycoside resistance genes; along with *mph* and two classes of *erm* macrolide resistance genes. TISARG's training process involves two main steps. First, a mask language model is employed within the encoder to learn the semantics of a highly diverse range of bacterial proteins. Then, an extended dataset of ARGs is utilized to fine-tune both the encoder and the neural network, enabling them to effectively identify fragments of resistance genes, Figures 4.4 A and B. Given that deep learning models require substantial datasets and the number of experimentally validated ARGs is limited, we expanded the pool of ARGs by searching for highly similar genes within existing repositories using hidden Markov models, Figures 4.4 A and B.

TISARG demonstrated exceptional performance with an average sensitivity as high as 96% across the 20 classes of ARGs. However, it is worth noticing that the sensitivity varied between classes, with the aminoglycoside acetyltransferases (*aac(2')*, *aac(3)*, and *aac(6')*), and the macrolide phosphotransferases *mph* exhibiting the highest sensitivity. In contrast, the aminoglycoside phosphotransferase *aph(2")* showed the lowest sensitivity. Furthermore, the average specificity across ARG classes reached 99.8%

Given that ARGs are fragmented into random lengths within metagenomic data, we conducted an analysis of TISARG's performance with respect to the length of the input peptide sequences. Our findings indicated that TISARG's performance improved as the length of the peptide sequence increased. Nevertheless, even for shorter sequences, TISARG consistently delivered a high performance. TISARG exhibited a median sensitivity of 88% when handling sequences as short as 20 to 25 amino acids. This sensitivity rapidly increased to 93% for sequences ranging from 26 to 31 amino acids and reached 98.4% for sequences spanning 55 to 63 amino acids.

Recognizing the importance of detecting novel resistance forms, particularly given the considerable diversity of ARGs even within the same class, we further investigated TISARG's capability to identify ARGs based on their sequence similarity to previously known resistance genes. TISARG demonstrated a minimum sensitivity of 98% for sequences having a minimum of 71% sequence

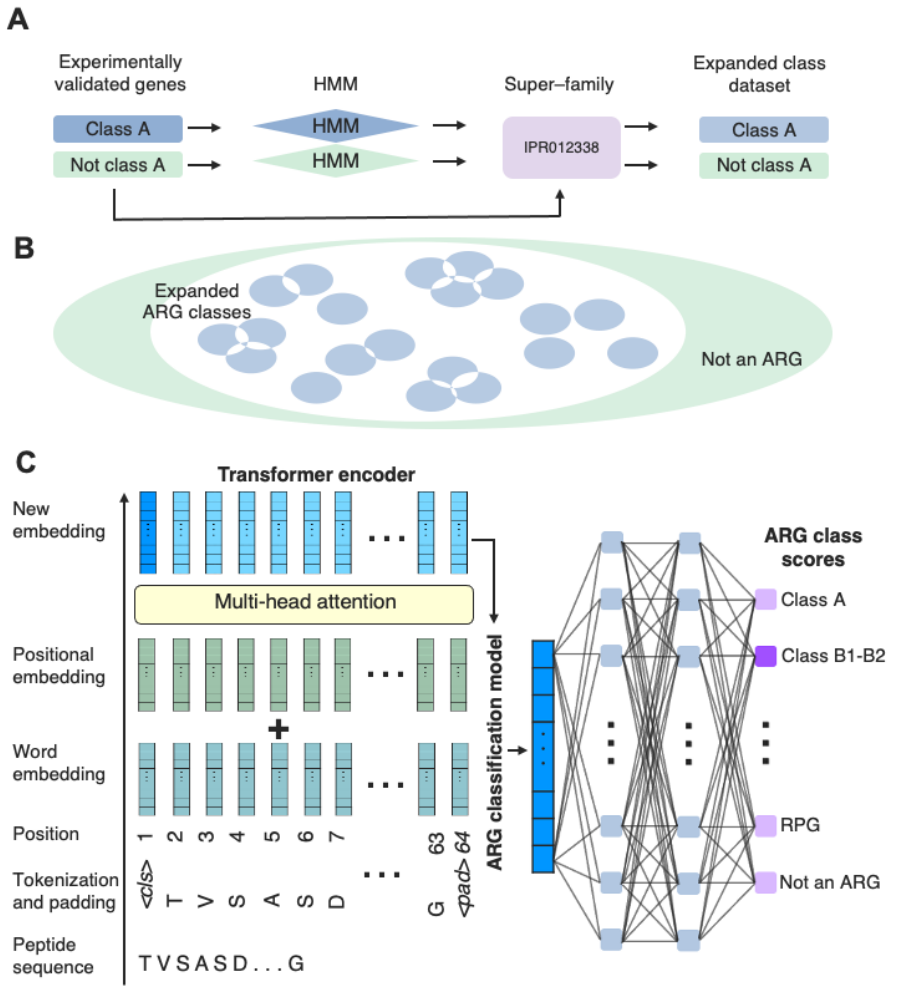

Figure 4.4: A) The principle of data expansion was used to increase the number of labeled sequences for the fine-tuning model. Hidden Markov models were built for each gene class and its corresponding negative gene set. Thereafter, the class-specific positive and negative HMMs were applied to their corresponding Interpro protein family dataset, and proteins from the families that had high domain scores to the class-specific ARG HMM or the negative HMM formed the expanded gene classes. B) All Intepro super-families were clustered and proteins identified by only one ARG HMM were included in their corresponding expanded class dataset, proteins identified by any negative HMM were merged in the "not an ARG" class, and proteins identified by two or more ARG HMMs were disregarded. The expanded ARG class datasets and the "not an ARG" class were used for fine-tuning TISARG. C) an overview of the transformer model. The amino acid sequence is complemented with the *cls* token at the start and padded with the token *pad* to a specified length. The input sequence and the ordered positions undergo word and positional embedding transformations, and the embedded vectors are added and subjected to multi-head attention layers. The first vector of the output of the attention mechanism is fed to a neural network for the multi-class categorization of the input sequence.

identity to previously known genes, along with a specificity ranging from 99.7% to 100%. In cases where sequence identity was low (0% to 30%), TISARG exhibited a sensitivity of 30%. However, this sensitivity substantially improved to 76% and 91% for sequences with similarities of 31% to 50% and 51% to 70%, respectively.

To provide a comparative analysis, we assessed TISARG's performance against established tools in the field, specifically fARGene (Berglund et al., 2019) and deepARG (Arango-Argoty et al., 2018). These tools rely on hidden Markov models and homology-dependent neural networks, respectively. The benchmark revealed TISARG's capability to identify sequences of short length and high dissimilarity to known genes. For instance, when analyzing peptide sequences ranging from 32 to 50 amino acids in length, TISARG achieved a sensitivity of 96%, a notable difference of over 20 percentage points compared to its closest competitor, fARGene. In the case of dissimilar sequences, exhibiting sequence similarities between 31% and 50%, TISARG demonstrated a sensitivity of 79%, nearly 40 percentage points higher than that of fARGene, Figure 4.5.

In summary, our findings highlight the applicability of artificial intelligence techniques, particularly natural language processing tools like transformers, for the annotation of short protein sequences derived from metagenomic data. The introduction of TISARG, the model outlined in this study, marks the pioneering development of a transformer designed exclusively for the identification of antibiotic resistance genes within metagenomic datasets. This advancement contributes significantly to the ongoing battle against antibiotic resistance and holds promise for integration into surveillance initiatives.

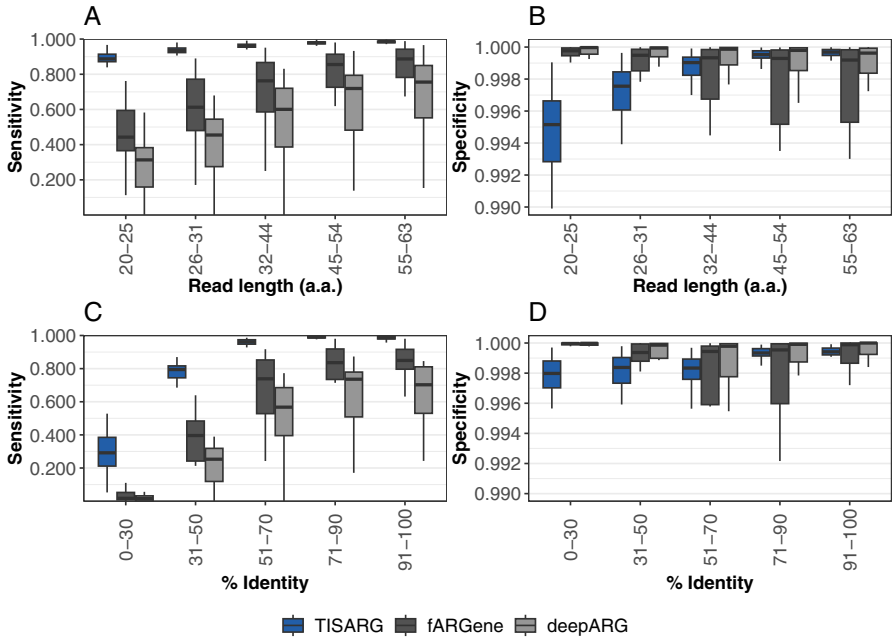

Figure 4.5: The distribution of the sensitivity and specificity of TISARG, fARGene, and deepARG, with respect to the number of amino acids in the input sequence (panel A and B) and the sequence identity between the ARG and resistance genes present in the positive dataset (panel C and D). The color of the boxes represents the different methods. The boxes span from the first to the third quartile of the performance over the ARG classes, respectively, while the black line within the boxes corresponds to the median. The length of the whiskers is set to 1.5 times the interquartile range. In (A) and (B), the peptide sequences were grouped according to five different input length intervals, 25 to 31, 26 to 31, 32 to 44, 45 to 54, and 55 to 63 amino acids. In (C) and (D), peptide sequences were instead grouped in five sequence identity level intervals, 0% to 30%, 31% to 50%, 31% to 70%, 71% to 90%, and 91% to 100%, calculated by comparing each ARG to the positive dataset.

Main contributions

1. The implementation and training of an alignment-free AI-based model based on transformers for the identification of ARG fragments in metagenomic short-read data.
2. An expanded capability of ARG detection with higher sensitivity than state-of-the-art alignment-based methods, enabling the accurate detection of ARGs in short metagenomic reads with various fragment lengths and for sequences with low sequence identity to known ARGs.
3. An implementation of an strategy to expand limited labeled ARG protein data for supervised learning.

4.3 Confidence-based Prediction of Antibiotic Resistance at the Patient-level Using Transformers (paper III)

The rise in antibiotic-resistant pathogens has stressed the need to enhance diagnostic approaches with susceptibility testing, enabling the identification of effective treatments for bacterial infections. While susceptibility testing plays a crucial role in treatment optimization, current methods often involve time-consuming processes that can be critical for patients. Furthermore, in situations where diagnostic information is lacking, treatment decisions often rely on the educated guesses of medical doctors, which can fail to achieve the desired outcomes.

In this study, we introduced a transformer-based method capable of accurately predicting antibiotic susceptibility test results using patient data and incomplete diagnostic information. This model can be applied in early phases when only limited information is available, facilitating treatment decisions based on the available data. Additionally, we incorporated an uncertainty control methodology based on conformal prediction, allowing us to ensure a predetermined level of certainty for each prediction, Figure 4.6.

We utilized the complex resistant dependencies using a dataset comprising

9,224,373 antibiotic susceptibility tests carried out between 2013 and 2017. These tests involved 261,378 *Escherichia coli* isolates collected from 30 European countries, sourced from The European Surveillance System. Each isolate underwent between seven and sixteen susceptibility tests against a range of antibiotics from four different classes: five penicillins, five quinolones, four cephalosporins, and two aminoglycosides. Additionally, patient demographic information, including country of origin, and gender, was recorded. To encapsulate the diagnostic information for each isolate, we employed a concise sentence format. For instance, "SV 30 M 2013_01 LVX_R AMC_S CAZ_S AMP_S CIP_S CTX_S GEN_S TZP_R", represents a bacterium isolated at a hospital in Sweden (SV), from a 30-years-old male patient in January 2013. The isolate exhibited resistance to levofloxacin (LVX) and piperacillin/tazobactam (TZP), and was susceptible to amoxicillin/clavulanic acid (AMC), ceftazidime (CAZ), ampicillin (AMP), ciprofloxacin (CIP), cefotaxime (CTX) and gentamicin (GEN).

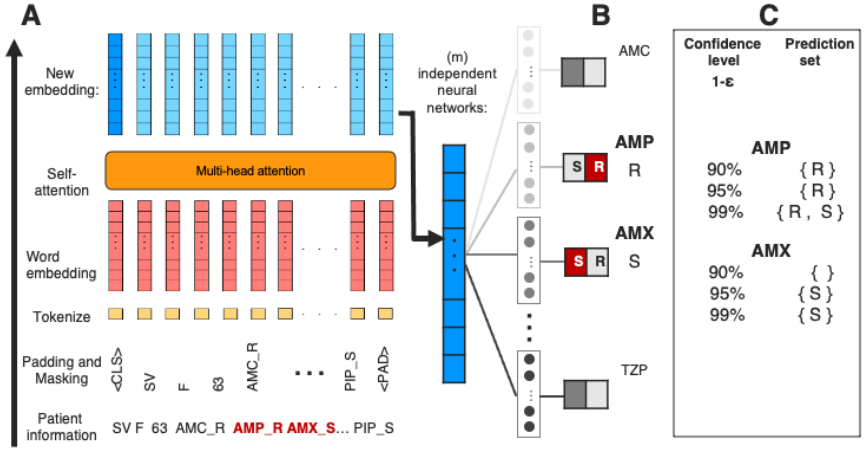

Figure 4.6: A) The patient information and susceptibility test results were fed to the transformer in the form of sequence data. A random number of susceptibility test results, here AMP_R and AMX_S in red, were removed from the input data. The sequence goes through the transformer and 16 antibiotic-specific neural networks. B) The output of each neural network is used to predict the susceptibility test result for the antibiotics not present in the input data. For training and performance evaluation, the test results that were removed from the input are compared to their corresponding neural network output. C) Conformal prediction returns prediction sets containing zero, one, or more possible results, and the prediction set will contain the true result with a pre-specified confidence level.

In both the training and testing of our model, we incorporated several input

variables, including the patient's age, gender, and country, along with a random selection of susceptibility test results ranging from 5 to 15 tests. The remaining susceptibility test results were reserved as targets to be predicted by the model.

To assess the model's performance, we computed two metrics: the major error (ME) rates, which signify the proportion of true susceptible isolates erroneously predicted, and the very major error (VME) rates, indicating the proportion of true resistant isolates erroneously predicted. Our evaluation also involved an assessment of the model's performance with respect to the specific antibiotic being predicted and the quantity of antibiotic susceptibility test results included in the input sequence.

We conducted a validation of the performance of the model using a dedicated dataset for testing, where we randomly excluded antibiotic susceptibility test results from the input sequences. The validation results indicate that the model performs well in correctly identifying susceptibility, with major errors as low as 1.7%, 3.5% 12%, and 13.5% on average for cephalosporins, quinolones, and penicillins, respectively. However, the model showed higher VME rates. Notably, the predictive performance of the model varied significantly depending on the specific antibiotic being predicted, with high performance observed for all cephalosporins and notably lower performance for penicillins. The model's performance improved with the inclusion of a greater number of susceptibility test results in the input sequence to the transformer, suggesting that the model's accuracy can be further enhanced with access to more information.

We integrated an uncertainty control algorithm based on conformal prediction to generate prediction sets. The algorithm's output for a single prediction can consist of a single label (either susceptible or resistant), multiple labels (both susceptible and resistant), or no label. The algorithm is designed to provide prediction sets that, on average, include the true label with a predefined level of confidence. This approach enabled us to independently manage the major and very major error rates, offering an advantage in scenarios where ineffective infection treatment i a concern. For cephalosporins and quinolones, the algorithm predominantly provided single-labeled sets, while for penicillins, a greater proportion of ambiguous sets containing multiple labels was observed, Figure 4.7.

In conclusion, the AI methodology presented here has the potential to enhance and supplement diagnostic information in clinical settings. This advancement could aid healthcare professionals in identifying antibiotic resistance at an earlier stage and serve as valuable support for promptly recommending effective antibiotic treatments.

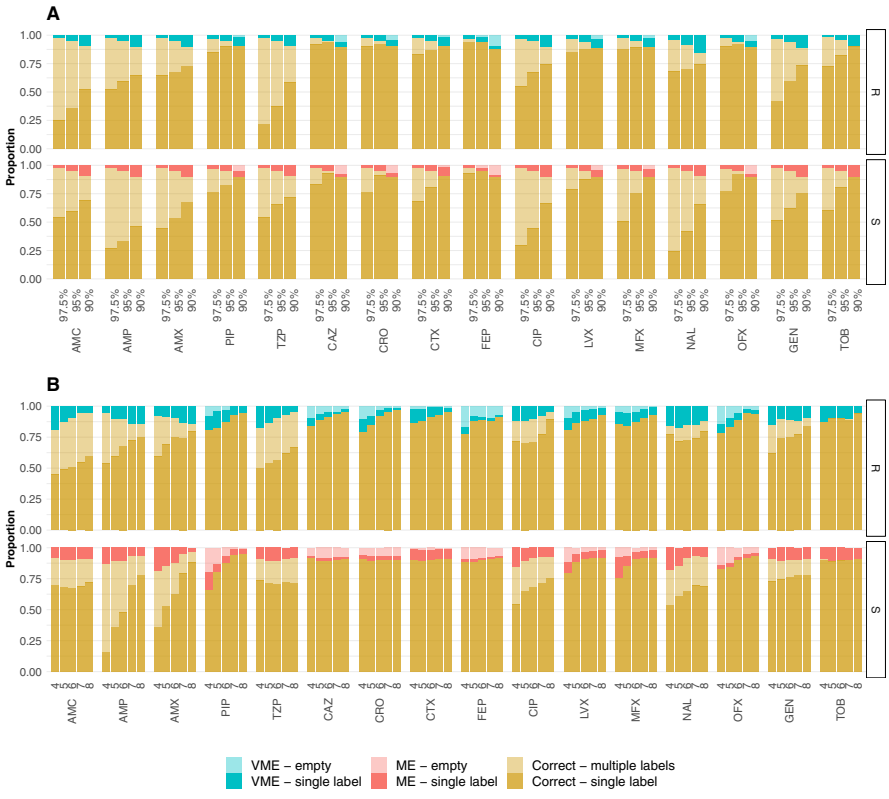

Figure 4.7: The proportion of correct predictions with a single label (opaque) and multiple labels (transparent), major errors (MEs), and very major errors (VMEs) with a single label (opaque) and empty set (transparent) predictions for resistant (R) and susceptible (S). A) The proportions are shown for each antibiotic using three different confidence levels: 90%, 95%, and 97.5%. B) The proportions are shown as a function of the number of input antibiotic susceptibility test results (90% confidence level).

Main contributions

1. The implementation of a new AI-based model that utilizes transformer and conditional inductive conformal prediction to predict diagnostic information of antibiotic resistance.
2. The utilization of patient data and a limited number of antibiotic susceptibility test results to accurately deduce a patient-specific comprehensive resistance profile for the infecting bacteria with pre-specified confidence values.

5 Conclusion

We live in a time where the battle against antibiotic resistance is a critical global healthcare challenge that threatens our ability to effectively prevent and treat bacterial infections. At the same time, we live in an era of information, where a massive amount of data is being constantly collected from almost all aspects of our lives. Moreover, rapid computational and methodological advances have produced innovative artificial intelligence applications that have been gradually implemented in everyday life. In this thesis, data-driven and AI methodologies designed to help mitigate the impact of antibiotic resistance have been presented.

A wide characterization of the resistome in host-associated and external environmental microbial communities, including over 10,000 metagenomes, is presented in paper I. Our knowledge of the resistome was expanded by including the study of both established and latent antibiotic resistance genes (ARGs). It was shown that latent ARGs are both more abundant and diverse compared to established ARGs. Although external environmental microbial communities, including soil and aquatic biomes, hosted a large diversity of ARGs, there were no indications found that specific ARGs were ubiquitously present in all the samples from these environments. In contrast, a large collection of both established and latent ARGs was found to be present in the majority of the samples of host-associated and wastewater bacterial communities. Moreover, these widely spread ARGs were, to a large extent, found in mobile genetic elements and pathogens. The results thus, show the necessity to include latent genes in future resistome studies and to elucidate the selection pressure mechanisms driving the spread of ARGs to pathogens and their prevalence is stressed. Paper I offers a new methodology to study the abundance and diversity of ARGs and a strategy to handle and analyze big and high-dimensional data, addressing aim 1a of this thesis: to provide a more comprehensive view of the resistome of bacterial communities, comprising both established and latent ARGs.

The limited knowledge of the origin of most ARGs together with the vast genomic data of human pathogenic and commensal bacteria, indicates that most ARGs likely originate in environmental bacterial hosts. Therefore, there is a need to detect both known and novel forms of resistance from environmental bacteria. In paper II, a new AI-based model using transformers designed to identify protein sequences from metagenomic data is presented. The mechanism behind transformers allows to capture of both sequence similarities and long distance dependencies between amino acids in peptide sequences, adding, thus, structural information to the prediction process on antibiotic resistance function. The model has higher performance compared to state-of-the-art alignment-based models, which are bound to only identify ARGs similar to genes present in antibiotic resistance repositories, especially for sequences with low sequence similarity to genes in the reference database. The model has the potential to be implemented in surveillance programs and could complement future studies on the resistome as the one performed in paper I. The model presented answers aim 1b of this thesis: to develop and evaluate an AI method for the detection of novel and uncharacterized ARGs.

In paper III, we present a new AI-based method developed to predict unavailable diagnostic information. The method, combined with inductive conformal prediction, provides complete profiles of antibiotic susceptibility at the patient level with a pre-specified confidence level. In the model, patient information and antibiotic susceptibility test (AST) results are combined using a transformer within the model. The evaluation of the model showed that only a few ASTs are sufficient to accurately derive a more complete resistance profile of infecting bacteria isolates. The model could be used in clinical settings by physicians as a support decision tool for the choice of efficient antibiotic treatments for bacterial infections, addressing aim 2 in this thesis: to develop and evaluate an AI method for making personalized predictions of antibiotic susceptibility test results based on incomplete diagnostic data.

The importance of data-driven approaches in the battle against antibiotic resistance cannot be overstated. Here, not only AI-based methods provide valuable tools for antibiotic resistance surveillance but also have the potential to serve as decision-support instruments for clinicians in the treatment of bacterial infections. The potential of AI to enhance traditional bioinformatics tasks has been shown. AI methods, in particular NLP tools, could help us overcome limitations of conventional methods such as sequence alignment. This would in turn, allow for more comprehensive analysis of complex data, from genes and proteins to genomes and metagenomes, even beyond the realm of antibiotic resistance. Moreover, AI has permitted us to perform personalized diagnosis by efficiently integrating different data types. In conclusion, it is expected that the continuous increase and availability of biological and patient data together

with the development of a new generation of AI-based methods will have the potential to significantly advance bioinformatics and health-care routines.

Bibliography

Allen, H. K., Donato, J., Wang, H. H., Cloud-Hansen, K. A., Davies, J., and Handelsman, J. (2010). Call of the wild: antibiotic resistance genes in natural environments. *Nature Reviews Microbiology*, 8(4):251–259.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Andersson, M. I. and MacGowan, A. P. (2003). Development of the quinolones. *J Antimicrob Chemother*, 51 Suppl 1:1–11.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6:1–15.

Barba, M., Czosnek, H., and Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1):106–136.

Bassetti, M., Rello, J., Blasi, F., Goossens, H., Sotgiu, G., Tavoschi, L., Zasowski, E. J., Arber, M. R., McCool, R., Patterson, J. V., et al. (2020). Systematic review of the impact of appropriate versus inappropriate initial antibiotic therapy on outcomes of patients with severe bacterial infections. *International journal of antimicrobial agents*, 56(6):106184.

Battle, S. E., Bookstaver, P. B., Justo, J. A., Kohn, J., Albrecht, H., and Al-Hasan, M. N. (2016). Association between inappropriate empirical antimicrobial therapy and hospital length of stay in gram-negative bloodstream infections: stratification by prognosis. *Journal of Antimicrobial Chemotherapy*, 72(1):299–304.

Bengtsson-Palme, J., Angelin, M., Huss, M., Kjellqvist, S., Kristiansson, E., Palmgren, H., Larsson, D. J., and Johansson, A. (2015). The human gut

microbiome as a transporter of antibiotic resistance genes between continents. *Antimicrobial agents and chemotherapy*, 59(10):6551–6560.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., and Larsson, D. J. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in india. *Frontiers in microbiology*, 5:648.

Bengtsson-Palme, J., Hammaren, R., Pal, C., Östman, M., Björlenius, B., Flach, C.-F., Fick, J., Kristiansson, E., Tysklind, M., and Larsson, D. J. (2016). Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Science of the Total Environment*, 572:697–712.

Bengtsson-Palme, J., Kristiansson, E., and Larsson, D. G. J. (2017a). Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*, 42(1).

Bengtsson-Palme, J., Larsson, D. J., and Kristiansson, E. (2017b). Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy*, 72(10):2690–2703.

Berglund, F., Johnning, A., Larsson, D. J., and Kristiansson, E. (2021). An updated phylogeny of the metallo-β-lactamases. *Journal of Antimicrobial Chemotherapy*, 76(1):117–123.

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. G. J., and Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, 7(1):52.

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, 13(1):42–51.

Bonin, N., Doster, E., Worley, H., Pinnell, L. J., Bravo, J. E., Ferm, P., Marini, S., Prosperi, M., Noyes, N., Morley, P. S., et al. (2023). Megares and amr++, v3. 0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic acids research*, 51(D1):D744–D752.

Botts, R. T., Apffel, B. A., Walters, C. J., Davidson, K. E., Echols, R. S., Geiger, M. R., Guzman, V. L., Haase, V. S., Montana, M. A., La Chat, C. A., Mielke, J. A., Mullen, K. L., Virtue, C. C., Brown, C. J., Top, E. M., and Cummings, D. E. (2017). Characterization of four multidrug resistance plasmids captured from the sediments of an urban coastal wetland. *Frontiers in microbiology*, 8:1922–1922. Publisher: Frontiers Media S.A.

Brown, C. L., Keenum, I. M., Dai, D., Zhang, L., Vikesland, P. J., and Pruden, A. (2021). Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Scientific reports*, 11(1):3753.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Burmeister, A. R. (2015). Horizontal Gene Transfer. *Evol Med Public Health*, 2015(1):193–194.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Davenport, T. and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dias, S. P., Brouwer, M. C., and van de Beek, D. (2022). Sex and gender differences in bacterial infections. *Infection and Immunity*, 90(10):e00283–22.

Dolejska, M. and Papagiannitsis, C. C. (2018). Plasmid-mediated resistance is going wild. *Plasmid*, 99:99–111. Antimicrobial Resistance and Mobile Genetic Elements.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ebmeyer, S., Kristiansson, E., and Larsson, D. J. (2021). A framework for identifying the recent origins of mobile antibiotic resistance genes. *Communications Biology*, 4(1):1–10.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195.

Farrar, W. E. and Eidson, M. (1971). Antibiotic resistance in" shigella" mediated by r factors. *The Journal of Infectious Diseases*, 123(5):477–484.

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O., and Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *science*, 337(6098):1107–1111.

Fricke, W. F., Cebula, T. A., and Ravel, J. (2011). Chapter 28 - genomics. In Budowle, B., Schutzer, S. E., Breeze, R. G., Keim, P. S., and Morse, S. A., editors, *Microbial Forensics (Second Edition)*, pages 479–492. Academic Press, San Diego, second edition edition.

Friedman, N. D., Temkin, E., and Carmeli, Y. (2016). The negative impact of antibiotic resistance. *Clin Microbiol Infect*, 22(5):416–422.

Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hutchings, M. I., Truman, A. W., and Wilkinson, B. (2019). Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51:72–80. Antimicrobials.

Inda-Diaz, J. S., Johnning, A., Hessel, M., Sjoberg, A., Lokrantz, A., Helldal, L., Jirstrand, M., Svensson, L., and Kristiansson, E. (2023). Confidence-based prediction of antibiotic resistance at the patient-level using transformers. *bioRxiv*, pages 2023–05.

Jaktaji, R. P. and Mohiti, E. (2010). Study of mutations in the dna gyrase gyra gene of escherichia coli. *Iranian journal of pharmaceutical research: IJPR*, 9(1):43.

Johnning, A., Karami, N., Hallbäck, E. T., Müller, V., Nyberg, L., Pereira, M. B., Stewart, C., Ambjörnsson, T., Westerlund, F., Adlerberth, I., et al. (2018). The resistomes of six carbapenem-resistant pathogens–a critical genotype–phenotype analysis. *Microbial genomics*, 4(11).

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S.,

Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol*, 1399:207–233.

Kumar, A., Ellis, P., Arabi, Y., Roberts, D., Light, B., Parrillo, J. E., Dodek, P., Wood, G., Kumar, A., Simon, D., et al. (2009). Initiation of inappropriate antimicrobial therapy results in a fivefold reduction of survival in human septic shock. *Chest*, 136(5):1237–1248.

Lai, K., Twine, N., O'brien, A., Guo, Y., and Bauer, D. (2018). Artificial intelligence and machine learning in bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1(3).

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*.

Linusson, H., Norinder, U., Boström, H., Johansson, U., and Löfström, T. (2017). On the calibration of aggregated conformal predictors. In Gammerman, A., Vovk, V., Luo, Z., and Papadopoulos, H., editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 154–173, Stockholm, Sweden. PMLR.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*, 113(21):5970–5975.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7):3348–3357. Edition: 2013/05/06 Publisher: American Society for Microbiology.

Mohr, K. I. (2016). History of Antibiotics Research. *Curr Top Microbiol Immunol*, 398:237–272.

Mukherjee, S., Stamatis, D., Li, C. T., Ovchinnikova, G., Bertsch, J., Sundaramurthi, J. C., Kandimalla, M., Nicolopoulos, P. A., Favognano, A., Chen, I.-M. A., et al. (2023). Twenty-five years of genomes online database (gold): data updates and new features in v. 9. *Nucleic acids research*, 51(D1):D957–D963.

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655.

Murray, K. A., Preston, N., Allen, T., Zambrana-Torrelio, C., Hosseini, P. R., and Daszak, P. (2015). Global biogeography of human infectious diseases. *Proceedings of the National Academy of Sciences*, 112(41):12746–12751.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., and Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont carsonella. *Science*, 314(5797):267–267.

NCBI (2023). NCBI. National Center for Biotechnology Information Genome Browser. https://www.ncbi.nlm.nih.gov/genome/browse/. [Online; accessed 2023-10-02].

Nikaido, H. (2009). Multidrug resistance in bacteria. *Annual review of biochemistry*, 78:119–146.

Pal, C., Bengtsson-Palme, J., Kristiansson, E., and Larsson, D. (2016). The structure and diversity of human, animal and environmental resistomes. *Microbiome*, 4(1):1–15.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka.

Paulsen, I. T., Brown, M. H., and Skurray, R. A. (1996). Proton-dependent multidrug efflux systems. *Microbiological Reviews*, 60(4):575–608.

Peterson, E. and Kaur, P. (2018). Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens. *Frontiers in Microbiology*, 9:2928.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C., and San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, pages 1–13.

Ruppé, E., Ghozlane, A., Tap, J., Pons, N., Alvarez, A.-S., Maziers, N., Cuesta, T., Hernando-Amado, S., Clares, I., Martínez, J. L., et al. (2019). Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature microbiology*, 4(1):112–123.

Schloss, P. D. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol*, 6(8):229.

van den Bosch, C., Hulscher, M. E., Akkermans, R. P., Wille, J., Geerlings, S. E., and Prins, J. M. (2017). Appropriate antibiotic use reduces length of hospital stay. *Journal of Antimicrobial Chemotherapy*, 72(3):923–932.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. Proceedings of Machine Learning Research.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Ye, L., Rochan, M., Liu, Z., and Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

Yelin, I., Snitser, O., Novich, G., Katz, R., Tal, O., Parizade, M., Chodick, G., Koren, G., Shalev, V., and Kishony, R. (2019). Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature medicine*, 25(7):1143–1152.

Yin, X., Zheng, X., Li, L., Zhang, A.-N., Jiang, X.-T., and Zhang, T. (2022). Args-oap v3. 0: Antibiotic-resistance gene database curation and analysis pipeline optimization. *Engineering*.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., and Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11):2640–2644.

Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., et al. (2021). Advances in metagenomics and its application in environmental microorganisms. *Frontiers in microbiology*, 12:766364.