Integration of Multi-Omic Datasets on Antimicrobial Resistance for Large-Scale Biomedical Data Science

Dissertation zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

vorgelegt dem
Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg

vorgelegt von
Sebastian Spänig, M.Sc.
geboren in Hildburghausen

Erstgutachter: Prof. Dr. Dominik Heider
Zweitgutachter: Prof. Dr. Alexander Goesmann

Eingereicht: 24.02.2022
Tag der Disputation: 14.06.2022
Erscheinungsort: Marburg
Erscheinungsjahr: 2022
Hochschulkennziffer: 1180

Acknowledgments

My greatest thanks go to my supervisor and mentor, Prof. Dr. Dominik Heider. Without his support, this work would not have been possible. His personal and scientific guidance has made me the scientist I am today. Thank you! Furthermore, I am deeply grateful to Roman Martin and Youngjun Park for their fruitful comments and suggestions. You have helped to put the finishing touches on this thesis! Furthermore, I thank all colleagues and fellow members of the working group for almost four beautiful years in Marburg. You have made the journey to this thesis a lot more pleasant. A big thank you goes to my family and friends, who also privately gave me the necessary support to get through the long, sometimes rocky, road to the end. Finally, I would like to thank Hannah, who is always there for me with her endless patience and energy.

Urheberschaftserklärung

Hiermit versichere ich, dass ich die vorgelegte Dissertation selbständig und ohne fremde Hilfe verfasst, nicht andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt, alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet, sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht habe.

Marburg, 08.06.2022

Sebastian Spänig

Integration of Multi-Omic Datasets on Antimicrobial Resistance for Large-Scale Biomedical Data Science

Zusammenfassung

Antimikrobielle Resistenz (AMR) führt zu einem enormen Gesundheitsrisiko und wurde daher von der WHO als eine der größten Belastungen für die moderne Gesellschaft eingestuft. Aufgrund unwirksamer Antibiotika werden alltägliche Operationen zu lebensbedrohlichen Eingriffen. Strenge staatliche Maßnahmen sollen die Verabreichung antimikrobieller Mittel überwachen und so die Verbreitung von AMR kontrollieren. Das Eingreifen von Verantwortlichen aus dem Gesundheitswesen und eine verantwortungsvolle Anwendung in der Human- und Veterinärmedizin sind dringend erforderlich. Vor diesem Hintergrund wird die abwasserbasierte Epidemiologie eingesetzt, um verschiedene Umweltfaktoren zu untersuchen, die die Antibiotikaresistenz begünstigen und um die Entwicklung dieser in der gesamten Bevölkerung zu überwachen. Antibiotikarückstände in menschlichen Ausscheidungen sind ein wichtiger Faktor für AMR. Daher ist es naheliegend, die Zu- und Abwässer von Kläranlagen zu untersuchen. Das gereinigte Abwasser wird letztlich in Flüsse, Seen oder das Meer eingeleitet, wodurch AMR von einem lokalen zu einem globalen Gesundheitsproblem wird. Daher ziehen die Wissenschaftler zunehmend Süß- und Salzwasser für umfassende AMR-Untersuchungen in Betracht. Gewässer zur Erholung könnten ein erhebliches Gesundheitsrisiko darstellen, wenn sie mit resistenten Bakterien belastet sind. Tatsächlich wurden im Rahmen der Süßwasser-Epidemiologie Hotspots in asiatischen Seen festgestellt, was die Dringlichkeit einer zeitnahen und konsistenten AMR-Überwachung weltweit unterstreicht. Allerdings wird die Konsistenz der Daten durch eine große Vielfalt an bioanalytischen Methoden erschwert. Daher wurden im Rahmen dieser Dissertation standardisierte Proben aus zahlreichen europäischen Süßwasserseen integriert, untersucht und ausgewertet. Dies ermöglichte Basiswerte für AMR zu ermitteln und die künftige, umfangreiche Überwachung zu erleichtern.

Die Ergebnisse unterstreichen außerdem, dass multiresistente Krankheitserreger alternative therapeutische Optionen jenseits konventioneller Antibiotika erfordern. Daher untersuchen Wissenschaftler antimikrobielle Peptide (AMPs). Bis heute sind mehrere AMPs in klinischen Studien fortgeschritten oder haben sogar Marktreife erlangt. Der Erfolg ermutigte Wissenschaftler Methoden des maschinellen Lernens (ML) für das AMP-Screening im Hochdurchsatzverfahren zu einzusetzen. Die ML-basierte Integration von Peptidomics Daten setzt jedoch ein maschinenlesbares Format voraus, was die Optimierung der Hyperparameter weiter erschwert. Daher wurden im Rahmen dieser Dissertation auch die Leistung in Bezug auf Kodierungen, Modelle und den biomedizinischen Bereich untersucht. Schließlich wird in einer weiteren Studie dieser Thesis ein neuer Ansatz für die unüberwachte Auswahl von Kodierungen und die Konfiguration von Ensembles vorgestellt. Kurzum, diese Arbeit erörtert die Sammlung, Analyse und Integration von Multi-Omics-Daten, um den Weg für die datengesteuerte Forschung von AMR zu ebnen.

Integration of Multi-Omic Datasets on Antimicrobial Resistance for Large-Scale Biomedical Data Science

Abstract

Antimicrobial resistance (AMR) results in tremendous health risks, causing the World Health Organization (WHO) to designate it as one of the significant burdens for modern society. Owing to ineffective antibiotics, once everyday surgeries will become life-threatening interventions. Rigorous governmental measurements are supposed to supervise administration of antimicrobials, hence controlling AMR dissemination. The intervention of healthcare stakeholders and responsible application in human and veterinary medicine is urgently required. In this light, wastewater-based epidemiology has been established to examine various environmental factors promoting AMR and monitoring their development population-wide. Antibiotic residuals in human excrements are a significant driver for AMR, and assessing in- and effluent of wastewater treatment plants is evident. Treated wastewater is ultimately released in rivers, lakes, or the sea, elevating AMR from a local to a global health concern. Thus, researchers consider increasingly fresh and salt waters for comprehensive AMR surveys. In this light, recreational waters could be a significant health risk if strained with resistant bacteria. Indeed, freshwater-based epidemiology ascertained hot spots in Asian lakes, underpinning the urgency for timely and consistent AMR surveillance worldwide. However, data consistency is hampered due to a great variety of bioanalytical methods. For this reason, as part of this thesis, we integrated, examined, and evaluated standardized samples from numerous European freshwater lakes. Baseline levels of AMR have been detected, which facilitates future monitoring on a large scale.

The results further emphasized that multi-resistant pathogens require alternative therapeutic options beyond conventional antibiotics. Therefore, scientists study antimicrobial peptides (AMPs). To date, several AMPs advanced in clinical trials or gained market maturity. The success encouraged researchers to develop advanced machine learning (ML) methods for high-throughput AMP screening. However, ML-based integration of peptidomics assumes a machine-readable format, further challenging hyper-parameter optimization. Thus, we explored as part of this thesis the performance concerning encodings, models, and the biomedical domain. Finally, we contributed a novel approach for unsupervised encoding selection and ensemble configuration to this dissertation. In summary, this thesis addresses the collection, analysis, and integration of multi-omic data to pave the way for data-driven research on AMR.

Contents

1 Introduction

AMR claimed already millions of death, around 1.5 Million in 2019 alone[176]. Over- and misuse in human and veterinary medicine is the major cause[73,100,190]. To regain control of this severe pandemic, factors for local and global resistance events must be considered[100]. In this light, the examination of epidemiological aspects utilizing waste- and freshwater samples enables timely intervention by health care stakeholders[42]. Moreover, environmental epidemiology, specifically based on wastewater samples, recognizes local drivers of AMR and the association on a population scale[42,54]. Ultimately, AMR leads to ineffective antibiotics; thus, previously low-risk surgeries become life threatening[228]. The search for alternative antimicrobials is therefore a natural consequence, bringing AMPs more and more into focus[151]. Although AMPs belong to the ancient immune system of diverse organisms, various obstacles, which will be addressed later in this work, must be taken until AMPs achieve clinical relevance[125]. However, artificial intelligence paved the way for high-throughput prediction and optimization, finally bridging the gap between *in vitro* activity and *in vivo* application[71].

Many tools and technologies have been published to study the rise of AMR. To enable the integration of multi-model datasets, standardized approaches are required. As a proof of concept, we examined how streamlining different aspects could be achieved. First, we elaborated standardized analysis of freshwater samples to monitor AMR dissemination and revealed base line levels of AMR in putative pathogen-free environments. For the second part we applied artificial intelligence to predict peptide properties, focusing on antimicrobial amino acid sequences as an alternative strategy to conventional antibiotics. The case stud-

ies are embedded in a comprehensive literature review to motivate standardized integration of multi-omic datasets and to illustrate accompanied challenges. The standardization techniques could pave the way for the integration of multiple modalities for AMR; thus, enable a comprehensive picture of prevalence, spread, and alternative therapies.

The aim of this thesis is a presentation of the published studies embedded in a comprehensive literature review to identify various challenges for the integration of multi-omic datasets concerning AMR. Furthermore, considering the respective domains, the latest publications are analyzed to understand the diversity of the sampling methods and available data. Five research articles have been contributed to augment current efforts, and to provide perspectives concerning the limitation and future opportunities of streamlined data integration and processing. Ultimately, the study ensues the hypothesis that unified data collection and workflows ease integration; hence, the comparability across multiple studies, specifically addressing various facets of AMR.

The following chapters elucidate the relation of AMR, environmental epidemiology (EE), AMPs, and ML. The second chapter introduces the variety of AMR mechanisms, modes of dissemination, and drug classes (see Chapter 2). In addition, this chapter enlightens the challenge of increasing multi-drug resistant pathogens. The third chapter addresses measures to detect AMR dissemination based on EE, particularly, by sampling waste- and freshwaters (see Chapter 3), which involves the **first publication** (see Section 6.1). The requirement for alternative strategies becomes obvious. Thus, the fourth chapter covers the effects of host-defense peptides (HDPs), focusing on AMPs, which includes biochemical characteristics, modes of action, applications, and clinical relevance (see Chapter 4). The fifth chapter introduces a broad range of ML techniques for identifying novel AMPs from validated peptidomics datasets (see Chapter 5). The introduction to ML also encompasses encodings, crucial for representing amino acid sequences, which is covered by the **second publication**, presented in Section 6.2. Moreover, the ML chapter integrates the **third publication**, which examines the encoding performance on various applications (see Sections 6.3). Encoding selection remains challenging; thus, the **fourth publication** presents an unsupervised approach (see Section 6.4). Chapter 5 concludes with the presentation of biomedical applications, such as the high-throughput screening of peptide binding kinetics, as described in the **fifth publication** (see Section 6.5).

The articles published for the dissertation are listed in Table 1.1. The table also highlights the respective contributions per topic as well as a reference to the according background and result sections.

Table 1.1: List of papers annotated with the respective author contributions. Contributions by Sebastian Spänig (SS) are highlighted in bold. The background sections link the publications into the topical context. Refer to the Publications section for the actual article.

Publication	Contribution	Context	Results
A multi-omics study on quantifying antimicrobial resistance in European freshwater lakes	**SS** and DH developed the concept and designed the experiments. JB designed the sampling campaign. JKN, DB, and JB collected and preprocessed the sequencing data. **SS** and LE performed the experiments and analyzed the data. **SS**, LE, MI, and DH interpreted the results. **SS** and DH wrote the manuscript. JB and DH supervised the study.	3.2.1	6.1
Encodings and models for antimicrobial peptide classification for multi-resistant pathogens	**SS** developed the concept and wrote the manuscript. DH gave conceptual advice, supervised the study, and revised the final draft.	5.2.1	6.2
A large-scale comparative study on peptide encodings for biomedical classification	**SS** and DH developed the concept. **SS** designed and performed the experiments as well as gathered, curated and analyzed the data. SM implemented the Delaunay Triangulation and Distance Frequency encoding. GH supervised the data visualization aspect and created the logo as well as the overview figure. **SS**, ACH. and DH interpreted the results. **SS** wrote the manuscript. ACH and DH supervised the study.	5.3.2	6.3
Unsupervised encoding selection through ensemble pruning for biomedical classification	**SS** and DH developed the concept. **SS** designed and performed the experiments and analyzed the data. **SS** and DH interpreted the results. AM implemented the MVO algorithm. **SS** wrote the manuscript. DH supervised the study.	5.4.1	6.4
Multivalent binding kinetics resolved by fluorescence proximity sensing	Conceptualization: HMM, CS; Methodology: CS, AS, NA, IB, **SS**; Software: **SS**, DH, Formal Analysis: CS, AS, NA, IB, **SS**; Investigation: CS, AS, NA, IB;Writing - Original Draft: CS, HMM; Writing - Review & Editing: AS, **SS**, NA, IB, RS, DH; Visualization: CS, **SS**; Supervision: HMM, DH, RS, WS; Project Administration: HMM, RS, WS, DH; Funding Acquisition: HMM, RS, WS, DH	5.3.1	6.5

2
Antimicrobial Resistance

Bacteria are critical for various aspects of life and form the basis of many environmental, industrial, and metabolic processes[74]. The industry employs bacteria for sewage purification or food production; however, bacteria are also involved in digestive nutrient utilization[74]. Although it has been long supposed that microbes outnumber human cells tenfold, a recent study assumes a one-to-one ratio[208], still stressing the importance of microorganisms. The human gastro-intestinal microbiome harbors many microorganisms, including around 1000 species from 8 taxonomic families[145]. Bacteria colonize other ecosystems than the human digestive system and are ubiquitous in terrestrial and aquatic environments such as soils and waters[74]. As such, bacteria are endemic in ecologic niches, therefore possessing defense measures to protect against microbial competitors and abiotic molecules[188]. These organisms produce antibiotic substances, completing their intrinsic resistance system for defense against prokaryotic intruders[188]. Moreover, microbes acquire antimicrobial resistance (AMR) to cope with selection pressure and environmental changes, for instance, human-made antibiotic residuals or from antibiotic-producing species[134]. It should be highlighted that human-made pollution is merely one element contributing to AMR, and intrinsic resistance is long present[115]. Recently, researchers acknowledged the latter by referring to it as the “ancient” resistome[134], which is the entirety of AMR gene clusters encoded on microbial chromosomes or plasmids[122].

As mentioned above, AMR concerns the intrinsic or acquired resistance against antimicrobial drugs. It describes the transformation of formerly susceptible to resistant pathogens

in a clinical context. The European Centre for Disease Prevention and Control (ECDC) observed significantly increased AMR rates for several critical pathogens between 2015 and 2019[66]. Owing to this concerning trend, the WHO reckons about 50 million deaths by 2050 and tremendous financial burdens for public healthcare systems[115,155]. Without effective antibiotics, low-risk surgeries, such as caesareans, could become life-threatening interventions[228]. Blair *et al.* (2015) confirmed antibiotics as essential for modern medicine since they allow the treatment of microbial infections and mitigate, for instance, side-effects of complex operations[19]. Thus, rising multi-resistant pathogens urge novel antibiotics[19]. In addition, the inappropriate usage of antibiotics intensifies this problem[100].

More and more drug residuals enter sewage, and incomplete purification causes environmental pollution[134]. Generally, the spread of AMR via hospital or municipal effluent or antibiotic release due to livestock farming is a significant concern[100]. The resulting selection pressure increases the probability that susceptible species acquire AMR; consequently, resistant bacteria become more prevalent[134]. Nevertheless, the subsequent section addresses resistance mechanisms before more details on AMR dissemination are described.

2.1 Mechanisms

Effectively, the genetic blueprints of AMR are encoded on the microbial chromosome or the plasmid. Chromosomal- and plasmid-encoded genes ultimately provide for the resistance induced by diverse modes of action. To this end, microbiologists distinguish between intrinsic and acquired AMR.

Intrinsic AMR comprises efflux pumps, selective membranes, as well as complex genomics rearrangements to mitigate antibiotic efficiency[47]. In particular, efflux pumps refer to proteins spanning the outer layer of the microbial cell wall and exclude toxic molecules, for instance, antimicrobials[47]. Removal via efflux pumps is specifically enabled through the resistance-nodulation-division (RND) complex[19,47]. Cox and Wright (2013) enumerate additional efflux pumps, namely “the ATP binding cassette”, “the major facilitator”, “the multidrug and toxic-compound efflux”, and “the small multidrug resistance”[47]. Efflux pumps exist in a wide range of bacteria, not necessarily for antibiotic removal, and can be classified in two groups: for a particular molecule class or with unspecific affinity[47]. To cope with environmental changes, higher expression of these transmembrane proteins result in the more effective discharge of antimicrobial agents[19].

Another intrinsic resistance mechanism refers to selective cell membranes. Gram-negative bacteria employ cell membrane selectivity to prevent infiltration of antibiotics[47]. The impermeability of gram-negative cell walls is due to physicochemical properties[47]. A cell wall consisting of compact phospholipids can form a tight mesh and hinder molecule transition[19,47]. In contrast, gram-positive bacteria are generally more sensitive to antimicrobials and possess

a looser, mostly composed of peptidoglycan molecules, outer membrane layer[47]. Furthermore, some species adapt proteins spanning the outer layer, specifically porins, resulting in decreased antibiotic selectivity and permeability[19]. In the review by Blair *et al.* (2015), the authors refer to several clinically relevant species, for instance, from the genus *Acinetobacter*, which reduce porin biosynthesis or employ different variants of this outer-membrane proteins to decrease permeability[19].

Choi *et al.* (2019) examined the role of different porine proteins and their role in AMR[43]. The study results indicated that mutations in the outer-membrane protein (omp)F increased insensitivity[43]. Altered ompA variants mitigated resistance to specific antibiotics[43]. In addition, mutations in ompC possess different effects against antibiotics, comprising increased resistance, enhanced efficiency, or retained resistance[43]. The authors conclude that proteins spanning the outer layer of the cell membrane contribute individually, underpinning the importance of porine proteins for AMR[43]. Blair *et al.* (2015) noticed that cell membrane modification resulting in biochemical alteration mitigates the efficiency of polymixins and daptomycins[19].

A further intrinsic defense mechanism concerns the adaption of the chromosomal-encoded genes, including mutations or mobile genetic elements (MGEs) such as transposons[47,225]. If antibiotics penetrate the cell membrane and reach intracellular compartments, the microbial cell expresses modified proteins, which retain functionality; however, inhibition due to antibiotics is mitigated[19]. An example is the resistance of *Pseudomonas* genera to triclosan, owing due to an additional gene encoding for a resistant homolog version of the actual target[19]. Transposable elements enable more complex alterations of the chromosome[225]. Blair *et al.* (2015) pointed out that genes encoding for efflux pumps can also be translocated to the plasmid, allowing faster adoption of multi-resistance[19]. Single nucleotide substitutions resulting in antibiotic resistance can be rapidly adapted in microbial communities[19]. For instance, *Staphylococcus aureus* developed resistance to the linezolid antibiotic in this way[19]. A further strategy is the uptake of free DNA, achieved via transformation, which leads to the development of genetic variants from concatenated homologous genes[19]. Blair *et al.* (2015) highlighted that novel genes, for example, acquired by horizontal gene transfer (HGT), encode for mutated proteins, likewise reducing susceptibility towards, for instance, β-lactams or oxacillin antibiotics[19].

In summary, efflux pumps, cell wall selectivity, and chromosomal adaption protect microbial organisms from biocidal threats[47]. Nevertheless, bacteria developed various mechanisms to inactivate antibiotics, which already passed the membrane[19]. Methylation of the antibiotic binding site in proteins or molecules increased insensitivity to multiple antibiotics, which contrasts to resistance systems based on DNA alterations since genes encoding for the antibiotic destination are not altered[19]. A further strategy concerns the protection of the DNA synthesis, which is the primary target of the quinolone antibiotic[19]. Bacteria can also directly tackle antibiotic agents to hamper the mode of action[19]. Such defense mechanisms,

comprising premature disintegration of the antibiotic and chemical modification, ultimately lead to binding inhibition of the antibiotic and its inactivation[19]. Antibiotics affected through direct modulation include various β-lactams, rifamycins, and aminoglycosides[19].

It is crucial to detect and understand the great variety of AMR mechanisms to introduce novel potent drug candidates[156]. To this end, in a study realized by Manna *et al.* (2021), the authors first determined the mutation, which is responsible for resistance against the unmodified trimethoprim (TMP)[156]. Afterward, they designed the derivate antibiotic, namely 4'-DTMP, which retained target affinity, despite the amino acid substitution[156]. In addition to increased susceptibility than the wild type, the authors observed a reduced AMR development using the modified version of TMP[156]. This stressed the significance of profound comprehension of AMR to develop improved antibiotics[156].

Furthermore, antibiotics can tackle over-expressed efflux pumps by reactivating suppressor genes[19]. Targeted silencing of the gene complex encoding for the efflux pump mechanism resulted in significant sensitivity of former resistant *Pseudomonas aeruginosa* strains[47]. However, some bacteria confer complex resistance using multiple mechanisms. For instance, Cox and Wright (2013) underpinned reciprocal effects of multiple resistance mechanisms[47]. To this end, the authors refer to a study conducted by Vaara (1992), who treated *Escherichia coli* with a membrane-disrupting antimicrobial peptide prior to antibiotic dispensation[47,233]. The constant susceptibility indicates that the cell wall is merely one part of the resistance cascade[47,233]. Lázár *et al.* (2018) demonstrated that peptides with antimicrobial efficiency increase sensitivity to conventional antibiotics if jointly administered[135]. Parallel administration of multiple antibiotics is thought to elude chromosome-encoded resistance via the inhibition of genes, for instance, encoding β-lactam insensitivity[19].

MGEs, including the transposon machinery mentioned above, enable other species to acquire resistance and bypass antibiotic effects[47]. Insensitive bacteria can integrate antimicrobial resistance genes (ARGs) into their plasmid and exchange the genetic information with conspecifics using HGT[19]. Furthermore, vertically inherited ARGs; hence, dissemination within the same species occurs[47]. Enhanced AMR rates through acquired resistance, in particular fostered by MGEs, constitute aggravated predictivity of environmental resistance flow[134] and is subject to diverse local and global conditions[100], which will be described next.

2.2 Dissemination

AMR originates from a nearby environmental mutual exchange, fostered by local ecosystems, referred to as "One Health" domains[100]. These domains encompass, for instance, antibiotic administration in animal husbandry resulting in pollution of freshwater[100]. Hernando-Amado *et al.* (2019) shed light on different aspects of AMR dissemination as well as relevant geographical and socio-economic factors[100]. The authors stressed the severity of the AMR

pandemic, which is further highlighted by economically, politically, and scientifically acknowledgment[100]. Additionally, they noted that ARGs emerging from non-pathogenic bacteria drive AMR in One Health environments[100]. In contrast, "Global Health" determinants consider One Health at a large scale, hence, the transmission of AMR via worldwide interactions, including traveling, trading, or animal fluctuation[100]. ARGs could be verified even in the arctic[134]. Larsson and Flach (2021) underpinned the Global Health aspect by referring to multiple ways of ARG carrying feces dissemination, for instance, through bird migration[134].

Furthermore, Hernando-Amado *et al.* (2019) pointed out various local conditions ultimately impacting global AMR spread[100]. AMR acquirement is facilitated through monocultural farming, hence, the loss of target variability, which eases the migration of bacteria to different hosts[100]. More diverse animal husbandry would have been a more significant obstacle for dissemination and health implications of the multi-resistant *Staphylococcus aureus* (MRSA)[100]. Furthermore, Kriegeskorte and Peters (2012) showed that *Staphylococcus aureus* obtained resistance mainly through HGT of MGEs[127].

Although more research is required to determine the actual transmission mode of action for *Staphylococcus aureus*[127], plasmid genes encoding for resistance against multiple antibiotics, including *β*-lactams and colistins, are readily shared among pathogens[100]. Bacteria conduct HGT via conjugation (DNA exchange), transformation (deoxyribonucleic acid (DNA) uptake), and transduction (viral transmission)[74]. Some species prefer a particular mechanisms[188]. For instance, *Streptococcus pneumoniae* employ transformation or *Staphylococcus aureus* prefer transduction, with conjugation being the most prevalent[188]. Winter *et al.* (2021) enlightened the relation of transformation and acquirement of ARGs across distinct species[249]. Specifically, the work stressed that AMR dissemination owing to natural transformation is enhanced by cross-species spread and transmissible gene clusters[249]. With this respect, Jian *et al.* (2021) highlighted phages carrying ARGs concerning particular acquirement through transduction[74,115].

As stated by Larsson and Flach (2021), the dissemination of MGEs contributes significantly to the rise of AMR[134]. A natural source of MGE carrying ARGs are antibiotic-producing bacteria, which utilize self-produced antibiotics to protect against competitive organisms; thus, they are intrinsically resistant[34,134]. Self-resistance in antibiotic-producing bacteria is generally chromosomal encoded, whereas, in pathogens, MGEs encode resistance mechanisms[188]. Nevertheless, AMR acquirement is governed by manifold environmental relations and complex interactions[134]. Larsson and Flach (2021) addressed the role of the environment to increased selective pressure and resistance[134]. The authors also considered the relation of anthropogenic contaminants and AMR development and finally recommend additional *in vitro* experiments to reconstruct selection patterns to predict upcoming resistance acquirement events[134].

Since MGE dissemination are characteristic for AMR development, Hernando-Amado *et*

al. (2019) highlighted countermeasures to tackle AMR spread on local and global levels[100]. An essential tool is the controlled application of antibiotics, which have been prescribed increasingly in the last decades, with a presumable growth until 2030[100]. Although the overuse of antibiotics in the first world is problematic, enhanced antibiotic application, partly due to self-prescription in some countries, has been observed[100]. The situation in third-world areas is also aggravated on account of regional sanitation standards[100]. Therefore, advances in wastewater treatment are crucial to prevent further AMR spreading and to stem antibiotic residue dispensation[100]. Antimicrobial usage in livestock farming outweighs human application, specifically in low or middle-income countries[100]. However, high rates are also reported for the USA[100]. Winter *et al.* (2021) underpinned the impact of biocides in the environment, which increases the availability of DNA residues for bacterial uptake[249].

Antibiotic administration of infected hosts, hence, humans or livestock, significantly contributes to environmental dissemination of ARGs[134]. Misuse of antimicrobial drugs enables DNA transformation and mutation events, which is additionally enhanced by close spatial bacteria in biofilms[134]. Research to understand and predict AMR spreading events is essential to establish trajectories, revealing ARG flow[134]. Although multiple ARGs can be tracked to the source organism, the initial species for a great variety of resistance genes is unknown, which underpins the concern of diverse microbial communities as ARG pools[134].

The intrinsic resistome of antibiotic producer-organisms contributes to AMR negligible, whereas anthropogenic pollution, including patient and livestock excrements, insufficient removal of drugs in manufacturers waste, and industrial residues, has a far greater impact[134]. It is also notable that environmental antibiotic concentrations are below lethal doses; however, samples of, for instance, clinic, community, and industrial wastewater indicated the contrary[134].

Moreover, transmission events are fostered by physicochemical properties of the environment, for instance, surrounding temperature[134]. Physicochemical attributes are putatively more critical for selection and AMR levels than non-lethal antibiotic residues concentrations[134]. Larsson and Flach (2021) suggested the reduction of any external factors initiating or enhancing microbial adaption processes[134]. Additional environmental pollutants enforce selection, and the impact of antibiotics residues on AMR might be exaggerated[134]. Albeit governments have already introduced measurements to govern antibiotic exposure in an industrial context, including pollution by drug manufacturers or agricultural applications, the authors demanded more profound actions[134].

As mentioned above, Cox and Wright (2013) concluded that although human-derived antibiotic pollution drives selective pressure, AMR is long-present in non-pathogenic bacteria, which must be considered in future research on the “intrinsic resistome”[47]. A promising method has been introduced by Ellabaan *et al.* (2021)[65]. The authors developed a computational model employing genome data from 56.716 bacteria to predict ARG dissemination

networks[65]. The study revealed 152 MGEs in a subset of 22.963 genomes[65]. Moreover, the authors identified transmissible genes encoding for resistance against various drug classes, including *β*-lactams or aminoglycosides[65]. Based on their findings, Ellabaan *et al.* (2021) further confirmed the severity of AMR spread, including resistance gene transmission in several ESKAPE species, for instance, *Staphylococcus* species[134]. *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* spp. (ESKAPE) possess different strategies to elude or mitigate antibiotic stress; hence, they are a significant threat, and timely development of effective countermeasures is critical[155]. According to Mancuso *et al.* (2021), the World Health Organization (WHO) designated Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter spp. (ESKAPE) organisms as critical multi-drug resistant pathogens[155]. Therefore, the upcoming section addresses resistance mechanisms concerning different antibiotic drug classes in more detail.

2.3 Drug Classes

Peterson and Kaur (2018) enumerated drug classes addressing multiple cellular compartments[188]. The authors listed crucial agents according to the targets: *β*-lactams, lipo-, and glycopeptides disintegrate the cell wall, aminoglycosides, tetracyclines, as well as macrolides perturb the protein synthesis cascade[188]. Moreover, platensimycin affects lipid synthesis[188]. As mentioned above, multiple modes of action allow bacteria to evade their own manufactured antimicrobials: efflux pumps, loss of antibiotic function via inhibitor proteins in the cell, or premature modification in the cell wall[188]. For increased efficiency, numerous species, such as *Streptomyces peucetius*, activate multiple resistance mechanisms in parallel[188]. Peterson and Kaur (2018) argued that in the long term, bacteria potentially elude all antibiotics owing to the diversity of natural AMR[188]. As an example, they referred to Dantas and Sommer (2012), who described the effect of the ampC gene acquirement on plasmids, resulting in a *β*-lactam resistance pandemic[53]. Furthermore, Pages *et al.* (2009) examined the relation of efflux pumps and *β*-lactam resistance[183]. The results indicate that efflux pump are critical for *Klebsiella pneumoniae* insensitivity[183]. In contrast, albeit several studies examined efflux pumps as a potential drug target, effectivity is hampered due to the low specificity of bacterial and human cells[47].

To sum up, Table 2.1 enumerates ten essential drug classes and is based on the WHO list of “Critically Important Antimicrobials for Human Medicine”[250]. The structure is according to the WHO prioritization factors and distinguishes between “Highest priority” (rows one to five) and the first five “High priority” antimicrobials (rows six to ten). The taxonomy weights crucial antibacterial drugs stronger, possessing an increased risk of resistance development. More

information about the critical antimicrobials can be found at the Comprehensive Antibiotic Resistance Database (CARD) using the Antibiotic Resistance Ontology (ARO) term[3].

Table 2.1: Based on the WHO list of "Critically important antimicrobials"[250], the table enumerates drug classes as well as respective examples, the source bacteria, the mode of action, and the ARO term from the CARD database[3]. The first five antibiotics are "Highest Priority", and the last records are "High priority".

Drug class	Example	Source	Mode of action	CARD
Cephalosporins (3rd+ generation)	Ceftriaxone	Semi-synthetic[1]	Cell wall synthesis inhibition[227]	ARO:0000062
Glycopeptides	Vancomycin	*Amycolatopsis orientalis*[254]	Cell wall synthesis inhibition[164]	ARO:0000028
Macrolides and ketolides	Azithromycin	Semi-synthetic[185]	Protein biosynthesis inhibition[185]	ARO:3000158
Polymyxins	Colistin	*Bacillus colistinus*[18]	Cell wall lysis[18]	ARO:0000067
Quinolones	Ciprofloxacin	Synthetic[109]	DNA replication pertubation[109]	ARO:0000036
Aminoglycosides	Gentamicin	*Micromonospora purpurea*[83]	Protein biosynthesis inhibition[242]	ARO:0000014
Ansamycins	Rifampicin	*Amycolatopsis rifamycinica*[11]	RNA polymerase perturbation[226]	ARO:3000169
Penems	Meropenem	Synthetic[193]	Cell wall synthesis inhibition[193]	ARO:0000073
Glycylcyclines	Tigecycline	Semi-synthetic[181]	Protein biosynthesis inhibition[181]	ARO:0000030
Lipopeptides	Daptomycin	*Streptomyces roseosporus*[133]	Cell wall lysis[96]	ARO:0000068

3
Environmental Epidemiology

Reconstructing antimicrobial resistance gene (ARG) dissemination is critical for investigating local and global resistance events[100]. The examination of human fecal to quantify antibiotic residuals is therefore essential[134]. Since stools ultimately enter wastewater processing, epidemiology-based quantification on a large-scale eases tracking[134]. Larsson and Flach (2021) compared wastewater-based epidemiology (WBE) and targeted examination of patient samples to track antimicrobial resistance (AMR) dissemination[134]. According to the authors, the advantage of WBE concerns large-scale screening capabilities with timely notification of severe AMR developments[134]. Patient-wise analysis is more costly, and different regional sampling policies ultimately hamper comparability[134]. Choi *et al.* (2018) specified additional advantages of WBE[42]. WBE can be utilized to track concentrations of chemicals and biologicals in normalized wastewater samples, comprising the initial concentration, the removal efficiency during treatment, and finally, residuals in treated wastewater; thus, the resulting exposure to the environment[42]. In addition, large-scale deployment of WBE facilitates the inference of drug consumption and dissemination of pathogens for a given time and place[42]. Besides quantification of drugs and medicines or the exposure to pathogens, WBE is employed to investigate the nutrition status of a population or to detect industrial pollution[42]. Finally, Choi *et al.* (2021) underpinned the collaboration of national or international institutes, which employ WBE, for instance, to monitor population-based drug consumption[42].

3.1 Wastewater-based Epidemiology

WBE is utilized to measure microbial contamination; hence, to track AMR development in the environment[42]. Consequently, multiple studies examined antibiotic pollution using WBE. In the following, representative studies will be presented, which reflect the diversity of WBE. For instance, Yuan *et al.* (2016) analyzed wastewater influent from four large Chinese cities, including Hong Kong and Beijing[260]. Although the results revealed the least antibiotic residual concentration for Hong Kong and Beijing among the validated cities, the authors conclude that Chinese antimicrobial prescription compared to Italy is significantly higher[260]. The authors used Liquid Chromatography - tandem Mass Spectrometry (LC-MS/MS) to obtain the rate of antibiotic molecules[260]. Afterward, the final concentration has been calculated, incorporating the total per-day wastewater volume and population size[260].

Galani *et al.* (2021) suspected that prescription-free antibiotics potentially contributed to a 1.5-fold rise of antibiotic usage in Athens (Greece)[79]. Specifically, the study surveyed wastewater to monitor prescription rates of various drugs and associate them with the coronavirus disease 2019 (COVID-19) pandemic[79]. The authors demonstrated the strength of WBE since it provided a detailed picture of drug prescription in Athens[79]. Concerning hydroxychloroquine, a malaria drug with antiviral effects, Galani *et al.* (2021) detected a five-fold increase compared to the pre-pandemic era[79]. Initially, physicians utilized hydroxychloroquine to treat COVID-19 infections[79]; however, the European Medicines Agency (EMA) assigned it as insufficient effective in June 2020[173]. The authors integrated mass spectrometry data, and the subsequent statistical analysis included the Wilcoxon signed-rank test to measure the effect between the two time points[79]. The drug intake estimation is based on molecule volume, wastewater quantity, and population size[79].

An advantage of WBE concerns monitoring antibiotics removal efficiency of wastewater treatment plants (WWTPs). In particular, Watkinson *et al.* (2009) identified antibiotic residuals in East Australian WWTPs and adjacent freshwaters[244]. The verified WWTPs are capable of removing four-fifth of the antibiotics from the wastewater influent[244]. Low quantities of residuals have been detected in receiving freshwaters; however, one river, free from WWTP effluent, revealed significantly lower rates[244]. The authors employed HPLC Tandem Mass Spectrometry (HPLC-MSMS) to detect antibiotic-related molecules[244]. For the difference between sample sites, the Analysis of Variance (ANOVA) has been employed[244]. Furthermore, Spearman's rank correlation was utilized to determine the effect of patient excretion and hospital sewage pollution[244]. The authors concluded that WWTPs contribute to antibiotic pollution in freshwaters; nevertheless, animal husbandry and stormwater runoff could also be crucial drivers[244].

In this light, Mirzaei *et al.* (2019) applied WBE to compare levels of various antibiotics in the in- and effluent of two WWTP in Teheran (Iran)[169]. The study approved inadequate

neutralization of antibiotic agents with various efficiency[169]. The authors further hypothesize that the removal efficiency depends on their chemical properties[169]. In this work, High Performance Liquid Chromatography (HPLC) has been employed for sample analysis[169]. The authors calculated the antibiotic burden through normalized antibiotic weight by population and WWTP flow[169].

In another study, researchers examined wastewater samples to determine levels of antimicrobial pollutants from the hospital and urban sewage in Kenya[178]. Ngigi *et al.* (2020) detected higher quantities in hospital wastewater and concluded that WBE is essential to monitor antibiotic consumption and timely notify healthcare stakeholders[178]. The authors obtained their data using LC-MS/MS[178]. The final concentration detection was based on the acidity of the 17 studied antimicrobial agents[178].

Mtetwa *et al.* (2021) examined ARGs regarding tuberculosis from WBE data from three South African WWTPs[174]. Most of the ARGs are poorly degraded[174]. In addition, the authors examined antibiotic removal efficiency for critical tuberculosis drugs[174]. The results indicated increased concentrations for individual drugs[174]. Mtetwa *et al.* (2021) explained the higher rates by the fact that the sewage processing disrupts bacteria, releasing genetic material[174]. The authors enriched the ARGs using polymerase chain reaction (PCR)[174]. Subsequently, the Kruskal–Wallis variance analysis followed by Dunn's posthoc assessment has been used to detect significant differences in ARG concentration across the sampling sites[174]. Moreover, the authors applied the Mann–Whitney U statistics to compare ARG levels in the in- and effluent[174].

Finally, Hutinel *et al.* (2019) observed a high correlation between AMR of *Escherichia coli* isolates from patients and WWTP influent, underpinning the significance of sewage for epidemiology studies[160]. In this study, the taxonomy classification has been conducted using Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI-TOF) mass spectrometry, followed by minimum inhibitory concentration (MIC) testing[160]. To statistically compare pathogen burden between the collecting points, a Welch's *t*-test has been applied[160]. Moreover, the authors used linear regression to verify correlation among hospital and WWTP samples[160].

3.2 Freshwater-based Epidemiology

Detectable concentrations of antibiotics and AMR in receiving freshwaters are critical for dissemination[100]. Concerningly, the referenced publications in the section above indicate that antibiotic residuals are verifiable in freshwaters. Larsson *et al.* (2021) highlighted recreational waters as sources for opportunistic infections and AMR, although admitting the necessity for more studies in this direction[134]. In addition, the authors suggested examining current AMR in natural ecosystems, potentially allowing inference towards local clinical relevance[134]. In this

light, Bueno *et al.* (2019) investigated the impact of aquaculture on AMR spread[27]. Specifically, the authors compared the AMR contamination of the in- and effluent of fish farms and collected samples with varying distances to the breeding station[27]. A significant finding concerns the confirmation of ARG pollution in the effluent[27]. However, the authors reported the decrease in the initial concentrations for some of the ARGs at the last sampling point[27]. These observations prove the low impact of the examined trout farming; however, they cannot rule out the AMR contribution of such facilities in general[27]. For the computational part, the authors integrated normalized gene scores based on 44 reference ARGs enriched by quantitative PCR (qPCR)[27]. The evaluation comprised linear mixed models to predict ARG concentration and Principal Component Analysis (PCA) to depict sampling site similarity[27].

Regina *et al.* (2021) examined the relation of human activity and ARG pollution using samples collected around the greater Rio das Ostras area (Brazil)[195]. The authors detected typical environmental species; however, minor levels of pathogens close to the city[195]. ARGs have been verified in all samples, whereby municipally specimens revealed increased rates[195]. In addition, the authors indicated genes encoding for *β*-lactam and carbapenem insensitivity[195]. The data integration comprised PCR for DNA enrichment, being the input for further analysis[195]. In particular, the PCR facilitated taxonomy classification using public databases and the Naïve Bayes classifier (NBC)[195]. Sample site clustering has been enabled through Principal Coordinates Analysis (PCoA)[195]. The results have been statistically revisited applying Wilcoxon signed-rank test on the taxonomy features and permutational Multivariate Analysis of Variance (MANOVA) to calculate the significance of the PCoA[195].

Schar *et al.* (2021) revisited environmental studies from the past 20 years to quantify AMR levels in Asia[203]. The authors aggregated the reported resistance as well as meta information and detected multiple AMR hot spots in Asian naval- and freshwaters[203]. Based on a machine learning (ML) model, they predicted various locations for which future monitoring is recommended[203]. Schar *et al.* (2021) obtained AMR of isolates from fish and seafood for the data integration[203]. They defined AMR with the P50 value, which is the percentage of antibiotics possessing no effect on half of the examined pathogens[203]. The authors modeled the P50 evolution of the past 20 years as a regression using generalized linear models[203]. In addition, ANOVA has been used to estimate the significance of various environmental and technical covariates[203]. AMR prevalence has been determined between the actual freshwater sample sites using a stacked ensemble approach[203]. The interpolated P50 values enabled the authors to quantify AMR for further geographic proximities[203]. In particular, Schar *et al.* (2021) used Boosted Regression Trees (BRTs) as base classifiers and Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection[203]. Finally, the coordinates of interest for further surveillance have been established by incorporating industry and population density along the geospatial P50 distribution[203].

Further researchers acknowledged the importance of computational tools to assess the

risk of ARGs. Specifically, Zhang *et al.* (2021) mined recent literature on antibiotic volume in the environment, including freshwater samples from Italy, Spain, and Sweden, among others, to illustrate the urgency of *in silico* models for ARG prevalence[262]. The authors integrated genomic and metagenomic datasets from patients and public resources as a bioinformatics pipeline to determine the ARG burden[262]. The statistical framework automatically assigns ARGs to four categories reflecting the clinical relevance[262]. The first level comprises ARGs, significantly enriched on mobile genetic elements (MGEs) from multiple ESKAPE species[262]. The authors observed 122 level one ARGs, potentially contributing to resistance events in the future[262]. To validate their results, Zhang *et al.* (2021) grouped putative level one resistance genes family-wise and found a high agreement with AMR gene families designated by the World Health Organization (WHO)[262].

Zhu *et al.* (2017) examined ARG residues at 18 transition points of river outfalls and marine waters along the eastern China coast[267]. The authors connected ARG dissemination and additional meta-information, for instance, population or industrial intensity, reflected by the gross domestic product[267]. The results demonstrated that anthropogenic activity impacts ARG pollution[267]. Human effects on AMR development are further supported by the fact that bacteria composition is similar among the investigated sites[267]. Data integration based on qPCR included AMR and 16S rRNA genes quantification[267]. Furthermore, the authors utilized 16S rRNA amplification to generate the Operative Taxonomic Unit (OTU) table for taxonomy classification[267]. Statistical analyses on ecological aspects were conducted using the R-library vegan[57,267].

Schar *et al.* (2021) independently observed that marine animals examined close to the eastern China coastland host multiple resistant bacteria[203]. Consequently, increased pathogen rates led to higher ARG levels in east Chinese sea water[203]. The observed ARG prevalence indicates that estuary AMR pollution could drive further dissemination in the ocean, potentially affecting Global Health issues[100]; however, the origin of marine AMR is not comprehensively clarified[95].

Besides estuary samples, Hooban *et al.* (2021) examined various origins across Ireland, including lakes, rivers, the sea, treated and untreated wastewater, as well as human feces[104]. The goal of the study was the detection of ARGs from clinically relevant microorganisms[104]. To this end, Hooban *et al.* (2021) isolated 211 species and verified genes encoding for resistance to various antibiotics, such as β-lactam and carbapenem[104]. The results revealed the prevalence of ARGs in all isolates, independent of the origin[104]. However, resilience to antibiotics differs according to the origin[104]. For instance, tetracycline resistance is more common in presumably clean water, whereas wastewater isolates were resistant against other antimicrobial agents[104]. Moreover, whole-genome sequencing (WGS) of *Escherichia coli* and several *Klebsiella* species exhibited high agreement of the core genome from waste- and freshwater cultures[104]. Hooban *et al.* (2021) concluded that AMR development progressed

across ecological borders and suggest ongoing AMR monitoring in the future[104]. The authors used PCR for ARG detection[104]. Isolates for WGS were selected based on the ARG content as well as a preceding antibiotics sensitivity screening[104]. Regarding the data analysis, the authors implemented a bioinformatics pipeline, conducting different steps to screen the assembled genomes for ARGs, virulence factors, and plasmid DNA[104].

Thus far, aside from the references cited above, numerous studies dealt with waste-[41,99,158] or freshwater-based epidemiology[40,161,180] to monitor AMR. The different data integration and analysis protocols demonstrate the diversity and urge for standardized approaches. Consistent measures and thorough surveillance to identify the risk for humans has been also demanded in a recent study about AMR pollution in European recreational waters[67]. Czekalski *et al.* (2015) provided a significant contribution in this direction[49]. The authors surveyed 21 Swiss lakes to measure background levels of AMR[49]. According to the study, treated wastewater and anthropogenic effects are related to sulfonamide resistance[49]. In the first step, qPCR has been utilized for antibiotic resistance and 16S rRNA gene detection[49]. Afterward, linear regression estimated the correlation between ARGs and socio-economic data[49]. The authors used the R-package vegan to statistically verify the diversity of the microbial communities[49].

3.2.1 A Multi-omics Study on Quantifying Antimicrobial Resistance in European Freshwater Lakes

To comprehensively reflect the environmental AMR diversity, many sampling sites beyond individual countries or covering even a whole continent are necessary. Since comprehensive studies are based on unified sampling and analyses protocols, temporal and quantitative conformity can be ensured. Consequently, in the first publication of this dissertation, we examined 274 European freshwater lakes to investigate ARGs, specifically encoding for resistance against four critical antimicrobials, comprising sulfonamides, tetracyclines, cephalosporin, and fluoroquinolones[221]. In addition, we related various farming facilities, for instance, animal husbandry or cereal cropping, to unveil potential correlation with AMR[221]. The results are based on multi-omic data, comprising 16S rRNA amplicon sequencing of all and the metagenomes of 39 waters[221]. First, we utilized the former technology to examine the genera[221]. Among others, we detected *Acinetobacter*, *Mycobacterium*, *Pseudomonas* amplicons, partly spanning ESKAPE species; nevertheless, non-pathogenic members are widespread in the lakes[221]. The OTU table, which describes the lakes (samples) by the abundance of bacteria (features), has been used as input for the PCoA to visualize the sampling sites in two dimensions[221]. Although taxonomy variation between the locations is seemingly low, a non-parametric MANOVA revealed a significant difference[221].

For the second part of this study, the metagenome of 39 lakes has been sequenced and analyzed[221]. Using the metagenome data, we refined the taxonomy classification from genus to species level, and verified common and clinical relevant bacteria[221]. Next, the metagenome data was used for ARG detection, carried out through the Resistance Gene Identifier (RGI) tool provided by the Comprehensive Antibiotic Resistance Database (CARD)[221]. In short, the RGI tool predicts ARG families and drug classes to which a strain at hand possesses resistance[3,221]. One German lake, located in Thuringia (N261LU), hosts bacteria with putative resistance against sulfonamides, tetracyclines, cephalosporin, and fluoroquinolones[221]. ARGs have also been identified in further German, Romanian, Italian, and French lakes[221]. Species there are potentially insensitive to tetracyclines, among others[221]. Moreover, we detected members of the TEM β-lactamase gene family for various lakes in Romania and Germany as well as lower quantities in French waters[221]. Notably, β-lactamase production confers resistance to cephalosporins[217]. An insignificant low correlation between sulfonamides and agricultural usage within 20 kilometers around the sampling points has been reported[221]. In this light, nearby farming has seemingly no effect on tetracycline, cephalosporin, or fluoroquinolones resistance[221]. However, higher ARG levels in particular countries, for instance, Germany, are following the grade of farming intensity reported by the European Union (EU)[221].

These findings indicate the progress of AMR against conventional antibiotics. Fortunately, research on alternative therapies is an active topic[190]. Qu *et al.* (2019) referred to a study that leveraged synthetic biology to confer antimicrobial abilities to host cells of mammalians[144,190]. Other researchers, as noted by the authors, investigated the role of antimicrobial peptides (AMPs)[190]. AMPs have also been covered in a recent review, reporting efficiency even on ESKAPE organisms[175]. To understand the significance of host-defense peptides (HDPs) and shed light on their mode of action and biochemical properties, particularly AMPs are covered in the following chapter in detail.

4
Host-Defense Peptides

Host-defense peptides (HDPs) are peptides that are capable of pleiotropic immune response regulation, including antimicrobial peptides (AMPs)[101,179]. AMPs are short molecules consisting of 10 to 50 amino acids, and the main modes of action are membrane disruption of pathogens and translocation into the cell[125,231]. Effectively in the year 2019, 34 AMPs were in preclinical, and 27 peptides were in the clinical phase[125]. Only one peptide has reached market maturity at this time[125]. The low amount is out of all proportion to the actual number of AMPs, namely almost 5000, possessing a great effect on pathogens and host-defense modulation[125]. The *in vivo* selectivity is hampered by toxicity, salt susceptibility, or premature degradation[125,231]. Since side effects ultimately compound translation into effective antibiotics, current research focuses on countermeasures. The greatest impact is obtained by adapting physicochemical properties of the peptide's primary sequence, such as amino acid composition and order[231]. To this end, the current chapter describes various aspects of antibiotic activity and sequence modulation in more detail.

4.1 Antimicrobial Peptides

For a peptide's antimicrobial activity, hence, membrane interaction, the amino acids sequence depends on cationic, hydrophobic, and amphipathic properties[56]. Hence, AMPs are composed of the positively charged (cationic) arginine, histidine and lysine, as well as water repellent (hydrophobic) amino acids[15]. Hydrophobicity is provided by the non-polar amino

acids glycine, alanine, proline, valine, leucine, isoleucine, methionine, tryptophan, and phenylalanine[15]. Other amino acids are the negatively charged aspartate or glutamate and those featuring hydrophilicity, including the polar serine, threonine, tyrosine, cysteine, asparagine, and glutamine[15]. In addition, the amino acid sequence of AMPs should possess an amphipathic character, hence, a non-polar hydrophobic and a polar hydrophilic face[237]. In aqueous environments, the structure of AMPs is undifferentiated; once the peptides bind to a membrane, a secondary structure formation occurs[231]. In particular, AMPs can develop α-helical, β-strand, combined $\alpha\beta$, or random coil segments[71,105]. The structure and amino acid composition are essential for membrane selectivity[125]. For instance, Deslouches *et al.* (2005) stressed the importance of arginine to increase the affinity of α-helices to hydrophobic cell walls[56].

AMPs are of a manifold and natural origin[71] and are part of the innate human immune system[31]. These peptides are responsible for diverse interactions with exogenous threats, such as microorganisms, or the regulation of the immune response[31]. Beyond human origin, the genetic information for AMPs is encoded on the DNA of various species, including vertebrates, invertebrates, or plants[229,231]. To neutralize pathogens and other microorganisms, the host responds with the synthesis of AMPs or pre-cursor proteins, which are post-translationally cleaved to mature AMPs[231]. Other modifications comprise disulfide bond formation or glycosylation, implying the augmentation of side chains with carbohydrates[13,231]. Generally, post-translational modification increases antimicrobial activity and durability of the peptide[13].

As mentioned above, the main mode of action of AMPs is the targeted aggregation on cell walls, whereby the selectivity depends on the physicochemical properties of the amino acid sequence and the cell membrane[231]. In particular, bacterial membranes are negatively charged, and eukaryotic cell walls are zwitterionic, hence, both positive and negative charged[231,237]. The membrane potential is due to different phospholipid structures[229]. Bacteria exhibit mainly phosphatidylethanolamine, and eukaryotic host cells are composed of phosphatidylcholine[229]. Different biochemical properties are essential since AMPs interact directly with the membrane wall[231].

However, the physicochemical properties also determine the toxicity of AMPs[231]. For instance, mellitin, a bee venom, and pardaxin, targeting both mammals and bacteria[202,231]. In contrast, magainins and cecropins are highly specific to bacteria[231].

Besides lipid composition, thus, membrane charge, the electric potential along the primary structure define the selectivity[231]. In particular, the cationic, positively charged AMP interacts with anionic, negatively charged cell wall[231]. Deslouches *et al.* (2005) highlighted that antimicrobial effects could be increased by directed integration of amino acids in the sequence[56]. Additionally, the study evaluated the length and ratio of helices on the antimicrobial effect[56]. Albeit sequence extension and helix propensity led to increased antimicrobial activity, the cor-

relation solely demonstrated a significant increase for up to 24 amino acids[56]. Furthermore, the authors studied the effect of AMPs containing only arginine and valine, reinforcing the amphipathic character and helicity[56]. The results revealed good antimicrobial activity, partly confirming the observation that some AMPs develop a secondary structure on membrane binding[231]. The effect could be increased specifically concerning *Pseudomonas aeruginosa* and *Staphylococcus aureus* by substituting valine with tryptophan, an important amino acid for protein-membrane interaction[55,56].

The affinity is also facilitated by particular physicochemical properties of both conjugates, ultimately defining the antimicrobial mode of action[231]. Disruption of the lipid bi-layers, pore-forming (permeability of the cell wall), and dissolution of the membrane potential, hence, disruption of ion flow, results in neutralization of the pathogen[231]. Besides cell wall disruption also translocation is known[81]. Considering that microbes have been exposed to AMPs ever since, only a few species developed resistance[71,231]. The resistance mechanisms comprise cell wall modification, proteolysis, or non-specific efflux pumps[85,231].

Nevertheless, AMPs are still regarded as highly effective antibiotic agents, and physicochemical parameters determining the membrane interaction are under continuous investigation[136]. In this light, researchers identified membrane aggregation and disruption as a two-step process[184]. First, as mentioned above, AMPs possess a random coil before binding to the membrane surface[184]. Afterward, structural development occurs on binding to the anionic half of the phospholipid[184]. The hydrophilic part interacts with phospholipids on the cell membrane and the hydrophobic section with hydrophobic carbon chains, allowing the peptide to extend into the lipid layer[231].

Papo and Shai (2005) synthesized two different peptides, only differing in the chirality of four leucines[184]. The author's objective was the examination of physicochemical properties, which determine membrane binding and permeation[184]. Papo and Shai (2005) observed that the one containing proteinogenic L-leucine, only binds to the lipopolysaccharide layer without membrane penetration[184]. In contrast, D-leucine enantiomers additionally disrupt the cell membrane[184]. Consequently, the authors suggested that different physicochemical properties are responsible for binding and insertion[184].

The findings of Papo and Shai (2005) stressed the amphipathic character of AMPs[184], hence, hydrophobic and hydrophilic segments[231]. Thus, the positively charged amino acids lysine and arginine are common in natural AMPs[231]. Besides amphipathicity, the amino acid composition also defines the secondary structure propensity[231]. Accordingly, AMPs are classified in linear peptides, possessing α-helix tendency, and non-linear peptides, including individual numbers of disulfide bonds to strengthen the β-sheet formation[231]. However, proline-rich linear peptides mitigate a helical structure aggregation, ultimately retaining random coil structure[231].

It is difficult to derive the mode of action or an antimicrobial effect from the shape since

AMPs vary in amino acids, secondary structure propensity, and biochemical features[231]. To shed light on the relation of peptide shape and particularly pore-forming, Sato and Heix (2006) examined the antimicrobial activity of the AMPs cecropin and a cecropin-mellitin-hybrid[202]. The hybrid employed positively charged N-terminal amino acids from cecropin and non-polar residues from mellitin[202]. The authors confirmed that upon α-helix formation, the hydrophobic and hydrophilic faces are located on opposite sides[202]. Moreover, experiments demonstrated that the alignment of helical AMPs occurs parallel, oblique, or vertical to the membrane surface[202]. Upon reaching the required peptide concentration, transmembrane insertion results in cell wall lysis[202,231]. More details and the description of additional modes of action follow in the next section.

4.2 Modes of Action

Different models have been observed to describe the mode of action of AMPs, including the barrel-stave, toroidal, and carpet model[202]. The mechanism of the barrel-stave model is as follows. Hydrophobic peptides permeate into the phospholipids in a parallel manner[231]. Subsequently, multiple AMPs merge to form the required peptide concentration and create the pore[202,231]. The hydrophilic face is oriented inward, hence, towards the pore core, whereas the hydrophobic side interacts with the phospholipids[231]. The hydrophobic and hydrophilic faces are oppositely oriented, thus, increasing the stability of the pores[202].

In the carpet model, the AMP head interacts with the glucosamine part of the membrane lipids, hence, the exterior of the cell wall[231]. Afterward, oligomerization of horizontal aligned AMPs to the membrane increases the force on the lipid bi-layer, eventually leading to disruption of the membrane[231].

Regarding the toroidal model, the peptide's amphipathic segments lead to the collapse of the lipid bi-layer[231]. In particular, the hydrophilic and hydrophobic moieties induce horizontal pressure on the outer membrane, therefore, forcing lipid molecules from the outer layer to interact with molecules from the inner layer, ultimately forming a pore with a toroidal shape[231]. Lipid heads and the peptide's hydrophilic face stabilize the interior of the toroid[202,231]. The cell wall lysis perturbs the cell's interior and exterior ion gradient, ultimately neutralizing the pathogen[231]. Since AMPs can also affect interior targets, more research is necessary to identify the definitive cause for inactivation[231].

Toxicity is an undesired side effect and concerns, for instance, erythrocytes[191]. The "therapeutic index" describes AMPs by their tendency of prokaryotic membrane lysis and their impact on red blood cells[264]. Findlay *et al.* (2010) pointed out that the primary structure could also indicate toxicity and antimicrobial activity[70]. In particular, the authors observed that disintegrating α-helix segments reduced hemolysis without affecting thereby bactericidal[70]. Other research specifically addressed the clinical relevance of natural AMPs, owing to

their high selectivity and broad-spectrum activity[26]. Brogden and Brogden (2011) suggested that the peptide's feature of ready optimization should be leveraged to reduce hemolysis and premature disintegration to increase effectivity[26]. Various studies investigated the role of the α-helix tendency and toxicity and concluded that helix forming is relevant for hemolytic activity[231]. However, amino acids can be substituted with proline to mitigate α-helix propensity and simultaneously retain antimicrobial activity,[140,219,231].

The hydrophobic moment also influences toxicity[231]. It is defined by the vectorized amino acid sequence with a length equal to its hydrophobicity value and direction corresponding to the helical orientation[64]. Afterward, all values are accumulated to obtain the hydrophobic moment of an AMP[231]. Researchers also observed that hydrophobicity is more critical for hemolysis than for antimicrobial activity[231]. Thus, selectivity and hydrophobicity are related, meaning that reducing hydrophobicity and retaining a constant hydrophobic moment increases selectivity without mitigating antimicrobial potential[231].

To further correlate the effect of physicochemical properties and the mode of action, Yin *et al.* (2012) developed different protocols for testing antimicrobial and hemolytic activity of synthetic peptides with varying parameters[259]. In particular, the authors adapted hydrophobicity and amino acid composition through the insertion of lysine and alanine[259]. The results indicate that hydrophobicity is essential for secondary structure formation, and β-strand aggregation was higher for more hydrophobic peptides[259]. Insertion of alanine increased hemolysis[259]. Furthermore, amino acid replacement with leucine showed similar antimicrobial potential with reduced toxicity[259].

4.3 Pleiotropic Applications

AMPs are broadly applicable to pleiotropic targets, comprising microorganisms, fungi, such as yeast, or protozoa[81]. Toke *et al.* (2005) stressed the antiviral and anticancer effect of AMPs[231]. The authors also referred to some AMPs, which can be employed for conveying active substances into the bacterial cell[231]. Cell-penetrating peptides carry molecules, usually hydrophilic compounds, to overcome the hydrophobic cell membrane[258]. Furthermore, Mader and Hoskin (2006) examined the potential of some cationic AMPs for cancer treatment[150]. Specifically, cecropins and melletin, among others, exhibit cytotoxicity against tumor cells by prevention of angiogenesis, disruption of the tumor membrane, or initiating programmed cell death (apoptosis)[150]. However, a significant challenge is premature peptide digestion due to blood proteases[150]. Encapsulation in lipid vesicles or viral vectors could improve delivery efficiency[150]. In particular, modified adenoviruses carrying the gene encoding for melittin have been used to study the effects on tumor cells[150]. In addition, anticancer peptides can support conventional treatment by alteration of the tumor cell membrane for more effective chemotherapy[150]. A recent study revealed that tumor cells contain phosphatidylser-

ine, which could explain the affinity of some AMPs[229]. However, tumor tissue can consist of different lipids, which influence the membrane charge, resulting in increased or decreased AMP attraction[229].

Furthermore, Zaiou (2007) specified antiviral, antifungal, and antiprotozoa as well as antiinflammatory, antirespiratory, and antiparadontal efficiency[261]. Antiviral activity has been reported against human immunodeficiency virus (HIV) and influenza viruses[261]. In particular, AMPs can perturb HIV's co-receptor tropism mechanism, either by inhibiting a co-receptor or binding to the viral gp120 protein[261]. Antifungal effects are due to membrane lysis and modulation of intracellular targets[261]. Multiple species from *Candida* are of interest here, owing to infections of immune-compromised patients on intensive care stations[35,261]. Moreover, CZEN-002, the dimerization of the tri-peptide *α*-melanocyte-stimulating hormone (*α*-MSH), can be administered against yeast infections of the female genital tract as soon it accomplishes clinical trial[125,209].

Antiprotozoa peptides cause cell wall disintegration, for instance, in *Trypanosoma brucei*, a parasite causing the African trypanosomiasis[120,261]. Zaiou (2007) also stressed the potential of AMPs considering antiinflammatory effects[261]. AMPs mitigate pro-inflammatory diseases, such as psoriasis or acne vulgaris[261]. In addition, researchers observed an effect of AMPs in Crohn's disease, Ulcerative colitis, or atherosclerosis development[261]. Zaiou (2007) also highlighted that multiple AMPs can be employed against respiratory tract disorders[261]. However, diseases affecting ion exchange between cell compartments are challenging, owing to the salt susceptibility of AMPs[261].

Moreover, AMPs are involved in cytokine synthesis, therefore, stimulating the innate defense system[261]. In this light, Yeung *et al.* (2011) underpinned that AMPs can improve the immune response to vaccination[258]. Human defensins, expressed in white blood cells, particularly neutrophils, can attract cytokines, stimulating endogenous defense[258]. Vaccination based on such an antigen therapy could be more effective[258].

AMPs additionally possess regulatory capabilities, impacting various immunomodulatory pathways[92]. In particular, Hancock *et al.* (2016) revealed multiple interactions of the AMP LL-37 with different genes and proteins, demonstrating the importance of this cathelicidin for metabolism[92]. Notably, the authors stated that due to enzymatic modification, the host-microbiome eludes AMPs[92]. Hilchie *et al.* (2013) suggested another class of AMPs, denoted as innate defense regulators[101]. Three groups of interactions are described: targeted anticellular activity, including microbial or tumor cell wall lysis, peptide-mediated immune response, including antigen-presenting AMPs, and immunomodulation, such as cytokine biosynthesis[101]. For instance, LL-37 binds on the targets cell surface, resulting in affinity to white blood cells and immunomodulation[101]. Hilchie *et al.* (2013) also reported that albeit some species elude cell lysis by premature peptide degradation or adapting physicochemical properties of the cell wall, host-immunoregulation is not affected[101,187]. Furthermore, AMPs ac-

tivate antigen-presenting cells to notify the adaptive immune system about the presence of pathogens[81].

In another study, Mannoor *et al.* (2010) bound AMPs on a solid phase to quantify pathogens[157]. In particular, pathogen levels are measured through the chemical attraction of the peptide's polar face to the non-polar bacteria's lipid bi-layer[157]. The application of AMPs also demonstrated potential for the food industry[26]. Hence, the attachment of AMPs to the inner surface of containers extended the content's edibility[26]. Furthermore, small vesicles prepared with AMPs can be mixed into cosmetical products to extend storage life[26].

4.4 Synthetic Modification

Physicochemical properties determine the activity and toxicity of AMPs. Toke *et al.* (2005) testified an overall positive charge for active peptides; hence, the cationic character is critical to interact with the negatively charged cell membrane[231]. The actual number of positively charged amino acids is secondary[231]. Specifically, the charge correlates only to a certain degree with interaction tendency[231]. Membrane penetration has a maximum and is independent of additional positively charged residues[231]. However, the hydrophobic moment, hence, amphipathicity, is crucial for antimicrobial activity and is, according to the authors, more critical than helicity and hydrophobicity[231].

AMPs with evenly distributed hydrophilic and hydrophobic faces tend to arrange vertically to the phospholipids, hence, parallel to the cell membrane[231]. In contrast, peptides exhibiting overall hydrophobic face form pores, that is, prefer parallel alignment to the phospholipids[231]. Peptide orientation also determines the binding phases[231]. Peptides interact with the cell wall, followed by the insertion into the membrane[231]. As mentioned above, the structure is also significant for selectivity and efficiency, which could be verified by artificially secondary structure lysis, resulting in inactivity and hindered selectivity[231].

Impeding premature lysis is essential since AMPs are prone to protease and to salt[81,231]. Low salt susceptibility is crucial at endogenous conditions, including tolerance to the physiological salt concentration[170]. AMPs can also possess undesired side-effects, for instance, bleeding disorders, modification of the innate immune response, or hemolytic activity, which must be mitigated by synthetic sequence modification[81,231]. Regarding hemolysis, Giuliani *et al.* (2007) suggested the substitution of arginine with lysine to improve selectivity to microorganisms and hamper interaction with red blood cell[81]. Nicolas (2009) underpinned the importance of arginine prevalence to support membrane translocation[179].

Brogden and Brogdon (2011) recited a variety of chemical modifications to improve AMPs[26]. These modifications include substitution with non-proteogenic amino acids or enantiomers conjugates, sequence truncation, amino acids deletion, or hybridization of AMPs[26]. The fundamental characteristics, hence, polar or non-polar moieties and hydrophobicity, are re-

tained[26]. The attachment of the side chain to the amide group instead of the α-carbon also significantly reduced susceptibility and retained antimicrobial activity[26,168]. Examples for hybridization include the peptides cecropin-mellitin or pyrrhocoricin-drosocin-apidaecin (A3-APO)[26,202].

A3-APO leverages multiple properties from its parent peptides, resulting in a dual mode of action, which involves membrane lysis and translocation[26]. Substitution of individual amino acids, including the modification of the secondary structure, is also essential[26]. Cyclization enhances the stability of AMPs, confirming the effectivity of plant-derived cyclic AMPs[26]. In terms of stability, hence, pharmacokinetics, Findlay *et al.* (2010) suggested incorporating chiral amino acids to reduce premature peptide digestion, for instance, exchanging L-leucine with D-leucine mitigates proteases[70]. Rathinakumar *et al.* (2009) observed that replacing L-amino acids with their D-enantiomers results in disorganization of the helix and simultaneously reduces antimicrobial activity against some species[191]. A further strategy to enhance activity concerns the employment of more voluminous side chains to inhibit the protease's active site[70]. Amino acid modification using fatty acids increases the interaction with the cell membrane[70].

Schmidt *et al.* (2011) narrowed down the membrane disruption progress of multiple defensins to the "sattle-spray curvature" and underpinned the role of a specific amino acid composition[205]. More precisely, the authors discovered that fewer arginines could be neutralized by additional lysines and, in general, hydrophobic residues[205]. Notably, more lysines and fewer arginines highly correlate with the overall hydrophobicity and confirm the significance of lysine for antimicrobial activity[205]. The results demonstrated that the microbial membrane, composed of negatively charged phospholipids, and choline-rich eukaryotic membranes, is crucial for selectivity[205]. The authors observed membrane bending solely for phosphatidylethanolamine-containing prokaryotic cell walls[205].

Another determinant of antimicrobial activity is the peptide length. Seo *et al.* (2012) collected multiple short AMPs with less than 12 amino acids and described their clinical relevance[209]. Remarkably, even AMPs with three amino acids possess antimicrobial activity[209]. For instance, the three amino acids long, C-terminus truncated α-MSH, still demonstrated pleiotropic effects[209]. Moreover, Mikut *et al.* (2016) designed peptide libraries of varying amino acid composition derived from a large cohort of short, active peptides for high-throughput screening[167]. The results revealed that antimicrobial characteristics, such as hydrophobic moment and positively charged amino acids, are difficult to transfer to short peptides[167]. However, the findings suggest highly active candidates with a low minimum inhibitory concentration (MIC)[167]. Findlay *et al.* (2010) noted that short AMPs are cheaper to synthesize[70].

Furthermore, Teixeira *et al.* (2012) investigated the contribution of cell membrane characteristics to AMP selectivity[229]. The authors listed the percentage of various compartments

of prokaryotic and eukaryotic membranes to elucidate the importance of the target membrane[229]. Since the interaction between the positively charged AMP moiety and negatively charged phospholipids occurs via variations in the potential, the electrostatic of the lipid-bilayer is crucial[229]. The charge also hampers interaction with eukaryotic cell membranes due to more uncharged phosphatidylcholine or sterols[229]. Teixeira *et al.* (2012) pointed out that phosphatidylserine also contributes to selectivity, and albeit the inner section of eukaryotic membranes partly consists of this lipid, AMPs would require exposition for toxicity[229].

A major obstacle is the *in vitro*, time-consuming screening for novel AMPs[77]. To tackle this challenge, Fjell *et al.* (2012) underpinned advantages using "virtual screening" techniques[71]. Researchers can employ machine learning (ML) algorithms to predict AMPs virtually. These algorithms are trained with known AMPs, for instance, derived from public databases[71]. Many ML models require a fixed-length and numeric input[137]; thus, preprocessing of the amino acid sequences is necessary. To this end, the subsequent chapter thoroughly describes encoding libraries, ML for peptide classification and the diversity of ML models utilized in this context. The ML chapter is completed by publications made as part of this dissertation.

5
Machine Learning

Various aspects of antimicrobial resistance and data integration of multi-omic resources have been presented. Specifically, environmental epidemiology utilizes next-generation sequencing data or different bioanalytical techniques, including polymerase chain reaction (PCR). Afterward, the data is used for bioinformatics and statistical evaluation. In contrast to such prospective studies, artificial intelligence retrospectively finds unknown patterns in the data. The current section continues the work of the previous chapters and illuminates various parts of machine learning (ML) workflows for host-defense peptide (HDP) classification. First, the present section describes the necessity of public peptide databases for data generation. The second part covers libraries for peptide encodings and an overview of available descriptors. The third section introduces various models and applications for ML on peptidomics datasets, including antimicrobial or cell-penetrating peptides. Since researchers applied numerous encodings on various biomedical prediction tasks, this part concludes with our encoding benchmark and the introduction of sophisticated ML models. In this light, a novel method for ensemble performance and unsupervised encoding selection is also outlined.

5.1 Databases for Antimicrobial Peptides

Sufficiently large datasets are critical *in silico* peptide screening studies. A binary classification problem is expected; hence, the dataset consists of sequences from two classes. The sequences from the positive class carry the property to be predicted, such as antimicrobial effi-

ciency. Peptides from the negative class are ineffective. A straightforward manner is to query the UniProt database[12]. Sequences annotated with, for instance, “antimicrobial” are assigned to the positive class; otherwise, to the negative class[253]. Multiple dedicated databases for antimicrobial peptides (AMPs) have been published over time. These databases enable researchers the ready acquirement of experimentally validated, for instance, biocidal, amino acid sequences[238].

Wang *et al.* (2004) established the Antimicrobial Peptide Database (APD)[243]. The latest version of the APDs is a collection of around 2,700 HDPs, including the majority being antibiotic and other peptides, comprising antiviral or antifungal effects[240]. The APDs provides biological information, such as the target species, or further details on biochemical modifications[240].

Zhao *et al.* (2013) founded the Linking Antimicrobial Peptides Database (LAMP), which contains approximately 5,500 natural and synthetic AMPs[265]. The maintainer incorporated associations between various AMP databases, linking the entries to other repositories, providing comprehensive sequence information[265]. In addition, the LAMP covers AMP cell specificity and the toxicity of AMPs[265].

Waghu *et al.* (2015) created the Collection of Antimicrobial Peptides (CAMP)[238]. The CAMP aggregates over 10,000 sequences, partly wet-lab verified[238]. A unique feature concerns the family-wise grouping[238]. In particular, the authors trained Hidden Markov Models with sequences belonging to known AMP families, for instance, cathelicidins, to categorize further peptides[238]. Moreover, the database provides structures of around 750 AMPs[238].

Jhong *et al.* (2019) developed dbAMP, a database consisting of more than 12,000 HDPs, with nearly one-third being laboratory reviewed[113]. A unique feature is the possibility to upload multi-omic data from the genome, transcriptome, or proteome for automatic AMP screening[113]. The latest update provides 26,447 HDPs[114].

The dataset construction concludes with the removal of similar sequences, as it has been conducted by Chung *et al.* (2020)[44]. Thus, the authors utilized the CD-HIT algorithm to remove overlapping amino acid sequences and to detect AMPs in various species subsequently[44]. In particular, CD-HIT counts *k*-mers to quantify the similarity between, for instance, peptides[107]. Afterward, the algorithm determines relevant sequences, hence, clusters, and assigns related ones to the respective cluster[107]. In the present context, CD-HIT is generally utilized to clean the initial AMP datasets before model training[44,58,252].

Golbraikh *et al.* (2014) introduced the Modelability Index (MODI), which describes the number of similar sequences within and across classes as a single metric[82]. Consequently, the MODI can be used to estimate the discriminability of a given dataset *a priori*[215].

5.2 Encodings

Datasets for binary classification tasks consist of amino acid sequences, represented as strings of amino acids in the one-letter code of differing lengths. The initial preprocessing step covers the elimination of similar sequences (see Section 5.1). Afterward, the workflow encompasses the sequence encoding[220]. Encoding algorithms translate the raw input to numerical sequences with equal length[220]. One differentiates among a direct mapping from amino acids to an index, for instance, using the hydrophobicity, or the incorporation of multiple amino acids, including the *k*-mer count[220]. Various packages have been published in the last years, easing the programmatic access to these algorithms and underpinning the great variety of available encodings.

Cao *et al.* (2013) implemented PyDPI, a library for the numerical description of chemical and biological sequences[30]. The packages include six amino acid encodings based on, for instance, the composition or auto-correlation[30]. In addition, the authors introduced further encodings to represent protein-protein interaction and molecule-protein interaction[30].

Ruiz-Blanco *et al.* (2015) developed ProtDCal, an interactive Java tool providing various physicochemical encodings to study the sequence-structure relationship[200]. In a recent update, the working group also provides a web-based version (ProtDCal-Suite)[199].

Furthermore, Wang *et al.* (2017) published a PSSM-based encoding framework, namely the POSSUM web service[241]. In particular, the algorithm leverages the position-specific scoring matrix (PSSM) of the Position-Specific Iterative BLAST (PSI-BLAST) algorithm as input, which represents the query peptides using the calculated evolutionary stability[4,241]. Moreover, the iFeature package by Chen *et al.* (2018) aggregated 18 encoding methods and various preprocessing algorithms, including clustering and feature reduction[39].

PyBioMed, developed by Dong *et al.* (2018), introduces further encodings for amino acid and nucleotide sequences[59]. The package offers interfaces to further sequence databases[59]. In the light of PyDPI[30], another feature of PyBioMed are encodings, which describe the interaction between macromolecules, such as DNA and proteins[59].

Li *et al.* (2019) integrated, besides various DNA, RNA, protein encodings, ML models in a comprehensive web-based tool denoted as BioSeq-Analysis2.0[142]. The developers grouped the encodings by residue- and whole-sequence-derived descriptors and introduced a novel for the former class[142]. A similar project, namely ProPythia, has been published by Sequeira *et al.* (2021)[210]. The authors implemented a user-friendly framework, which includes multiple amino acid sequence encodings for preprocessing and various ML components[210]. Additionally, ProPythia provides access to numerous Deep Learning (DL) modules[210].

Recently, Bonidia *et al.* (2021) augmented the encoding domain with a novel descriptor group, which the authors denote as “mathematical descriptors”[21]. As part of their MathFeature library, the authors added, for instance, the “complex networks” encoding, initially devel-

oped by Ito *et al.* (2018)[21,111]. This encoding represents a biological sequence as a graph, with *k*-mers being the nodes, and the edge weight reflects the *k*-mer count[111]. The final encoding is derived of various measures from graph theory, such as the betweenness centrality, which summarizes the number of shortest paths through a particular node[111]. Users can access the MathFeature package via the command line or graphical user interface[21].

The objective of the packages above is to capture the biological meanings of the input sequences involving potential interactions of non-adjacent amino acids. However, individual amino acids can also be regarded as categorical data, which is a common challenge in ML[32,189]. To this end, McGinnes *et al.* (2018) published the category_encoder library, a collection of various encodings for non-numeric scales[163]. The one-hot encoding is a straightforward example since it is simply a binary vector with the length equal to the number of instances per category, and one bit set, respectively[189].

Nevertheless, the one-hot encoding is only one proxy of a pool of many encodings. This circumstance is additionally reflected by the broad choice of encoding packages mentioned above. Some of the presented libraries also interface multiple ML models. Ultimately, researchers are faced with numerous encodings and models to tackle a biomedical classification task. Thus, in a recent review, we comprehensively examined a wide range of encodings and models[220].

5.2.1 Encodings and Models for Antimicrobial Peptide Classification for Multi-resistant Pathogens

As part of this dissertation, we reviewed various encodings, specifically for the prediction of AMPs[220]. Encodings are derived from the primary structure, or secondary and tertiary structure[220]. Thus, we classified encodings in two main groups: sequence-based encodings (SeBEs) and structure-based encodings (StBEs)[220]. In addition, model-based encodings (MoBEs) involve intrinsic, model-dependent representations[220]. MoBEs are derived either from the primary or tertiary structure[220].

SeBEs process individual or multiple amino acids at once[220]. For instance, the binary encoding or physicochemical properties are functions of single amino acids and return a binary vector or a float value per amino acid, respectively[220]. The AAindex database[119], a collection of various experimentally derived amino acid indices, can be used to retrieve physicochemical properties[220]. A drawback of a one-to-one mapping concerns peptides of varying length[220]. For this purpose, Heider and Hoffmann (2011) developed Interpol, a package that provides algorithms for normalizing the input sequences to a common length[98,220].

The second group of SeBEs summarizes multiple amino acids, which is beneficial for reflecting the interaction of non-adjacent amino acids[220]. With these SeBEs, sequences of different lengths can be conveyed in a fixed-length feature vector, eluding the necessity of in-

terpolation[220]. The amino acid composition (AAC) is a basic example[220]. Here, each amino acid is counted, and the ratio concerning the total residue number is calculated[220]. An advanced encoding first groups the amino acids using their chemical characteristics, such as the charge, before calculating the composition[220]. Various SeBEs encode the interaction of distributed amino acids statistically[220]. Essential proxies comprise auto-correlation-based encodings, which measure repeating patterns within a peptide or protein[220]. The individual amino acids are firstly represented by their physicochemical properties, followed by the actual correlation analysis[220].

In contrast to SeBEs, encodings derived from the secondary or tertiary structure (StBEs) describe the conformation of the amino acid chain in the three-dimensional space[220]. StBEs require the amino acid coordinates in advance[220]. The Protein Data Bank (PDB) is a central repository for peptide and protein structures, which also enables the retrieval of AMP structures[29]. The CAMP additionally provides access to hundreds of AMP conformations[238]. The peptide structure is employed for calculating various encodings, for instance, describing the secondary structure content or fold propensity[220]. Another example is the electrostatic hull, which encodes electrostatic properties at distinct points across the solvent-accessible surface[146,220]. This surface is the reference to determine the final feature vector using various distances[146,220].

A further StBE is the Delaunay triangulation[220]. First, the coordinates of all amino acids, specifically of the C_α-atoms, are used as the triangulation constraints, which states that the circumscribed sphere of four connected atoms, hence, a tetrahedron, must not contain additional points[220]. If the condition is fulfilled, the respective amino acids are connected via edges[220]. Subsequently, the Delaunay triangulation is passed to several aggregation methods[220]. For instance, the average distance calculates the average length between two individual amino acids, ultimately used as feature[220].

We concluded with an overview of additional encodings[220]. One example is the chaos game representation (CGR), which depicts the primary sequence as two-dimensional images[220]. Recently, Löchel *et al.* (2021) also highlighted the broad applicability of this encoding[147], including constraint resolution towards synthesis and sequencing in the field of DNA storage[148].

We also listed several models, for instance, the Random Forest classifier (RFC), which has been successfully employed for the prediction of AMPs[220]. Support Vector Machines (SVMs) are a special case since the default kernel can be used for classification and custom kernels[220]. An example for the latter is the string kernel, a MoBE, which measures the similarity between two amino acid sequences[220].

Although the review explicitly focused on AMP prediction, amino acid encodings and ML models have been applied in many biomedical domains. Thus, more details on applications are provided in the following section.

5.3 Biomedical Applications

The classification of AMPs is crucial; however, encodings enable general peptide classification tasks. Specifically, Section 4.3 provides an overview of the pleiotropy of HDPs. In the following, selected biomedical studies on ML and applications are introduced, highlighting peptides' broad applicability and the importance of encodings.

Lee *et al.* (2016) employed various encodings to investigate the cell membrane activity of AMPs[138]. The employed encodings comprise dipeptide composition (DPC), peptide charge, and auto-correlation-based algorithms, among others[138,220]. In the first step, the authors trained an SVM model to predict novel, active AMPs[138]. To reduce the search space, Lee *et al.* (2018) included other properties, for instance, sequence similarity or α-helix propensity[138]. Selected peptides have been synthesized to examine membrane affinity experimentally[138]. Lee *et al.* (2016) also confirmed the "sattle-spray curvature" formation (see Section 4.4)[138,205]. Finally, the authors applied the developed algorithm to determine membrane interaction and predicted several peptides with significant activity, including neuropeptides[138].

Gupta *et al.* (2017) examined interleukin-17 (IL17) inducing effects of peptides[88]. Specifically, IL17 is a cytokine naturally involved in the host defense; however, over-expression results in various autoimmune diseases[88]. First, the authors collected sequences, specifically epitopes, hence, amino acid patterns located on antigens[15], with and without an IL17-inducing effect[88]. Afterward, the authors employed the AAC, DPC, and amino acid pairs (AAP)[88]. The AAP encoding is based on the DPC, whereby the dipeptides are additionally weighted using their frequency in the dataset[88]. Finally, the authors deployed an SVM, trained on the DPC-encoded dataset due to superior performance[88].

Simeon *et al.* (2017) investigated the effect of various composition-based encodings for HDP prediction[215]. In particular, the authors employed the AAC, DPC, and composition/transition/distribution-composition (CTDC) encodings for the subsequent identification of crucial sequence components[215,220]. The CTDC encoding, a special case of reduced AAC descriptors, groups amino acids using their physicochemical properties and returns the percentages per group as the final feature vector[220]. Important features have been constituted with the biological meaning, including the significance of threonine in antibacterial peptides, due to the involvement of this amino acid in manifold host-defense processes[215]. The final algorithm is denoted as PepBio and uses C4.5 decision trees and the RFC as models[215].

In contrast, Fuchs *et al.* (2018) encoded peptides on the molecular level using quantitative structure-activity relationship (QSAR) descriptors[76,220]. The goal of the study was the prediction of lipophilicity[76]. This characteristic describes the uptake efficiency of drugs and is specifically important in pharmaceutical research[108]. The authors utilized two ML regression models, namely SVM regression and Least Absolute Shrinkage and Selection Operator (LASSO)[76]. The final output is based on the average prediction of the SVMs using the

feature subsets obtained by the LASSO and Principal Component Analysis (PCA)[76]. Notably, Fuchs *et al.* (2018) applied their model to successfully estimate the lipophilicity of several newly synthesized peptides[76].

Grisoni *et al.* (2018) studied the efficiency of cancer-toxic peptides through DL[84]. The authors encoded the peptides through the binary encoding[84]. In particular, the initial amino acid sequences have been transferred into *k*-dimensional binary input space, where *k* is the length of the longest peptide[177]. The model was pre-trained on generic peptides fulfilling properties of anticancer peptides (ACPs)[84]. Afterward, *in vitro* verified ACPs were used for the conclusive training, and the final model was employed to generate novel ACPs[84]. Compared to the experimental data, the predictions possessed similar physicochemical properties[84]. Agreeing overlapping characteristics demonstrated the generative capability of the applied model[84]. Ultimately, 12 peptides have been synthesized, of which six showed high *in vitro* selectivity in breast cancer cell lines[84].

In this light, Shoombuatong *et al.* (2019) developed an online-accessible tool, namely TH-Pep, to predict ACPs[213]. The authors implemented multiple RFCs employing AAC, DPC, and the pseudo amino acid composition (PAAC)[213]. The PAAC utilizes the AAC and autocorrelation of non-adjacent amino acids using a distance of λ[220]. The hybrid model trained with the AAC and PAAC encodings achieved the highest accuracy[213]. Additionally, the authors observed that tryptophan is an import feature for ACP classification[213]. The importance of tryptophan is following Simeon *et al.* (2017)[215] and other studies[213].

Another study concerning ACPs has been conducted by Gabernet *et al.* (2019)[78]. The authors collected the ACPs from the CancerPPD database[232], and peptides possessing a known secondary structure from UniProt[78]. As sequence representation, QSAR-derived molecular features have been employed[78,220]. First, RFC and SVM models have been used for classification[78]. Afterward, the authors verified putative ACPs and confirmed that around four-fifth of the synthesized peptides have *in vitro* activity[78]. Furthermore, Gabernet *et al.* (2019) implemented an evolutionary algorithm to ameliorate selectivity[78]. The resulting peptides demonstrated increased affinity towards tumor cells and mitigated toxicity[78].

Manavalan *et al.* (2019) leveraged artificial intelligence to classify peptides possessing hypertension decreasing effects[153]. Hypertension causes cardiovascular diseases and is widespread among the population; thus, additional treatment options are of great interest[153]. The authors employed various encodings to represent the peptides in a machine-readable format and to enable the training of multiple models[153]. Besides the AAC, DPC, and the CTDC encoding, the authors employed physicochemical properties of the amino acids and binary transformation of the N- and C-terminus[153]. Additionally, the authors included the "overlapping property" encoding[153]. Briefly, each amino acid is depicted as a binary vector of length 10, whereby each bit represents a certain biochemical property[153]. Amino acids potentially share attributes; thus, properties are overlapping[153]. In contrast, the "twenty-one-bit"

encoding prohibits amino acids in the same group; therefore, seven physicochemical properties partitioned in three groups result in 21 features per amino acid[153]. The final prediction ensues from the mean probability of several base models[153]. Peptides exceeding the 0.44 threshold are putative antihypertensiv[153].

Furthermore, Armenteros *et al.* (2019) implemented a DL algorithm to detect signal peptides, which define the affinity of proteins to intracellular targets[6]. The authors encoded the sequences using the BLOSUM62 scoring matrix[6]. This matrix reflects the probability of ineffective sequence alterations within a certain period[220]. Armenteros *et al.* (2019) utilized the final model to examine amino acids with significant contribution to signaling cascades[6]. The authors revealed that an alanine at the second position of the peptide is an essential, particularly for chloroplast targeting[6].

Damiati *et al.* (2019) employed ML to examine the interaction of cell-penetrating peptides and model membranes[52]. The authors trained an Artificial Neural Network (ANN) featuring the amino acid count, sequence length, weight, charge, hydrophobicity, and composition of hydrophilic amino acids, as well as the examined membrane[52]. The final model revealed a high agreement between the observed and predicted membrane compression[52]. A feature importance survey indicates that tryptophan and less hydrophilic amino acids are significant contributors to membrane interaction[52]. According to the authors, this is due to the relevance of the molecular weight and low hydrophilicity for cell penetration[52].

Wei *et al.* (2020) developed another framework to predict signal peptides, in particular focusing on quorum-sensing peptides (QSPs), essential for subcellular interactions[245]. The authors utilized several encodings, such as the CTDC[220], similar encodings as Manavalan *et al.* (2019)[153], and additional ones[245]. The g-gap dipeptide composition (GPC) is based on the DPC and comprises non-adjacent dipeptides[245]. Moreover, the adaptive skip dipeptide composition calculates the auto-correlation between adjacent and non-adjacent amino acid pairs[245]. The 188-bit encoding combines the AAC and CTDC encodings using various physicochemical properties[245]. The authors trained RFCs for each encoding class and fused the predictions using stacked generalization[129,245]. Thus, the final feature vector depicts the QSPs as a binary vector[245].

Yamashita *et al.* (2020) examined the impact of amino acid substitution on the inhibitory activity of enzymes[257]. The concerning peptides inhibit critical enzymes, for instance, *α*-amylase and *α*-glucosidase, involved in digestive processes[257]. The authors generated the initial dataset by substituting amino acids on various positions and experimentally verified the efficiency[257]. Subsequently, amino acid-, and sequence-wise, physicochemical properties have been utilized for the encoding[220,257]. Finally, Yamashita *et al.* (2020) used a Random Forest regressor (RFR) to predict peptides with increased inhibitory activity[257]. The authors observed that although the expected activity was higher, a molecular docking-based validation demonstrated still improved efficiency[257].

Charoenkwan *et al.* (2021) developed a DL model to classify the bitterness of peptides[36]. The authors motivated their research with the significance of drug flavor since bitter agents could enhance aversion[36]. For the DL part, the authors used a Word2Vec-derived method to encode the sequences[36]. Word2Vec embeds sentences; hence, amino acid sequences in the present case, in a vector space, retaining the context[36,166]. However, the final model uses the "bidirectional encoder representation from transformers" method, which can capture a more fine-grained representation of the context[36]. The authors encoded the sequences with, for instance, AAC and amino acid indices[36]. The encoded dataset has been utilized for training further models[36]. Nevertheless, the DL algorithm, namely BERT4BITTER, revealed superior performance[36].

Janairo (2021) published a study concerning the prediction of antioxidative peptides[112]. Antioxidative agents are essential to alleviate oxidation, resulting in cell damage due to free radicals[112,214]. The author described antioxidative tripeptides employing various SeBEs and StBEs[112,220]. The encoded sequences are subsequently used to train an SVM regression model[112]. The model using the BLOSUM encoding achieved the highest accuracy[112]. Janairo (2021) highlighted the significance of *in silico* models for more environment-friendly peptide synthesis since manufacturing is targeted[112].

Shen *et al.* (2021) classified antioxidative peptides[212]. In this study, the authors used the PAAC in combination with a motif count scheme[212]. The algorithm searches the peptides with and without antioxidative activity for characteristic motifs[212,236]. Besides the actual amino acids, the motif search also considers various physicochemical features[236]. The predicted pseudo probabilities are increased, if a motif is present[212]. Albeit the authors claim that the motif encoding resulted in higher performance, validation results on an independent test set indicated non-significant differences[212]. However, the final model was used to predict novel peptides[212]. The top peptides have been synthesized and tested, revealing one with good antioxidative activity[212].

In another study, Manavalan *et al.* (2021) compared models from different working groups to predict anti-severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) peptides and epitopes[152]. The majority of the applied encodings have been reviewed by Spänig and Heider (2019)[220], including multiple SeBEs, for instance, auto-correlation, and various StBEs, such as distance distribution[152,220]. The authors encoded the amino acid sequences to create a benchmark dataset[152]. Using this dataset, all classifiers revealed poor performance, as only one model exceeded the Matthews correlation coefficient (MCC) of 0.3[152]. Although the developers of the evaluated algorithms utilized various encodings, the association of the encoding selection and model performance is not discussed.

The prediction of active peptides depends on *in vitro* or *in vivo* screening results. ML requires verified data, hence, the true label; likewise, artificial intelligence drives novel hy-

potheses, ultimately experimentally tested and reassessed[139]. This self-reinforcing cycle of wet-lab results, ML, and peptide synthesis generates more and more data, bridging the gap between biology and computer science. A critical component concerns artificially manufactured amino acid chains, with the solid-phase peptide synthesis being an important proxy of these methods[172].

However, "difficult sequences", resulting from the aggregation of non-adjacent amino acids, impair the synthesis[45]. Since determinants for aggregation events are unclear, ML models could be developed to enable prediction[171]. To this end, Mohapatra *et al.* (2020) conceived a DL-based framework to estimate the synthesis performance[171]. Specifically, the authors utilized the peptide sequence, the amino acid to be coupled, and the parameters of the synthesis machine as input[171]. The fingerprint encoding, binary-encoded molecular properties similar to the QSAR descriptor, has been used to describe the amino acids[171,197,220]. The binary encoding has been employed for categorical values[171]. The target score is the amplitude of the ultraviolet signal per reaction, hence, the coupling success[171]. The final model accurately predicted difficult couplings; thus, allowing improved peptide synthesis in the future[171].

5.3.1 Multivalent Binding Kinetics Resolved by Fluorescence Proximity Sensing

Mohapatra *et al.* (2020) created a model, which can be applied to optimize peptide synthesis in various domains[171]. Targeted synthesis decreases time and material overhead; thus, environmental pollution and expense[112]. In this light, a further study contributed to this dissertation stressed the significance of computer-aided predictions[206]. Specifically, we examined the binding affinity of peptides, modified with chemical conjugates[206]. It has been verified how different architectures, hence, di-, tetra-, and octamer linkers, could reinforce the interaction with the target protein[206]. The results demonstrated a positive correlation between linker count and the binding rate[206].

We used the modified peptides and binding affinities to train an ML model[206]. The amino acid sequences have been encoded with the AAC and the different architectures through binary encoding[206]. Although the RFR revealed good performance, more research is necessary to study the effect of the amino acid composition and diverse architectures[206].

The current chapter highlighted the great variety of encodings and applications in biomedical domains. Therefore, we gathered 48 encoding groups and 50 biomedical datasets and conducted a large-scale encoding comparison study as part of the thesis[223]. The experiments illuminated potential application-dependent encoding performance[223]. More details on this study will be addressed in the following section.

5.3.2 A Large-scale Comparative Study on Peptide Encodings for Biomedical Classification

Researchers implemented various peptide encodings for diverse applications. These biomedical domains include antimicrobial, antiviral, or signal peptides classification. Peptides possess additional targets, such as oxidation or hypertension regulators. However, encodings are applied without reasoning the selection more detailed. Thus, it is unclear whether certain encodings should preferably be used on certain biomedical applications[223]. Consequently, we conducted a comprehensive benchmark comprising various encoding groups and datasets from manifold domains[223]. We collected 50 datasets to evaluate 48 encoding groups divided into two groups: SeBEs and StBEs[220,223].

StBEs are based on the secondary or tertiary structure. Thus, structural data is required for the encoding algorithm; however, conformational information is unavailable for many peptides. To this end, researchers developed various structure prediction algorithms[186]. Due to the lengthy computation, embedding in a high-throughput framework is impractical[223]. Therefore, we presented a novel algorithm, which can approximate the tertiary structure[223]. The Basic Local Alignment Search Tool (BLAST)[4] has been employed to screen a database with known peptide and protein structures[223]. The best match on a sub-structure is extracted and returned as the tertiary structure approximation[223].

A further concern is the large number of encoded datasets[223]. The pseudo k-tuple reduced amino acids composition (PKRAAC), which accepts five parameters, contributes significantly[223]. The PKRAAC is based on the PAAC encoding and measures the auto-correlation between gap-divided *k*-mers or within a specific window using a reduced amino acid alphabet[39]. To obtain a representative encoded dataset, we calculated the correlation between each dataset and amino acid alphabet[223]. Afterward, this information is used to obtain a distance matrix to project the datasets into a two-dimensional space[223]. The representative dataset is the center of the cluster (medoid)[223]. In this way, the initial number of encoded datasets could be substantially reduced[223].

For the actual benchmark, RFCs are trained on the remaining datasets to calculate several metrics[223]. The metrics, for instance, MCC and sensitivity, are later utilized to rank and cluster the encoding groups[223]. However, we implemented various additional statistics to compare the encodings not only performance-wise[223]. The diversity indicates the agreement of the classifier outputs; hence, to what extent encodings resemble on their predictions[223]. The critical difference detects significant different classifier outputs[223], and has been initially developed to compare multiple classifiers on multiple datasets[201]. We applied it to compare multiple folds of multiple encoded datasets[223]. The adjusted RV-coefficient has been employed to calculate the correlation between encoded datasets, specifically within encoding groups and contiguous parameters[223]. This statistic has been originally introduced to calcu-

late the correlation among *k*-dimensional omics datasets, further aggravated due to distinct *k*'s per dataset[162].

We observed superior performance of the SeBEs[223]. Concerning the class ratio, the misclassification rate of balanced is lower than imbalanced datasets, which is also reflected by the ranks[223]. Only a few encodings, mostly SeBEs, are continuously among the best[223]. Multiple SeBEs and StBEs are rarely or never top-ranked[223]. Notably, the QSAR encoding demonstrated superior performance on the human immunodeficiency virus (HIV) datasets[223]. The clustering confirms the distinction of SeBEs and StBEs; however, a clear separation of biomedical domains is not visible[223]. The HIV and some AMP datasets are exceptions[223]. Also noticeable is the relationship of encoded datasets originated from the same group[223]. In particular, contiguous parameters yield similar performance[223]. This observation is additionally confirmed by the adjusted RV-coefficient, which revealed high similarity within encoding groups and the non-critical difference of the respective classifiers[223]. Moreover, unrelated encodings, including SeBEs and StBEs, indicate higher diversity[223].

Albeit we could not detect encodings superior on a particular biomedical domain, we proposed instructions for encoding selection[223]. Researchers should prefer SeBEs, due to faster computation[223]. Furthermore, we recommended SeBEs, which are more common among the top ranks[223]. In addition, it is advisable to limit the parameter space, owing to the high similarity of parameterized encodings[223]. Finally, we referred to Kunchevas' (2014) general[129] and Spänig and Heiders' (2019) specific[220] advice to fuse the predictions of base classifiers[223]. Ensemble classifiers mitigate weaknesses and reinforce advantages of individual classifiers[129]. Ensemble techniques are already widely applied in the domain. Consequently, the following section introduces selected examples in more detail and emphasizes the contribution of various encodings.

5.4 Base Models and Ensemble Classifiers

Various base models have been mentioned in the previous sections, including the Decision Tree classifier (DTC) or the SVM. The term "base" reflects that the predictions are aggregated in a meta-model, potentially improving the performance of the first-layer predictors[129]. Meta-models or ensemble classifiers correct misclassified training examples from the first layer. Ideally, some base models will predict correctly, whereas others will fail on respective instances[129]. Thus, an ensemble of three base models using the majority vote as fusion method will assign the correct label if at least two predictions agree with the actual class in the basis layer[129]. This example also demonstrates the importance of the diversity of the individual predictors[129]. If base models agree on their output, the ultimate performance will not improve. Although diversity is crucial, researchers should consider other metrics to find optimal first-layer models[129].

Combining different encodings to incorporate multiple amino acid sequence properties as ensemble classifiers is an active research topic. For instance, the RFC, an ensemble of DTCs, has been extensively applied[78,215,245]. Nevertheless, researchers developed various other approaches. For instance, Löchel *et al.* (2018) developed a classifier to predict co-receptor tropism of the HIV subtype A[146]. The model combined two base classifiers using stacked generalization[146]. In particular, Löchel *et al.* (2018) trained the first model with physicochemical properties of the primary structure[146]. The second model deployed the tertiary structure, numerically encoded by the electrostatic hull (see Section 3.2.1)[146]. The authors trained the meta-RFC via the predicted probabilities of the first-layer classifiers[146]. Although the ensemble performed, regarding the area under the curve (AUC), worse than the base models, the authors observed improved sensitivity and specificity under the partial, thus, diagnostic relevant, AUC[146,149].

Singh *et al.* (2019) developed a further ensemble method to predict HIV protease cleavage sites[216]. The HIV protease is a critical drug target since it is involved in the virus replication[216]. The algorithm is based on multiple SeBEs, for instance, amino acid and dipeptide composition[216]. Moreover, the authors used SVMs with four different kernels as base classifiers, and each model is trained on individually encoded datasets[216]. Furthermore, Singh *et al.* (2019) implemented a genetic algorithm consisting of genes, encoding for each of the four SVM types, the presence or absence, and the weight of a base model[216]. Hence, the first layer ranges from one to altogether 56 base classifiers, which are selected and improved over time using various genetic operators, such as mutation[216]. The final output results from a weighted majority vote; hence, a modification of the ordinary majority vote, which weights the respective predictions before voting[216]. According to the authors, automatic weight assignment emphasizes efficient models and encoding combinations, ultimately confirmed by the high accuracy[216].

Zhang *et al.* (2020) presented a multi-stage AMP ensemble classifier[263]. The architecture enables the first distinction between sequences possessing an antimicrobial effect and non-AMPs[263]. The second stage evaluates the specific subcategory, for instance, antiviral or antifungal[263]. The authors employed a classifier chain[263]. For each label, a separate binary predictor is trained; however, the predictions of the first classifier are used to train the next and so forth[194]. The majority vote aggregates the predictions of the individual chains, ultimately denoted as Ensemble Classifier Chains (ECCs)[194,263]. Zhang *et al.* (2020) used different tree-based models, leveraging the PSSM encoding and the propensity of amino acid interactions, as base classifiers[263]. The second-stage ECC exploits an over-sampling technique to tackle skewed class distribution. The proposed workflow outperformed various state-of-the-art methods[263].

Fu *et al.* (2020) assembled a cell-penetrating peptides predictor using stacked generalization[75]. The predictions of the base models are combined using a meta SVM[75]. The first-layer

classifiers included tree-based models, SVM, and k-nearest neighbors[75]. The individual classifiers are trained with the interaction energy of the amino acids[75]. Before the second-layer stacking, Fu *et al.* (2020) examined the performance of the base models, whereby the SVM revealed the best performance[75]. However, the ultimate meta-model was significantly better[75].

Ren *et al.* (2021) applied the CGR to encode genome sequences for antimicrobial resistance (AMR) prediction[196]. The study was based on whole-genome sequencing (WGS) data from hundreds of clinical *Escherichia coli* isolates and the respective drug sensitivities[196]. The great advantage of the CGR encoding is the broad-applicability on multiple sequence types, such as proteins or DNA in the current study[147]. Ren *et al.* (2021) trained an SVM, Logistic Regression classifier (LRC), ANN, and RFC with the encoded datasets[196]. The authors observed that ANNs and RFCs predicted susceptibility to various antimicrobials with improved performance[196]. According to important features, specific single nucleotide polymorphisms are relevant for AMR[196].

Guo *et al.* (2021) created an ensemble classifier to predict HDPs[86]. To this end, the authors collected peptides, covering multiple applications, for instance, antibacterial or -viral activity[86]. Afterward, several encoding algorithms have been utilized to transform the amino acid sequences[86]. The authors trained SVMs and RFCs on one encoded dataset, respectively[86]. The probability scores of the 18 base models are used as input for a genetic algorithm, which optimizes the contribution of the individual classifiers[86]. The final probability is the sum of the individually weighted predictions[86]. Compared to the gold standard, Guo *et al.* (2021) reported a slight AUC improvement on the training set; however, the difference on the test data is minor[86].

Finally, Xu *et al.* (2021) combined multiple DL algorithms to predict immune-modulating peptides[255]. The models comprised ANNs and Convolutional Neural Networks (CNNs)[255]. The dataset also contained the sequences of the accompanied T-cell receptor to enhance the performance[255]. The first layer predicted the interaction probability of epitopes with the receptor's α- or β-chain[255]. The average likelihood of all base models is the final score[255]. Based on the results, the model determines combined $\alpha\beta$-chain-binding[255]. By applying this algorithm on independent test data, Xu *et al.* (2021) demonstrated high AUC and accuracy[255]. The binary encoding, physicochemical properties, and PCA features have been utilized as input[255]. For the latter, the authors reduced the concatenation of all available amino acid indices from AAindex database[119] to the first principal components[255].

In summary, it could be demonstrated that ensembles are widely adopted, and the performance is superior to single classifiers. Additionally, the vast encodings underlying the meta classifiers reveal the abundance, thus, the complexity of end-to-end ML workflows. To pave the way for unsupervised pipelines, we conceived a tool that allows an automatic encoding selection and ensemble configuration. This work continues the initial groundwork on peptide

encodings[220,223] and will be introduced subsequently.

5.4.1 Unsupervised Encoding Selection through Ensemble Pruning for Biomedical Classification

Researchers have conducted tedious work to find the best encoding and model composition. Specifically, the hyperparameter search space is drastically increased due to the exploration of optimal encodings, models, and parameter configuration. To ease basic hyperparameter optimization, Feurer *et al.* (2015) presented auto-sklearn, a framework for automated ML[69]. The user solely passes a dataset, auto-sklearn conducts automatic hyperparameter configuration, and a trained model is returned[69]. Various other publications underpinned the importance of this research, for instance, Hyperopt, Auto-WEKA, and TPOT, all pursuing different approaches or bindings to ML libraries[16,182,230].

Nevertheless, these tools assume a single dataset as input. Scientists in the biomedical area are challenged with multiple encoded datasets derived from identical amino acid sequences. The encodings alone contribute primarily to the search space, even after filtering[223]. Furthermore, ensembles using diverse encoded datasets are potentially superior[220,223], which follows Kuncheva (2014), who suggested diverse base models to minimize the prediction error[129,159]. Optimal encodings and classifiers approach a theoretical boundary[129,159], which we leveraged for an unsupervised encoding selection[222]. The algorithm determines the best encodings, base models, and fusion methods[222]. As a proof-of-concept, we collected ten datasets from distinct biomedical domains, four base models, and three aggregation functions[222]. The base models encompass the Naïve Bayes classifier (NBC), LRC, DTC, and RFC, whose output is fused by majority voting (hard voting), averaging (soft voting), and stacked generalization (stacking)[222].

Kuncheva (2014) referred to multiple methods to select the optimal ensemble size for one dataset[129]. “Sequential forward selection”, as an example, successively adds one base model, provided the overall performance increases[129]. Such approaches become computationally challenging considering 100 to 200 encoded datasets[223]. A promising alternative is based on the kappa-error plot, which represents classifier pairs by their diversity and average error in a two-dimensional space[129]. Here, the convex hull and Pareto frontier pruning can be utilized[129] to mitigate the computational complexity.

In addition, we implemented an optimization algorithm, which is based on the multi-verse paradigm[2,222]. The multi-verse optimizer (MVO) receives a binary vector, where each bit describes the presence or absence of a base model trained on a specific encoded dataset. Hence, the *i*-th bit indicates whether the *i*-th encoding is part of the ensemble[222]. We also added the best and random base models for comparison besides the convex hull, Pareto pruning, and the MVO[222]. The best base classifiers are selected owing to the lowest error,

and the random models are distributed across the kappa-error diagram[222]. We also introduced the extended critical difference chart[201], which depicts the average performance and a statistical comparison of pruning methods and ensemble classifiers[222].

In conclusion, we developed a workflow that enables high-throughput ensemble generation and unsupervised encoding selection[222]. Specifically, the case study involved four base classifiers and three fusion methods created by five distinct ensemble selectors[222]. The results demonstrated that the ensembles improved individual performance[222]. The RFC, an ensemble per se, already performed good on single encoded datasets[222]. The performance gain as base classifier of ensembles is negligible[222]. The RFC saturation is also reflected by a relatively compact distribution in the kappa-error chart[222]. The Pareto frontier pruning creates the most efficient ensembles[222]. The pipeline follows an extensible design pattern; thus, users can add other base classifiers and ensemble methods[222]. The visualizations and evaluation will scale accordingly[222]. However, only base model-independent fusion methods are supported out-of-the-box in the current version[222].

6 Publications

Following the background part, the current chapter comprehensively presents the publications of this thesis. In particular, each section begins with an extended abstract, followed by the corresponding manuscript. The first study concerned the examination of multi-omic datasets from European freshwater lakes[221]. The objective was to detect baseline levels of antimicrobial resistance (AMR)[221]. The second publication discussed various encodings and models to describe and predict antimicrobial peptides (AMPs)[220]. The objective of the third publication was the evaluation of multiple encoding groups in additional biomedical areas[223]. The fourth article covered the prediction of the peptide-protein binding affinity[206]. Finally, the fifth publication addressed the unsupervised encoding selection leveraging ensemble pruning[222].

6.1 A Multi-omics Study on Quantifying Antimicrobial Resistance in European Freshwater Lakes

The World Health Organization (WHO) designated AMR as a significant thread for modern healthcare systems[110]. To prevent the estimated ten million deaths by 2050, the WHO recommended effective countermeasures immediately[110]. Several studies surveyed AMR employing wastewater-based epidemiology (WBE), specifically from hospital effluent[102]. Although many countries require hospital wastewater processing, resistant bacteria are verifiable[102]. Hendriksen *et al.* (2019) analyzed untreated sewage of 79 sample sites of 60 countries[99].

Albeit the authors found no relation between antibiotic dispensation and AMR prevalence, they identified sanitation standards to have a significant impact[99].

Cleaned wastewater ultimately enters rivers and lakes, potentially leading to further AMR dissemination[100]. Czekalski *et al.* (2019) conducted freshwater-based epidemiology (FBE) on 21 Swiss freshwater lakes to relate human activity to background levels of AMR[49]. The authors examined whether genes encoding for resistance against various antibiotics are detectable in the sampling sites, and verified treated sewage as source for AMR specifically against sulfonamides[49].

WBE and FBE are crucial for monitoring AMR[110]. Thus, we collected samples from 274 European freshwater lakes and conducted a comprehensive multi-omics study on AMR. We employed 16S rRNA amplicon sequencing for a taxonomic overview in the first step. Subsequently, metagenomes from 39 lakes have been analyzed to quantify and allocate AMR genes across Europe. We focused on resistance genes against the following essential human and veterinary drug classes: sulfonamides, tetracyclines, fluoroquinolones, and cephalosporins[63].

The amplicon analysis revealed various genera in the samples, including *Acinetobacter*, *Pseudomonas*, and *Mycobacterium*. However, predominantly nontuberculous mycobacteria are common in freshwater ecosystems[198]. Afterward, we utilized Operative Taxonomic Units (OTUs) for a Principal Coordinates Analysis (PCoA) to detect differences in the sampling sites. A non-parametric Multivariate Analysis of Variance (MANOVA) verified that sample sites differ significantly. The examined metagenomes confirmed the initial taxonomy assignment. Common freshwater genera are widespread, such as *Limnohabitans*[90]. Subsequently, we focused on putatively pathogenic genera and identified *Clostridium*, *Staphylococcus*, and *Corynebacterium*, among others.

We quantified AMR in Germany, Austria, and Romania, considering different drug classes and gene families of putative pathogenic genera. Certain samples indicate resistance against sulfonamides, tetracyclines, and fluoroquinolones, whereas cephalosporins were detected at higher rates. A low and insignificant correlation with animal farming nearby could be solely observed for sulfonamides. Further research is necessary to include additional anthropogenic factors for AMR besides livestock farming.

A large proportion of reads could be assigned to multiple gene families, such as TEM beta-lactamase[24] or elfamycin resistant EF-Tu[103]. However, bacteria acquire AMR differently, which is not reflected in the analysis. Precisely, the origin of reads, hence, whether the AMR gene is located on the chromosome or plasmid, or the distinction between intrinsic and acquired resistance is not considered[47,188].

The results provide a baseline reference to monitor AMR in Europe and other countries. Considering in particular China or India, where antibiotic abuse led to an advanced AMR

contamination in lakes[33,124], the findings enable healthcare officials to execute appropriate measures timely.

Environment International 157 (2021) 106821

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

A multi-omics study on quantifying antimicrobial resistance in European freshwater lakes

Sebastian Spänig [a], Lisa Eick [a], Julia K. Nuy [b], Daniela Beisser [b], Margaret Ip [c], Dominik Heider [a,1,*], Jens Boenigk [b,1]

[a] Department of Bioinformatics, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany
[b] Department of Biodiversity, University of Duisburg-Essen, Universitätsstr. 5, Essen 45141, Germany
[c] Department of Microbiology, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China

ARTICLE INFO

Handling Editor: Frederic Coulon

Keywords:
Pathogens
Antimicrobial Resistance
Multi-Omics
European Freshwater Lakes

ABSTRACT

The surveillance of wastewater for the Covid-19 virus during this unprecedented pandemic and mapped to the distribution and magnitude of the infected in the population near real-time exemplifies the importance of tracking rapidly changing trends of pathogens or public health problems at a large scale. The rising trends of antimicrobial resistance (AMR) with multidrug-resistant pathogens from the environmental water have similarly gained much attention in recent years. Wastewater-based epidemiology from water samples has shown that a wide range of AMR-related genes is frequently detected. Albeit sewage is treated before release and thus, the abundance of pathogens should be significantly reduced or even pathogen-free, several studies indicated the contrary. Pathogens are still measurable in the released water, ultimately entering freshwaters, such as rivers and lakes. Furthermore, socio-economic and environmental factors, such as chemical industries and animal farming nearby, impact the presence of AMR. Many bacterial species from the environment are intrinsically resistant and also contribute to the resistome of freshwater lakes. This study collected the most extensive standardized freshwater data set from hundreds of European lakes and conducted a comprehensive multi-omics analysis on antimicrobial resistance from these freshwater lakes. Our research shows that genes encoding for AMR against tetracyclines, cephalosporins, and quinolones were commonly identified, while for some, such as sulfonamides, resistance was less frequently present. We provide an estimation of the characteristic resistance of AMR in European lakes, which can be used as a comprehensive resistome dataset to facilitate and monitor temporal changes in the development of AMR in European freshwater lakes.

1. Introduction

The surveillance of wastewater for Covid-19 virus in our current pandemic and mapped to the distribution and magnitude of the infected in the population near real-time exemplifies the importance of tracking rapidly changing trends of pathogens or public health problems such as antimicrobial resistance that may involve large populations in a large-scale (Kitajima et al., 2020). According to the World Health Organization (WHO), the resistance to antibiotic agents is one of the major threats to modern society (de Kraker et al., 2016; UN Interagency Coordination Group (IACG) on Antimicrobial Resistance, 2019;). As of 2019, there are already around 700.000 deaths annually, with a potential increase to 10 Million in the next decades without appropriate measurements (UN Interagency Coordination Group (IACG) on Antimicrobial Resistance, 2019). Thus, it is crucial to understand how environmental pressure gives rise to new resistance mechanisms. It has been shown that the over- and misuse of antibiotics are a significant contribution to AMR (Holmes et al., 2016; Singer et al., 2016). Moreover, Hendriksen *et al.* observed a strong correlation between sanitation standards and health care conditions by analyzing global AMR distribution in urban sewage (Hendriksen et al., 2019). Even indications for the contamination of community sewage by hospital wastewater burdened with antibiotic-resistant bacteria is present (Hocquet et al., 2016). These epidemiological approaches measure the collective signature across a community, and they have the potential to enhance detection, contain, and mitigate an outbreak. At the same time, the application may be deployed within

* Corresponding author.
E-mail address: dominik.heider@uni-marburg.de (D. Heider).
[1] contributed equally

https://doi.org/10.1016/j.envint.2021.106821
Received 13 May 2021; Received in revised form 4 August 2021; Accepted 6 August 2021
Available online 14 August 2021

monitoring networks to provide inter-comparable data across countries (Daughton, 2020).

However, hospitals, urban transit, and sewage share a significant commonality: they pool sources of health risk, e.g., the risk of contagion through human interchange or the spread of diseases through human waste, respectively. Such urban sewage and hospital wastewater are more likely to be contaminated with multi-resistant pathogens, and the search for AMR in this direction is quite natural. Other studies investigated AMR distribution in the environment, e.g., the abundance of antimicrobial resistance genes (ARGs) in 21 Swiss lakes (Czekalski et al., 2015). Freshwaters are the recipients of the effluent of wastewater treatment plants (WWTP). Consequently, several studies proved the presence of pathogens in natural surface waters (Blaak et al., 2015; Franz et al., 2015; Yang et al., 2017). Accordingly, Czekalski *et al.* revealed a 200-fold increase of ARGs in the sediment close to sewage release points of freshwater lakes (Czekalski et al., 2014).

However, the dissemination of AMR is not solely associated with human behavior, as various dispersal processes in most ecosystems also contribute to the spread of AMR (Berendonk et al., 2015). Two-thirds of the global antibiotic usage is associated with treating farm animals and agriculture, having a significant effect on the rise of AMR (Done et al., 2015; Van Boeckel et al., 2017). Most of these antibiotics belong to the class of so-called "uncritical" agents, e.g., tetracyclines and penicillins (Annual report on antimicrobial agents intended for use in animals, 2018). Concerningly, other studies reported similar indications, i.e., isolates from indicator bacteria reveal medium to high AMR levels to tetracyclines, sulfonamides, and quinolones (European Union Summary Report on antimicrobial resistance, 2013). Most notably, restricted antibiotics, such as colistin and third- and fourth-generation cephalosporins, are widely applied in poultry farming (European Union Summary Report on antimicrobial resistance, 2013). Wang *et al.* (2020) even conclude that reducing antibiotic contamination and eutrophication reduces the risk of AMR (Wang et al., 2020).

To this end, we collected samples from multiple freshwater lakes following a standardized protocol to detect AMR levels within microbial communities and quantify the resistome of the environment. Whereas our findings suggest that all possible AMR classes can be observed within the samples, we focused on four important classes of antibiotics in animal husbandry and human healthcare, i.e., tetracyclines, cephalosporins, quinolones, and sulfonamides, for quantifying the resistance in European freshwater lakes.

2. Materials and methods

2.1. Sampling and sample preparation

The dataset consists of standardized samples from 274 lakes for which 16S rRNA has been sequenced. Moreover, for 39 of these lakes, shotgun metagenomic sequencing was performed. For homogeneity, all samples were collected within one month and followed a standardized

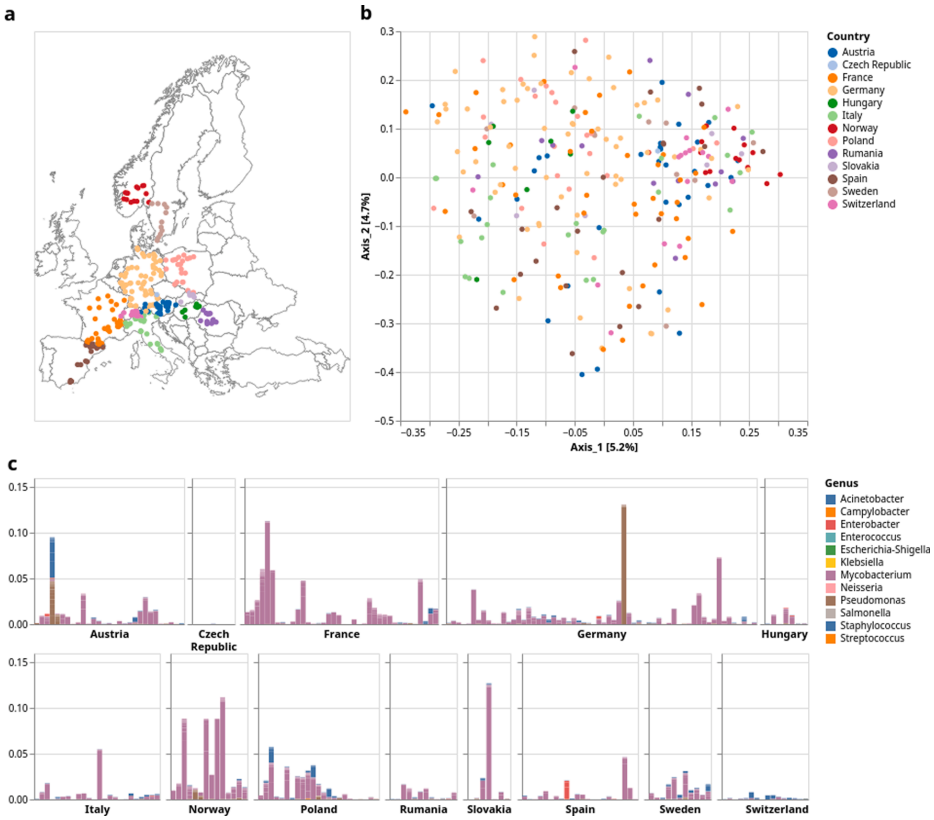


Fig. 1. Relative abundance of taxa, based on amplicon sequencing. **a** Map of the sample sites for the 274 freshwater lakes across Europe. **b** Principal coordinate analysis (PCoA) of operational taxonomic units. 5.2 and 4.7 % variation can be explained for the first and second components, respectively. **c** Relative abundance (y-axis) of genera, including genera with literature-known antimicrobial resistance for the different sample sites (x-axis).

protocol for sampling and analysis. Sampling sites are summarized in Fig. 1a. GPS coordinates of the different sampling sites are shown in Supplement 1. The sampling sites include lakes in 13 countries all across Europe. 274 European freshwater lakes were sampled, covering a broad latitudinal gradient ranging from Scandinavia to Spain. We have chosen a gradient design to cover a broader range of sampling points under varying environmental variables, including altitude or physicochemical factors, e.g., temperature, pH value, or chemical composition, instead of a replicated design, which is why no biological replicates were collected per lake. Samples were taken from the shore of each lake or pond, collecting epilimnial water up to 0.5 m depth. For genomic DNA extraction, samples were filtered onto 0.2 μm nucleopore filters until the filters were blocked to obtain similar amounts of biomass. Biomass filters were subsequently air-dried and preserved below −80 °C in a cryoshipper (Chart/MVE, Ball Ground, USA).

2.2. DNA extraction, PCR, and sequencing

For the amplicon analysis, genomic DNA was extracted from biomass filters using the my-Budget DNA Mini Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) following the manufacturer's protocol with minor adaptations. We changed the protocol as follows: Except that filters were homogenized in 800 μl Lysis Buffer TLS within lysing Matrix E tubes (MP Biomedicals, Santa Ana, California, USA) and homogenized three times for 45 s using FastPrep (MP Biomedicals, Santa Ana, California, USA) at 6 m/s followed by incubation for 15 min at 55° C. The DNA quality was checked using a NanoDrop™ ND-2000 UV–Vis spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and on 1% agarose gels. PCR amplifications targeted the V2-V3 region of the bacterial 16S rRNA gene using the primers 104F (5′-GGC GVA CGG GTG MGT AA-3′) and 515R (5′- TTA CCG CGG CKG CTG GCA C-3′) (Lange et al., 2015). The selected forward primer contains two wobble positions in order to cover a broad taxonomic spectrum. For each sample, two technical replicates of the extracted DNA were independently amplified using primers with different sample identifiers (Lange et al., 2015). For the PCR reaction, 1 μl of DNA template in 25 μl PCR reaction with 0.4 units of Phusion DNA polymerase (Thermo Fisher Scientific, Waltham, Massachusetts, USA), 0.25 μM primers, 0.4 mM dNTPs, and 1 × Phusion buffer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) was used. The PCR protocol consisted of 35 cycles, including a denaturation step at 98° C for 30 s, an annealing step at 72° C for 45 s (Lange et al., 2015), and an elongation step at 72° C for 30 s. Finally, the PCR was completed by a final extension step at 72° C for 10 min. Samples were pooled in equimolar ratios and sequenced using paired-end (2 × 300 bp) HiSeq 2500 Illumina sequencing in "rapid-run" mode at a sequencing provider (Fasteris, Geneva, Switzerland). Clean sequencing reads were in total 731,842,882 or on average 2,670,959 reads per lake. Finally, clean and demultiplexed samples were provided for further analyses.

Besides the amplicon analysis, we also carried out metagenomics analysis on 39 samples. The metagenomic samples were sequenced at BGI on an Illumina HiSeq XTen machine producing 150 bp paired-end samples. At BGI, genomic DNA was quality tested, and the qualified samples were used to construct the sequencing library. Therefore, purified DNA samples were first sheared into smaller fragments with the desired size by Covaris S/E210 or Bioruptor. Then the overhangs resulting from fragmentation were converted into blunt ends using T4 DNA polymerase, Klenow Fragment, and T4 Polynucleotide Kinase. Adapters were ligated onto both ends of the DNA fragments. The desired fragments were purified through gel-electrophoresis, then selectively enriched and amplified by PCR. Index tags were introduced into the adapter sequence to allow pooling. Finally, the libraries were quality tested and sequenced. Clean and demultiplexed samples were provided.

2.3. Amplicon analysis workflow

For the amplicon analysis, we used the standardized workflow Natrix, including (i) quality filtering, (ii) clustering, and (iii) taxonomy annotation (Welzel et al., 2020). The quality of the sequencing reads was re-checked using FastQC (v0.11.8), and low-quality tails were removed from the reads using PRINSEQ (Schmieder and Edwards, 2011) (v0.20.4). Trimmed reads with an average Phred quality score of less than 25 were discarded. Additionally, we removed all reads with at least one base with a quality of less than 15 and all reads that contained errors in the primer regions. Adapters containing primer, barcode, and poly-N sequences were removed, and the paired-end reads were subsequently assembled using PANDAseq (Masella et al., 2012) (v2.10). Chimeras were removed using UCHIME (usearch v7.0.1090) (Edgar et al., 2011). Subsequently, the sequences that passed quality and AmpliconDuo filtering (Lange et al., 2015) were clustered into Operational Taxonomic Units (OTUs) with SWARM (Mahé et al., 2015) (v2.1.9), using a local threshold since lineages evolve at variable rates. The local clustering threshold d was set to 1. For all OTUs, we used BLASTn (Altschul et al., 1990) (v2.7.1 +) with the NCBI nt and the Taxonomy Database (Dec 5, 2017) to annotate the OTUs with taxonomic information (see supplement 2 and 3).

2.4. Amplicon abundance analysis

We used the R package phyloseq (McMurdie and Holmes, 2013) for the relative abundance analysis of operational taxonomic units (OTUs). Specifically, the principal coordinate analysis has been conducted with the ordinate function using the "PCoA" method and Bray-Curtis dissimilarity as the distance metric. The OTU table is used for the principal coordinate analysis (PCoA). In order to statistically revise the PCoA, we used the non-parametric multivariate analysis of variance (MANOVA) (Anderson, 2001) provided by the R-Vegan package (v2.5–6) through the adonis function with Bray-Curtis dissimilarity (Oksanen et al., 2019).

For the bar chart, the following genera containing strains with literature-known antimicrobial resistance were filtered out from the original dataset: *Enterococcus*, *Mycobacterium*, *Staphylococcus*, *Streptococcus*, *Campylobacter*, *Neisseria*, *Escherichia-Shigella*, *Klebsiella*, *Enterobacter*, *Salmonella*, *Acinetobacter*, and *Pseudomonas*. This list is based on a reference list of pathogens of the Pathosystems Resource Integration Center (PATRIC) (Wattam et al., 2017). The visualization has been carried out with phyloseq's barplot function, using the genera for color filling. The final version of the plot is crafted with Altair, a visualization library for the Python programming language (VanderPlas et al., 2018).

2.5. Metagenomic analysis of antimicrobial resistance

Antimicrobial resistance (AMR) was analyzed using the resistance Gene Identifier (RGI tool) of the Comprehensive Antibiotic Resistance Database (CARD) (Jia et al., 2017), which can be used to predict the resistome from raw genome sequences. CARD is a database containing AMR drug classes and resistance mechanisms and intrinsic mutation-driven and acquired resistances. The basis consists of antibiotic resistance ontologies (ARO term), a networked and hierarchically controlled system of terms (Jia et al., 2017). Internally, the RGI tool uses Bowtie2 to align the metagenomic reads against CARD. We used the default settings for the analyses. For further evaluations, the focus was set on the AMR gene family and drug classes to achieve comparability and practical relevance. In addition, only those reads that have been entirely mapped to genes encoding for AMR factors were used for the subsequent analysis.

2.6. Metagenomic taxonomic analysis

In our analyses, we focused on reads related to resistance. Thus, only

those reads, which could be associated with AMR gene families or drug classes listed by the CARD database (see above), were used for further analysis. We used Centrifuge (Kim et al., 2016) to perform taxonomic analyses using Burrows-Wheeler transform (BWT) and Ferragina-Manzini (FM) index. The taxonomic analysis of the filtered FASTQ files was analyzed with the index containing all complete bacterial genomes (Kim et al., 2016). In order to cope with different sequencing depths among the samples, we normalized the mapped reads (numFragments) for gene length (geneLength) and sequencing depth (totalNumReads) (Chen et al., 2021):

$$FPKM = \frac{numFragments}{\frac{geneLength}{1000} * \frac{totalNumReads}{1.000.000}}$$

2.7. Data visualization and statistics

We used the function clustermap from the Python package seaborn for drawing the heatmaps (Waskom, 2021). All fragments, derived from the mapped reads of the CARD output, were displayed in one heatmap, each for the AMR gene families and the drug class resistance. We restricted the analyses and visualization to those gene families and drug class resistances that were most common among the lakes. We then clustered the lakes based on the country. AMR gene families or drug class resistances with over 500 fragments per lake accounted for less than 2% of all samples. Therefore, we set the limit for visualization to a maximum of 500 fragments per AMR gene family or drug class resistance, respectively, i.e., AMR gene families or drug class resistance with more than 500 fragments were capped. To finalize the heatmaps, we utilized Altair (VanderPlas et al., 2018). For Fig. 2, only the most common taxa are shown for the comparison. In order to analyze the pathogenic taxa, we used a filtered list of pathogens from PATRIC (Wattam et al., 2017).

Correlation analyses were carried out based on Pearson correlation to detect associations between resistance genes and taxa. To this end, we correlated the number of reads found by centrifuge with all mapped reads found by CARD and calculated the coefficient of determination. Moreover, we analyzed the association between resistance and farmland. We used SEDE-GPS for gathering socio-economic data (Sperlea et al., 2018). That is, we collected all data related to the term agriculture as defined by Eurostat (https://ec.europa.eu/eurostat/en/web/agriculture/data), for instance, agricultural products and organic farming, among others. SEDE-GPS takes a table with the GPS coordinates as input and collects information from different databases, such as Eurostat, within a user-specified radius. In our study, we used 20 km as the radius for SEDE-GPS. Correlations were calculated and reported based on the Pearson's product-moment correlation coefficient:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

with n the sample size, x_i and y_i the sample points, and $\overline{x}$ and $\overline{y}$ the corresponding mean; p values are calculated based on Student's t-distribution with $n-2$ degrees of freedom.

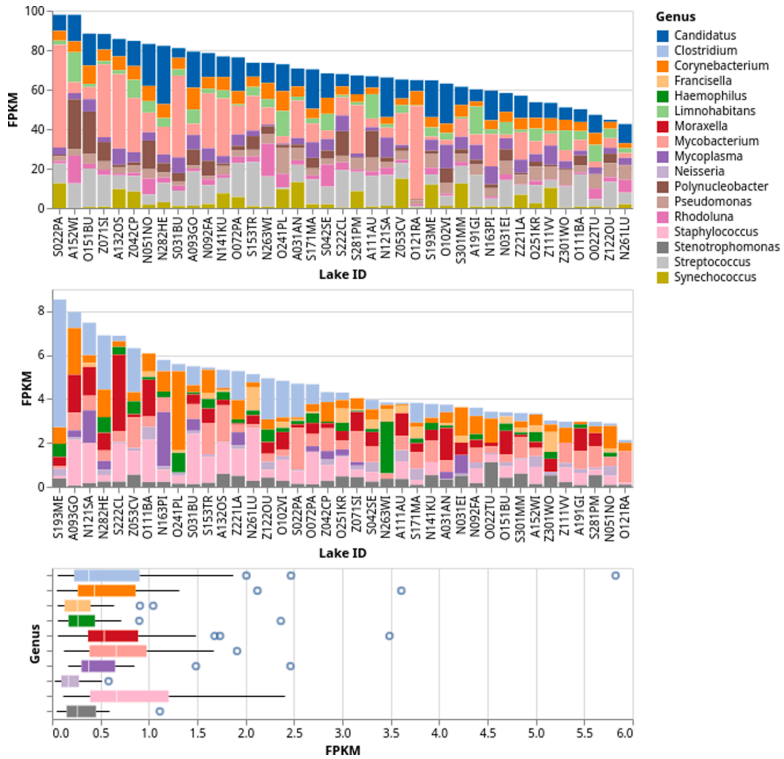


Fig. 2. Stacked bar charts of genera found in the metagenomic taxonomy classification. **Top:** The fragments mapped to a particular genus (y-axis) and for a sample site (x-axis). Only non-pathogenic Pseudomonads were found. **Center:** Non-AMR-related genera are removed, and only the top-10 of pathogenic genera are kept. **Bottom:** Total distribution of fragments mapped to pathogenic genera across all 39 lakes.

3. Results

3.1. Relative abundance of taxa

In total, we generated 731,842,882 clean sequencing reads from 274 freshwater lakes, i.e., 2,670,959 reads on average per lake, which we employed for subsequent 16S rRNA amplicon sequencing. Based on the 16S rRNA amplicon sequencing results, we analyzed the relative abundance of operational taxonomic units (OTUs) for each sample site (Fig. 1a). A summary of the lakes can be found in supplement 1. Moreover, a principal coordinate analysis reveals that seemingly no country-specific differences can be observed across the samples since only 5.2 and 4.7 % variation can be explained for the first and second components, respectively (Fig. 1b). However, a non-parametric multivariate analysis of variance unveils a significant difference across the samples ($p = 0.001$). We then focused on the genera known for antimicrobial resistance (AMR). The results indicated the presence of multiple genera, including species with known AMR (Fig. 1c). Representatives of the genus *Mycobacterium* can be observed in large quantities in almost all samples since it is widespread in aquatic ecosystems and is likely dominated by nontuberculous mycobacteria (NTM) (Roguet et al., 2018). Roguet and the coworkers pointed out that the

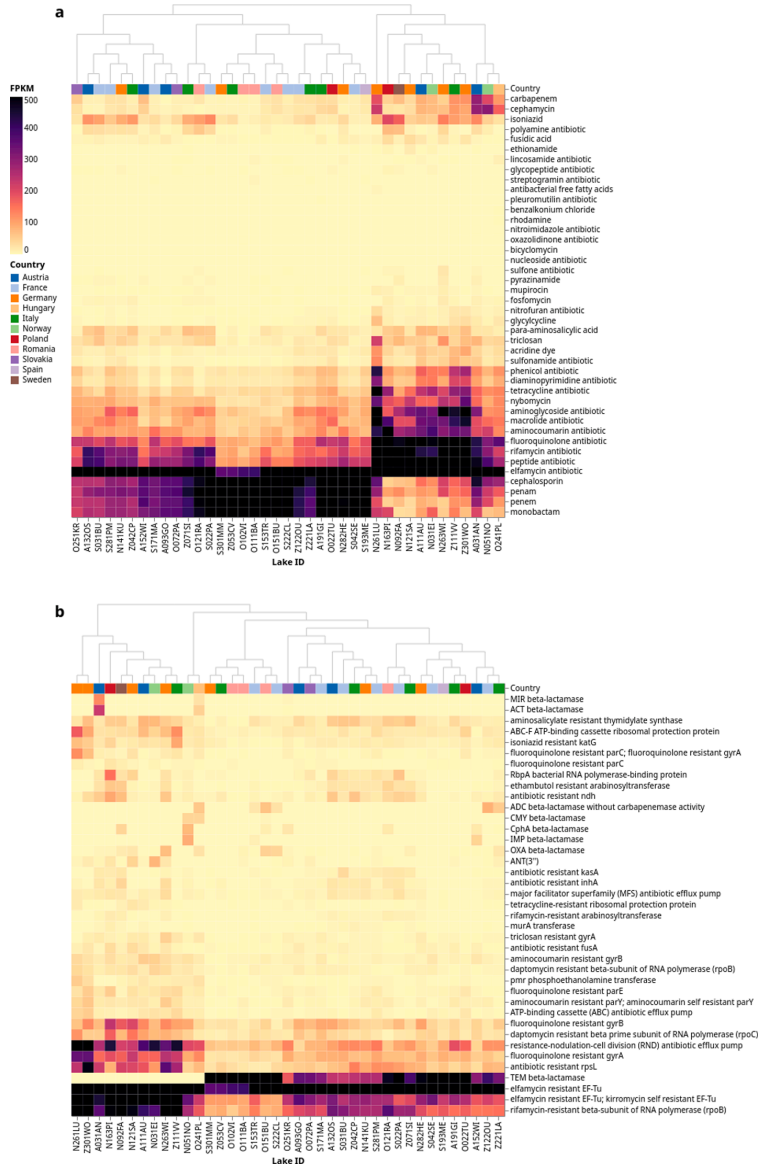


Fig. 3. Heatmaps are depicted of the respective lakes (x-axis) and resistances to drug classes (**a**) and AMR gene families (**b**) on the y-axis, respectively. In addition, the color density is determined by the quantity of fragments mapped against the respective category.

flowing of rivers into lakes appeared to strongly increase NTM densities, as opposed to lakes without connected rivers (Roguet et al., 2018). *Acinetobacter* species were also present in many samples, albeit in smaller quantities. Shao et al. (2019) suspected inflows polluted by stormwater runoff of sewage posed a possible source for *Acinetobacter* contamination (Shao et al., 2019). In addition, one lake in Germany seems to host larger quantities of *Pseudomonas* species (Fig. 1c). Species of these abundant genera are often associated with environmental waters and may be a source of opportunistic infections. The OTU table and the taxonomy data can be found in supplement 4 and 5.

3.2. Metagenomic taxonomy classification

Since amplicon sequencing only allows a straightforward overview of the present species, we conducted a more detailed taxonomic classification in a follow-up analysis based on metagenomic reads. We were able to increase the resolution of the found taxa based on the reads assigned to genes involved in antimicrobial resistance mechanisms. Specifically, we found multiple species related to AMR and a wide range of unrelated species, e.g., *Limnohabitans* strains (Fig. 2a). This analysis provides more details about the genus distribution in the samples. However, since we are interested in genera potentially related to AMR, we removed the remainders and focused on hits with AMR genes to research potential resistance. We detected fragments mapped to AMR-related genes, suggesting that AMR-related genera emerge in all samples, such as *Pseudomonas* or *Staphylococci* (Fig. 2b).

3.3. Antimicrobial resistance levels

We analyzed metagenome samples from 39 of the 274 freshwater lakes to identify AMR-related genes. The lakes were selected as a representative subset of the lakes analyzed in the first step, i.e., considering their geographical distribution to limit country-specific findings. Our results show indications towards various AMR genes in low to moderate quantities in the different samples (Fig. 3a). Specifically, we found a cluster of lakes exhibiting resistance against, e.g., β-lactam antibiotics, such as monobactams, cephalosporins, and penems, including lakes in France, Romania, and Germany. Genes encoding potential drug class resistances for quinolones and tetracyclines, as well as sulfonamides, can also be detected in the samples, albeit in smaller quantities (Fig. 3a). Moreover, an analysis on the abundance of AMR gene families reveals higher quantities of genes acting as antibiotic targets, mainly involved in protein biosyntheses, such as the elfamycin-resistant elongation factor thermo unstable (EF-Tu) gene (Parmeggiani and Nissen, 2006), the rifamycin-resistant β subunit encoding RNA polymerase (rpoB) gene (Goldstein, 2014), as well as the fluoroquinolone-resistant gyrases A and B (van der Heijden et al., 2012). Again, distinct clusters based on the quantity of mapped fragments can be observed for lakes in France, Germany, and Romania (Fig. 3b). Considering all mapped fragments for resistance against tetracyclines, cephalosporins, quinolones, and sulfonamides, it turns out that precisely, lakes in these countries exhibit diverse levels of putative resistance to drug classes stated above. Moreover, individual lakes in Germany, Italy, France, and Romania show a higher amount of mapped fragments to AMR-related genes than other countries (Fig. 4). We investigated a potential correlation with surrounding farmland or other possible factors in the following section.

3.4. Association to agriculture

As stated above, livestock farming is one of the main fields for applying antibiotics (Done et al., 2015; Van Boeckel et al., 2017). We employed SEDE-GPS to retrieve information on agriculture in an area of 20 km around the GPS coordinates of the lakes (Sperlea et al., 2018). Antibiotics or resistant bacteria from sewage with human excreta, in general, can enter freshwater in many ways. In a study conducted by the German Environment Agency, three possible pathways were identified: (i) the straight entering into surface water via excretion, for instance, sewage carrying human excreta with resistant bacteria, e.g., from stormwater runoff, (ii) the detour via the soil or (iii) via manure, supplied on fields and meadows (German Environment Agency, 2015). Our findings indicate only a low, non-significant correlation ($R = 0.28$, $p = 0.08$) between agricultural use and the frequency of antimicrobial resistance genes, particularly for sulfonamides. We also found non-significant correlations for the other three antibiotics investigated, i.e., tetracyclines, cephalosporins, and quinolones, which show no significant correlation with adjacent agriculture. Our findings suggest that human-made agricultural influences are low in Europe. However, we observed indications for genes encoding for resistance to one of the four drugs mentioned above, which will be discussed.

3.5. Tetracyclines

It has been reported that tetracyclines are among the most popular antibiotics in animal husbandry, with a share of around 30% (Annual report on antimicrobial agents intended for use in animals, 2018). Our findings are generally in support of this. Thus, resistance to this drug class can be observed in several European lakes in Austria, Germany, and Poland (Fig. 4). Furthermore, tetracyclines belong to frequently observed drug classes compared to others across all lakes (Fig. 3a).

3.6. Cephalosporins

In contrast to tetracyclines, cephalosporins, starting from the third generation on, are considered as critical antimicrobials (including carbapenems which are drugs of last resort), respectively, according to the WHO (Critically important antimicrobials for human medicine, 2019). However, their application is widespread, particularly in poultry farming (Annual report on antimicrobial agents intended for use in animals, 2018). Our study indicates the presence of resistance for this drug

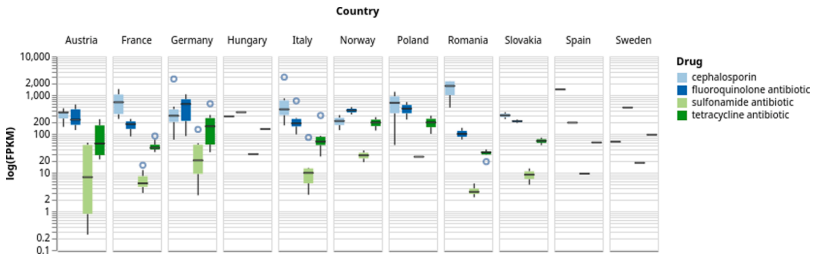


Fig. 4. Boxplot of the mapped fragments to genes encoding for resistance against different drug classes: tetracyclines, cephalosporins, (fluoro)quinolones, and sulfonamides. The following number of lakes were sampled in the respective countries: Austria (5), France (8), Germany (7), Hungary (1), Italy (6), Norway (2), Poland (2), Romania (4), Slovakia (2), Spain (1), and Sweden (1).

class. In particular, specific sample sites show higher exposure to microbes, potentially carrying associated resistance genes, compared to resistance against the remaining drug classes (Fig. 3a). If one considers the fragments mapped for individual countries, again lakes in Romania, France, and Germany reveal a descending order of fragments of genes encoding for resistance against this drug class (Fig. 4). The indications can be further validated by regarding individual lakes as shown in Fig. 5, where increased fragments mapped to genes encoding for cephalosporin resistance can be observed for lakes in Romania, Italy, and France.

3.7. (Fluoro)quinolones

These belong to another important class of antimicrobial agents, as stated by the WHO (Critically important antimicrobials for human medicine, 2019). The results also show resistance across individual lakes (Fig. 3a) as well as for specific countries (Fig. 4). Redgrave *et al.* (2014) observed a strong correlation between fluoroquinolones resistance and antibiotic consumption in Greece, France, and Sweden, i.e., the higher the intake, the higher the percentage of resistant *Escherichia coli* isolates (Redgrave et al., 2014). We endorse the findings of Redgrave *et al.*; i.e., fluoroquinolones resistance can be found from lakes in Germany, Italy, Romania, and France. Even for the sole Swedish lake, for which metagenomic samples have been sequenced, we observed indications for resistance to this drug class (Fig. 4), supporting the results from Redgrave *et al.* Interestingly, we observed a correlation between fluoroquinolones resistance and the presence of *Streptomyces albus*, a bacterial strain known for non-pathogenicity ($R = 0.52$, p less than 0.001). However, this might be due to cryptic gene clusters that are not expressed but are frequently found in *Streptomyces* (Xu et al., 2017).

3.8. Sulfonamides

Finally, the mapped fragments to genes encoding for sulfonamide resistance are lower than the other drugs but still present (see Fig. 3a). In addition, the quantity for individual countries is lower than the remaining drug classes (Fig. 4).

4. Discussion

We collected samples from 274 European lakes for a large-scale study on quantifying antimicrobial resistance (AMR) in freshwaters. To the best of our knowledge, this is the largest, standardized data set so far, employing 16S rRNA amplicon and metagenomic analyses on the bacterial composition and the resistome of these lakes. The standardized approach clearly distinguishes our study from others, relying heavily on non-standardized metagenomic data collected from public databases, differing in sampling protocols and analytic procedures, for instance, studies dealing with environmental or agricultural resistomes (Durso et al., 2012; Pal et al., 2016).

The present study used an integrative multi-omics approach using

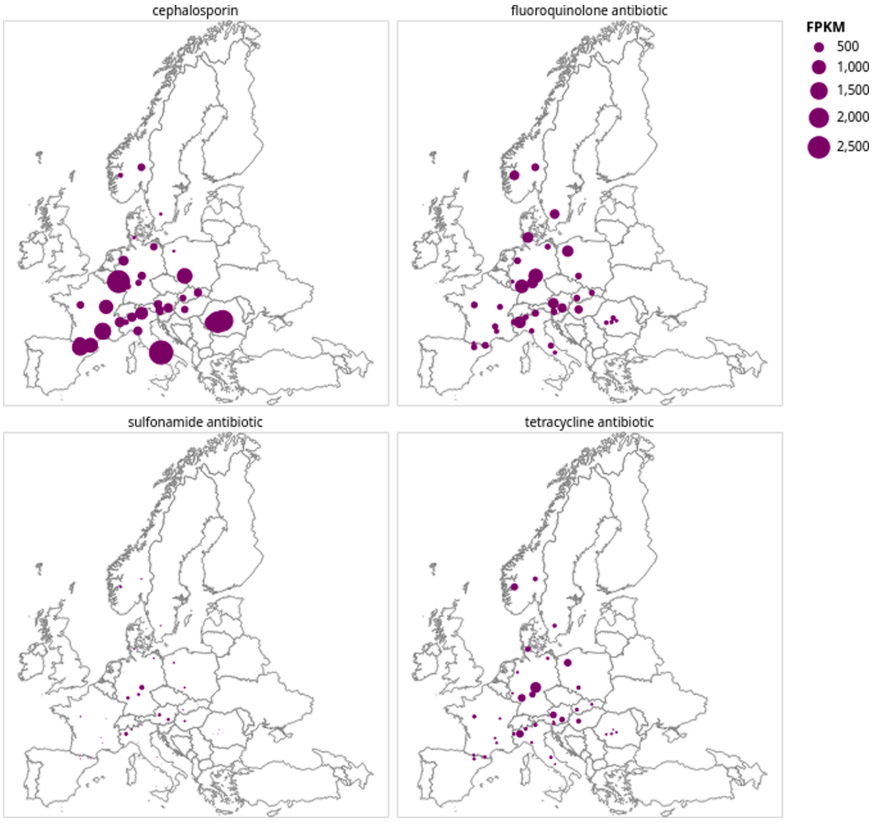


Fig. 5. Number of fragments per lake, which were mapped to genes encoding for resistance against tetracycline, sulfonamide, (fluoro)quinolone, and cephalosporin. Each dot denotes a specific lake, and the larger a dot, the more fragments were mapped to the respective resistance class. We observed increased numbers of fragments mapping to cephalosporin resistance for lakes in Germany, Italy, and Romania.

16S rRNA amplicon sequencing for a first-glance taxonomy identification, followed by a shotgun metagenomics analysis. Both approaches have strengths and weaknesses: taxonomical classification using 16S rRNA is more appropriate for various samples but offers a limited resolution in taxonomic classification depth. Contrary, shotgun metagenomics provides a detailed taxonomic resolution and the functional annotation of sequences, e.g., AMR genes (Jovel, 2016), however, at higher costs. Consequently, Hendriksen *et al.* argued that metagenomics offers the advantage of detecting transmissible resistance genes from a variety of bacterial species (Hendriksen et al., 2019).

Our findings quantify AMR among the analyzed lakes. We specifically focused on the association of resistance genes to four antibiotics, namely tetracyclines, cephalosporins, (fluoro)quinolones, and sulfonamides, with agriculture (Annual report on antimicrobial agents intended for use in animals, 2018). However, none of these show a significant correlation with agriculture. The data suggest that low human impact on AMR can be observed in the European freshwater lakes, and our findings may serve as a reference for monitoring AMR development in European freshwater lakes in the future.

We selected the 39 lakes as a representative subset of the overall lakes analyzed based on their geographical distribution, thus limiting the country-specific findings. Nevertheless, our methodology and data will be valuable as a reference to track the temporal development of AMR in Europe, and comparisons for those studied from countries of other continents. Our standardized approach contrasts from studies that already identified significant accumulation of AMR from lakes (Chakraborty et al., 2020; Kong et al., 2021; Ram and Kumar, 2020; Wang et al., 2020) and could avoid conditions, for instance, present in China and India, where, albeit governmental actions have been already taken, the environment is highly suffering from a mis- and overuse of antibiotics (Kakkar et al., 2017; Qu et al., 2019). One limitation in our current study is concerning the chromosomal and non-chromosomal elements such as plasmids, as the AMR genes are not necessarily vertically inherited, and the 16 s rRNA survey, therefore, most likely yields an incomplete list of AMR-related genera. Furthermore, characteristic mutations leading to resistance, e.g., in chromosomal genes gyrA and gyrB, were not considered in more detail nor correlated to phenotypic resistance of the bacteria. Moreover, Cox and Wright (2013) underpin the role of antibiotic-producing bacteria in soils (Cox and Wright, 2013) or species with chromosome-encoded elements, e.g., non-specific efflux pumps (Peterson and Kaur, 2018), which can be further disseminated by horizontal gene transfer (Cycoń et al., 2019). Thus, the exposure of intrinsically resistant bacteria to man-made environmental factors is not an explanation for their AMR and the natural resistome in European freshwater lakes.

In addition, we only considered agriculture, e.g., livestock farming, as an external impact on AMR levels in freshwater lakes. Hence our results might be biased towards agriculture (Collignon et al., 2018). Nevertheless, recent statistics about developments in agriculture in the European Union (EU) states that Romania, Italy, France, Poland, and Germany are among those countries with the most significant proportion of farming land (Agriculture, 2018). The findings by the EU coincides with the observations made by our study, i.e., our results verified not only indications for resistance against the four drug classes aforementioned but also specifically in these countries. Moreover, higher resistance against cephalosporins can be observed for lakes in France, Germany, or Romania, albeit their use is restricted in the EU (Fig. 4).

Our results support recent studies which reported increased levels of AMR resistance genes in various environments, e.g., against sulfonamides, in groundwater (Balzer et al., 2016), a further study which reported sewage as a source for AMR in the sediment of freshwater lakes (Czekalski et al., 2014), and in general, overuse of antibiotic agents in livestock farming (Hernando-Amado et al., 2019). We observed AMR in freshwater lakes, emphasizing AMR as a significant challenge for current and future healthcare systems. However, we could not rule out an overestimation of strain confidence completely, and the observed drug class resistances cannot be confidentially associated with present drug-resistant bacteria.

5. Conclusion

We comprehensively analyzed the resistome of freshwater lakes from European countries and focused explicitly on the antimicrobial resistance genes to four important classes of antibiotics, namely tetracyclines, cephalosporins, (fluoro)quinolones, and sulfonamides. Our findings provide a reference for the surveillance and monitoring of AMR development in European freshwater lakes and comparisons to those of other countries.

Funding Sources

This work has been financially supported by the Federal Ministry of Education and Research (BMBF) in project Deep-iAMR (FKZ 031L0209B) and the German Academic Exchange Service (DAAD) and BMBF under grant ID 57513593.

Author contributions

SS and DH developed the concept and designed the experiments. JB designed the sampling campaign. JKN, DB, and JB collected and preprocessed the sequencing data. SS and LE performed the experiments and analyzed the data. SS, LE, MI, and DH interpreted the results. SS and DH wrote the manuscript. JB and DH supervised the study. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106821.

References

Agriculture, forestry and fishery statistics. Eurostat; 2018. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Agriculture,_forestry_and_fishery_statistics.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol. 215 (3), 403–410.

Anderson MJ. A new method for non-parametric multivariate analysis of variance. Vol. 26, Austral Ecology. 2001. p. 32–46. Available from: https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.

Annual report on antimicrobial agents intended for use in animals. World Organisation for Animal Health; 2018.

Balzer F, Zühlke S, Hannappel S. Antibiotics in groundwater under locations with high livestock density in Germany. Vol. 16, Water Science and Technology: Water Supply. 2016. p. 1361–9. Available from: https://doi.org/10.2166/ws.2016.050.

Berendonk, T.U., Manaia, C.M., Merlin, C., Fatta-Kassinos, D., Cytryn, E., Walsh, F., et al., 2015. Tackling antibiotic resistance: the environmental framework. Nat Rev Microbiol. 13 (5), 310–317.

Blaak, H., Lynch, G., Italiaander, R., Hamidjaja, R.A., Schets, F.M., de Roda Husman, A. M., 2015. Multidrug-Resistant and Extended Spectrum Beta-Lactamase-Producing Escherichia coli in Dutch Surface Water and Wastewater. PLoS One. 10 (6), e0127752.

Chakraborty J, Sapkale V, Rajput V, Shah M, Kamble S, Dharne M. Shotgun metagenome guided exploration of anthropogenically driven resistomic hotspots within Lonar soda lake of India. Ecotoxicol Environ Saf. 2020;194:110443.

Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., et al., 2021. Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. Nat Commun. 12 (1), 1106.

Collignon, P., Beggs, J.J., Walsh, T.R., Gandra, S., Laxminarayan, R., 2018. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. Lancet Planet Health. 2 (9), e398–e405.

Cox, G., Wright, G.D., 2013. Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. Int J Med Microbiol. 303 (6–7), 287–292.
Critically important antimicrobials for human medicine, 2019. 6th revision. World Health Organization. Available from: https://www.who.int/foodsafety/publications/antimicrobials-sixth/en/.
Cycoń, M., Mrozik, A., Piotrowska-Seget, Z., 2019. Antibiotics in the Soil Environment-Degradation and Their Impact on Microbial Activity and Diversity. Front Microbiol. 8 (10), 338.
Czekalski, N., Gascón Díez, E., Bürgmann, H., 2014. Wastewater as a point source of antibiotic-resistance genes in the sediment of a freshwater lake. ISME J. 8 (7), 1381–1390.
Czekalski, N., Sigdel, R., Birtel, J., Matthews, B., Bürgmann, H., 2015. Does human activity impact the natural antibiotic resistance background? Abundance of antibiotic resistance genes in 21 Swiss lakes. Environ Int. 81, 45–55.
Daughton, C.G., 2020. Wastewater surveillance for population-wide Covid-19: The present and future. Sci Total Environ. 20 (736), 139631.
de Kraker MEA, Stewardson AJ, Harbarth S. Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? Vol. 13, PLOS Medicine. 2016. p. e1002184. Available from: https://doi.org/10.1371/journal.pmed.1002184.
Done, H.Y., Venkatesan, A.K., Halden, R.U., 2015. Does the Recent Growth of Aquaculture Create Antibiotic Resistance Threats Different from those Associated with Land Animal Production in Agriculture? AAPS J. 17 (3), 513–524.
Durso, L.M., Miller, D.N., Wienhold, B.J., 2012. Distribution and quantification of antibiotic resistant genes and bacteria across agricultural and non-agricultural metagenomes. PLoS One. 7 (11), e48325.
Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 27 (16), 2194–2200.
The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2011. Vol. 11. European Food Safety Authority (European Centre for Disease Prevention and Control); 2013 p. 3196. Available from: https://doi.org/10.2903/j.efsa.2013.3196.
Franz, E., Veenman, C., van Hoek, A.H.A.M., de Roda, H.A., Blaak, H., 2015. Pathogenic Escherichia coli producing Extended-Spectrum β-Lactamases isolated from surface water and wastewater. Sci Rep. 24 (5), 14372.
German Environment Agency, 2015. Pharmaceuticals in the environment-avoidance, reduction and monitoring. Available from: https://www.umweltbundesamt.de/sites/default/files/medien/378/publikationen/pharmaceuticals_in_the_environment.pdf.
Goldstein BP. Resistance to rifampicin: a review. Vol. 67, The Journal of Antibiotics. 2014. p. 625–30. Available from: https://doi.org/10.1038/ja.2014.107.
Hendriksen, R.S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., et al., 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun. 10 (1), 1124.
Hernando-Amado, S., Coque, T.M., Baquero, F., Martínez, J.L., 2019. Defining and combating antibiotic resistance from One Health and Global Health perspectives. Nat Microbiol. 4 (9), 1432–1442.
Hocquet D, Muller A, Bertrand X. What happens in hospitals does not stay in hospitals: antibiotic-resistant bacteria in hospital wastewater systems. Vol. 93, Journal of Hospital Infection. 2016. p. 395–402. Available from: https://doi.org/10.1016/j.jhin.2016.01.010.
Holmes, A.H., Moore, L.S.P., Sundsfjord, A., Steinbakk, M., Regmi, S., Karkey, A., et al., 2016. Understanding the mechanisms and drivers of antimicrobial resistance. Lancet. 387 (10014), 176–187.
Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., et al., 2017. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 45 (D1), D566–D573.
Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Vol. 7, Frontiers in Microbiology. 2016. Available from: https://doi.org/10.3389/fmicb.2016.00459.
Kakkar, M., Walia, K., Vong, S., Chatterjee, P., Sharma, A., 2017. Antibiotic resistance and its containment in India. BMJ. 5 (358), j2687.
Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L., 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26 (12), 1721–1729.
Kitajima, M., Ahmed, W., Bibby, K., Carducci, A., Gerba, C.P., Hamilton, K.A., et al., 2020. SARS-CoV-2 in wastewater: State of the knowledge and research needs. Sci Total Environ. 15 (739), 139076.
Kong M, Bu Y-Q, Zhang Q, Zhang S-H, Xing L-Q, Gao Z-Q, et al., 2021. Distribution, abundance, and risk assessment of selected antibiotics in a shallow freshwater body used for drinking water, China. Vol. 280, Journal of Environmental Management. p. 111738. Available from: https://doi.org/10.1016/j.jenvman.2020.111738.
Lange, A., Jost, S., Heider, D., Bock, C., Budeus, B., Schilling, E., et al., 2015. AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial Communities. PLoS One. 10 (11), e0141590.
Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M., 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 10 (3), e1420.
Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., Neufeld, J.D., 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics. 14 (13), 31.
McMurdie, P.J., Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 8 (4), e61217.
Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2019. Available from: https://CRAN.R-project.org/package=vegan.
Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. The structure and diversity of human, animal and environmental resistomes. Microbiome. 2016;4(1):54.
Parmeggiani, A., Nissen, P., 2006. Elongation factor Tu-targeted antibiotics: four different structures, two mechanisms of action. FEBS Lett. 580 (19), 4576–4581.
Peterson, E., Kaur, P., 2018. Antibiotic Resistance Mechanisms in Bacteria: Relationships Between Resistance Determinants of Antibiotic Producers, Environmental Bacteria, and Clinical Pathogens. Front Microbiol. 30 (9), 2928.
Qu J, Huang Y, Lv X. Crisis of Antimicrobial Resistance in China: Now and the Future. Front Microbiol. 2019;10:2240.
Ram B, Kumar M., 2020. Correlation appraisal of antibiotic resistance with fecal, metal and microplastic contamination in a tropical Indian river, lakes and sewage. Vol. 3, npj Clean Water. Available from: https://doi.org/10.1038/s41545-020-0050-1.
Redgrave, L.S., Sutton, S.B., Webber, M.A., Piddock, L.J.V., 2014. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. Trends Microbiol. 22 (8), 438–445.
Roguet, A., Therial, C., Catherine, A., Bressy, A., Varrault, G., Bouhdamane, L., et al., 2018. Importance of Local and Regional Scales in Shaping Mycobacterial Abundance in Freshwater Lakes. Microb Ecol. 75 (4), 834–846.
Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27 (6), 863–864.
Shao, Keqiang, Yao, Xin, Xie, Guijuan, Wu, Yuanyuan, Hu, Yang, Tang, Xiangming, Gao, Guang, 2019. Detectable Levels of Bacterial Pathogens in the Rivers of the Lake Chaohu Basin, China. Int J Environ Res Public Health 16 (23), 4857. https://doi.org/10.3390/ijerph16234857.
Singer, A.C., Shaw, H., Rhodes, V., Hart, A., 2016. Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators. Front Microbiol. 1 (7), 1728.
Sperlea, T., Füser, S., Boenigk, J., Heider, D., 2018. SEDE-GPS: socio-economic data enrichment based on GPS information. BMC Bioinformatics. 19 (Suppl 15), 440.
UN Interagency Coordination Group (IACG) on Antimicrobial Resistance. No Time to Wait: Securing the future from drug-resistant infections. World Health Organization; 2019.
Van Boeckel, T.P., Glennon, E.E., Chen, D., Gilbert, M., Robinson, T.P., Grenfell, B.T., et al., 2017. Reducing antimicrobial use in food animals. Science. 357 (6358), 1350–1352.
van der Heijden YF, Maruri F, Sterling TR, Kaiga A, Blackman A, et al. A Systematic Review Of Gyrase Mutations Associated With Fluoroquinolone-Resistant Mycobacterium Tuberculosis And A Proposed Gyrase Numbering System. B54. TUBERCULOSIS IN SPECIAL POPULATIONS. 2012. Available from: https://doi.org/10.1164/ajrccm-conference.2012.185.1_meetingabstracts.a3265.
VanderPlas, Jacob, Granger, Brian, Heer, Jeffrey, Moritz, Dominik, Wongsuphasawat, Kanit, Satyanarayan, Arvind, Lees, Eitan, Timofeev, Ilia, Welsh, Ben, Sievert, Scott, 2018. Altair: Interactive Statistical Visualizations for Python. Journal of Open Source Software. 3 (32), 1057. https://doi.org/10.21105/joss10.21105/joss.01057.
Wang, Z., Han, M., Li, E., Liu, X., Wei, H., Yang, C., et al., 2020. Distribution of antibiotic resistance genes in an agriculturally disturbed lake in China: Their links with microbial communities, antibiotics, and water quality. J Hazard Mater. 5 (393), 122426.
Waskom M., 2021, seaborn: statistical data visualization. Vol. 6, Journal of Open Source Software. p. 3021. Available from: https://doi.org/10.21105/joss.03021.
Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al., 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45 (D1), D535–D542.
Welzel, M., Lange, A., Heider, D., Schwarz, M., Freisleben, B., Jensen, M., et al., 2020. Natrix: a Snakemake-based workflow for processing, clustering, and taxonomically assigning amplicon sequencing reads. BMC Bioinformatics. 21 (1), 526.
Xu, F., Nazari, B., Moon, K., Bushin, L.B., Seyedsayamdost, M.R., 2017. Discovery of a Cryptic Antifungal Compound from Streptomyces albus J1074 Using High-Throughput Elicitor Screens. J Am Chem Soc. 139 (27), 9203–9212.
Yang, Y., Xu, C., Cao, X., Lin, H., Wang, J., 2017. Antibiotic resistance genes in surface water of eutrophic urban lakes are related to heavy metals, antibiotics, lake morphology and anthropic impact. Ecotoxicology. 26 (6), 831–840.

6.2 Encodings and Models for Antimicrobial Peptide Classification for Multi-resistant Pathogens

AMR led to countless deaths and burdened public health care systems drastically with the associated costs[235]. AMR refers to the loss of efficiency of antibiotics against pathogens through natural or acquired resistance mechanisms[235]. These mechanisms concern, for instance, removal of antibiotics from the cell interior or inactivation[19]. The emergence of AMR is fostered by various factors, including mobile genetic elements (MGEs), which can be disseminated by horizontal gene transfer (HGT)[100,73]. As a consequence, more pathogens acquire resistance; thus, more antibiotics become ineffective, ultimately compromising medical treatment[235]. In addition, the over- and misuse of antibiotics increase evolutionary pressure on the bacteria and promotes AMR[19]. The ability of biofilm-forming pathogens further enhances resistance to common antibiotics[80].

AMPs, a class of molecules belonging to the innate immune system of mammalians and other organisms, are an alternative strategy to conventional antibiotics[135]. AMPs are ubiquitously produced as a host defense mechanism in epithelial cells, including the skin, to neutralize microorganisms, such as bacteria and fungi[93,151]. The mode of action is mainly interfering with the pathogenic cell, hence, membrane disruption[71,258]. Another effect is translocation to perturb intracellular targets, including the whole gene expression process[71,258]. High concentrations of AMPs, generally with low activity per se, lead to increased antimicrobial efficiency and the potential of degrading biofilms[71]. Moreover, AMPs are highly specific to the bacterial cell wall[71]. The reason for the specificity are different molecular properties of prokaryotic and eukaryotic membranes[71]. In addition, resistance to AMPs is low[135].

Manual identification of AMPs is time-consuming[77]. To this end, researchers employed machine learning (ML) for automated classification and autonomous decision making, for instance, in the field of self-driving cars[9] or face recognition[46]. Advances in ML methods, greater peptidomics databases, and increased computational power paved the way for predicting novel AMPs[192,240]. However, most ML algorithms require numerical and fixed-length input[137].

Amino acids feature physicochemical properties, such as the hydrophobicity scale by Kyte and Doolittle (1982)[130]. Consequently, peptides can be represented as numerical vectors. Besides the physicochemical properties mentioned above, researchers developed various encodings. An example concerns the binary encoding, meaning each amino acid is represented as a vector of zeros with one bit set[121]. The length of the vector is 21, hence, equal to the number of natural amino acids plus one bit denoting a gap[121]. However, one can find additional encodings in the literature[39]. Thus, we collected and reviewed peptide encodings, specifically for AMP prediction. In particular, we elaborated the different types, algorithms, and origins. The study concludes with an overview of libraries and databases, which can be

used for dataset creation and encoding.

We specified sequence-based encodings (SeBEs), structure-based encodings (StBEs), alternative encodings, and model-based encodings (MoBEs). SeBEs are derived from the order of the amino acids (primary structure). StBEs rely on the folding of the amino acid sequence in higher dimensions (secondary or tertiary structure). Encodings neither fitting to the main categories have been assigned as alternative encodings. MoBEs result from an intrinsic model representation, for instance, filter layers from Deep Learning (DL) or custom kernels from Support Vector Machines (SVMs).

In summary, we collected encodings from a wide range of biomedical studies. We observed that SeBEs not only map amino acids to numerals but also reflect interactions between non-adjacent amino acids. Moreover, StBEs, although requiring a known peptide structure, have been applied successfully in multiple studies[22,146]. For completeness, models and the particular application are also described in detail. Encoding libraries, focusing on SeBEs, provide easy access to the underlying algorithms. Since encoding selection is challenging, we finally recommend to include the diversity of the model outputs for encoding selection. However, more research is necessary in this direction. Overall, the study provided an overview of all aspects of AMP prediction, hence, paving the way for future research.

Spänig and Heider *BioData Mining* (2019) 12:7
https://doi.org/10.1186/s13040-019-0196-x

BioData Mining

REVIEW Open Access

Encodings and models for antimicrobial peptide classification for multi-resistant pathogens

Sebastian Spänig and Dominik Heider*

* Correspondence: dominik.heider@uni-marburg.de
Department of Bioinformatics, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany

Abstract

Antimicrobial peptides (AMPs) are part of the inherent immune system. In fact, they occur in almost all organisms including, e.g., plants, animals, and humans. Remarkably, they show effectivity also against multi-resistant pathogens with a high selectivity. This is especially crucial in times, where society is faced with the major threat of an ever-increasing amount of antibiotic resistant microbes. In addition, AMPs can also exhibit antitumor and antiviral effects, thus a variety of scientific studies dealt with the prediction of active peptides in recent years. Due to their potential, even the pharmaceutical industry is keen on discovering and developing novel AMPs. However, AMPs are difficult to verify in vitro, hence researchers conduct sequence similarity experiments against known, active peptides. Unfortunately, this approach is very time-consuming and limits potential candidates to sequences with a high similarity to known AMPs. Machine learning methods offer the opportunity to explore the huge space of sequence variations in a timely manner. These algorithms have, in principal, paved the way for an automated discovery of AMPs. However, machine learning models require a numerical input, thus an informative encoding is very important. Unfortunately, developing an appropriate encoding is a major challenge, which has not been entirely solved so far. For this reason, the development of novel amino acid encodings is established as a stand-alone research branch. The present review introduces state-of-the-art encodings of amino acids as well as their properties in sequence and structure based aggregation. Moreover, albeit a well-chosen encoding is essential, performant classifiers are required, which is reflected by a tendency towards specifically designed models in the literature. Furthermore, we introduce these models with a particular focus on encodings derived from support vector machines and deep learning approaches. Albeit a strong focus has been set on AMP predictions, not all of the mentioned encodings have been elaborated as part of antimicrobial research studies, but rather as general protein or peptide representations.

Keywords: Machine learning, Antimicrobial peptides, Encodings

Introduction

Antimicrobial peptides are part of the inherent immune system of almost all organisms, such as plants, animals, and humans [1]. Owing to increasing rates of multi-resistant pathogens, the scientific community has reached out for novel strategies to tackle this threat [2, 3]. One of these approaches leverages the endogenous defense system mode of action, particularly on exposed surfaces, such as the skin,

commonly referred to as antimicrobial peptides (AMPs) [1]. To this end, researchers have shown that AMPs also have an effect even against multi-resistant pathogens and thus, can effectively employed as antibiotic agents. AMPs can also interfere intracellular mechanisms, which makes these potential candidates for cancer treatment or inflammatory diseases [4]. Owing to their broad fields of application and the demonstrated potential, the pharmaceutical industry pushes research ahead in order to discover and develop novel and highly effective AMPs, such as the approved polymyxins, which serve as last resort therapy, if the usual treatment fails [4]. In order to enable AMP detection with low costs and in high throughput, computational approaches offer the opportunity to explore the huge space of sequence variations in a timely manner. In particular, artificial intelligence, hence machine learning algorithms perform well in prediction and classification tasks, including computer vision [5], autonomous driving [6], or life science [7]. It is thus not surprising, that machine learning has been applied for fast and automated discovery of AMPs [8] and protein classification in general [9]. Two major issues arise here: firstly, biological information of the amino acid sequence has to be translated into a numerical representation and secondly, the input must not be of varying length, therefore sequence lengths have to be aligned. This is due to the intrinsic nature of machine learning models, i.e., the requirement of a numerical input with a fixed dimension. To this end, a variety of encodings has been developed over time. Each of these encodings are created to reflect biological relationships as well as intrinsic information of the primary sequence and higher order confirmations as accurate as possible. Since an informative encoding is very important and crucial for prediction accuracy, not only numerous encodings have been proposed, but also various strategies to combine existing ones. In order to shed light in this complex topic, literature has been mined for sequence and structure based encodings and elaborated as part of this review. The goal of the present study is the easing of the application of existing encodings for own projects and to encourage further research in the automated classification of antimicrobial peptides. The paper is structured as follows: in order to understand the rationale behind different encodings, we introduce the general effect of AMPs in the first section. Afterwards, prepared with the biological background, we summarize sequence- and subsequently structure-based encodings in the second section. Since the prediction task requires not only an expressive encoding, but also a performant classifier, we further highlight the employed machine learning algorithms in another section. Moreover, special encodings have been derived from support vector machines and deep learning. For this reason, we elaborate on these more detailed in another section. For the sake of completeness, tools for AMP prediction are uncovered, which includes different databases as sources for AMP sequences and packages, which provide implementations for many of the presented encodings.

Antimicrobial peptides

AMPs are part of the inherent immune system and can be especially found in exposed surfaces, such as mucosa and the skin [1]. At these sites, AMPs serve as a defence system and are expressed to protect the organism against microbial intruders. The defense measures encompasses different types of bacterial interaction, mostly due to the AMPs physicochemical properties and the resulting three-dimensional structure. That is, mostly positive charged and hydrophobic residues are constituted to 10 to 50 residues

long peptides, forming either α-helices, β-sheets or random coils [1]. Due to the "multi-hit mechanism", adaption against AMPs is difficult and thus, AMPs are effective even against highly resistant pathogens. To this end, active peptides are interacting with pathogens in two ways: on the one hand, they disrupt the bacterial membrane and on the other hand, they advance further into the cell, generally known as translocation [10]. Because of different characteristics of eukaryotic and prokaryotic membranes, the interaction of AMPs with their corresponding target is highly selective [11]. The membrane disruption leads to the loss of important ions and metabolites, which finally leads to cell lysis and subsequently to cell death [1]. Essentially, three membrane disruption models are known: the barrel-stave model for pore building, the carpet model for disintegration of the membrane, as well as the toroidal-pore model for arranging the membrane to build continuous pores [1, 11]. The further advancement to intracellular location, i.e., translocation, takes place without permeabilizing the pathogens membrane. Within the cell, AMPs aggregate in the cytoplasm and inhibit nucleic acid as well as protein synthesis [12]. Besides antimicrobial effects, antiparasitic, antivirus, and anticancer effects have been reported. In the case of the latter, AMPs can trigger apoptosis and prevent angiogenesis [4].

While most AMPs have the ability to kill microbial pathogens directly, other peptides, e.g., anticancer AMPs, have immunomodulatory capabilities to stimulate cells and tissues of the host defense system. More general, these class of peptides are known as host defense peptides (HDP). For instance, the well-studied HDP LL-37 [13] reveals its complex mode of action, due to direct and indirect interactions with a vast amount of genes and proteins of the host. Hence, HDPs are important signaling molecules, capable, for instance, to regulate autoimmune response in the case of inflammatory diseases or, as mentioned above, support tumor suppression [14].

Encodings

This section describes the different approaches and mechanisms to encode an amino acid sequence as a numerical vector and is divided in two main parts: the first deals with sequence-based encodings and the second part describes structure-based encodings. The former, summarized in Table 1, encompass sparse or binary encoding, followed by the general and the pseudo-amino acid composition. Afterwards, the reduced amino acid alphabet will be introduced as well as descriptors, which incorporate physicochemical as well as statistical properties of the respective amino acid and substitution matrices (which incorporate the substitution frequency of amino acids). Nevertheless, the function of a peptide is defined by its three-dimensional shape, hence structure-based encodings (Table 2) have been proposed in order to improve prediction performances. Thus the second part of this section introduces structure-based encodings. Besides the classical state-of-the-art approaches for encoding of peptides, novel, promising encodings have been developed, such as the Chaos Game Representation, which are described in the third section and summarized in Table 3. Hereinafter, each of these encodings are compared in detail and applications and method specific customizations are provided as well as, if possible, the relation between the biology behind the encodings and the antimicrobial effect.

Table 1 Summary of sequence based encodings

Encoding	Description	Summary	Used in	Used along with	Main Category
Sparse	each amino acid is represented as an one-hot vector of length 20, where each position, except one, is set to 0	Density: - Information: +	[15, 19–21]	Substitution Matrix, Amino Acid Composition	Sparse encoding
Amino Acid Composition	feature vector contains at each position the proportion of an amino acid in relation with the sequence length	Density: + Information: -	[22–24]	Distance Frequency, Quantitative Matrix, Dipeptide Composition, PseAAC	Amino acid composition
Distance Frequency	calculates the distance between amino acids of similar properties and bins the occurrence according to the gap length	Density: + Information: +	[22]		Amino acid composition
Quantitative Matrix	encodes the propensity of each amino acid at a position	Density: + Information: +	[23]		Amino acid composition
CTD	describes the composition (C), transition (T) and distribution (D) of similar amino acids along the peptide sequence	Density: + Information: +	[25]		Amino acid composition
Pseudo-amino Acid Composition (PseAAC)	computes the correlation between different ranges among a pair of amino acids	Density: + Information: +	[27–30]	Dipeptide Composition	Pseudo amino acid composition
Reduced Amino Acid Alphabet	similar amino acids are grouped together	Density: + Information: o	[9, 32–34, 36, 37]	N-gram Model, AAIndexLoc	Reduced amino acid alphabet
N-gram Model	occurrences of n-mers for an alphabet of size m, leading to a m^n dimensional, sparse representation of the initial sequence	Density: - Information: o	[9]		Reduced amino acid alphabet
AAIndexLoc	k-nearest neighbor clustering to aggregate amino acids into 5 classes using their amino acid index, i.e., amino acids with the respective highest(T), high (H), medium (M), low (L), and lowest (B) values of a particular physicochemical property are clustered together	Density: o Information: +	[37]	Dipeptide Composition	Reduced amino acid alphabet
Physicochemical Properties	translation of an amino acid to a	Density: o Information: +	[40, 42, 47–53]	z-descriptor, d-descriptor	Physicochemical properties

Table 1 Summary of sequence based encodings *(Continued)*

Encoding	Description	Summary	Used in	Used along with	Main Category
	particular physicochemical property			and many more	
z-descriptor	derived from the principal components of physicochemical properties by means of partial least squares (PLS) projections, PLS leads to a subset of five final features, capable to describe the 20 proteinogenic as well as 67 additional amino acids	Density: + Information: +	[42, 44]		Physicochemical properties
d-descriptor	amino acid sequence is squeezed between the y- (N-terminus) and the x-axis (C-terminus) with gradually bending of the single amino acids and subsequent vector summation	Density: + Information: +	[54]		Physicochemical properties
Autocorrelation	interdependence between two distant amino acids in a peptide sequence	Density: + Information: +	[57–61]		Autocorrelation
Substitution/ Scoring Matrix	provide accepted mutations between amino acid pairs, i.e., sequence alterations with either no or positive impact in terms of the protein function	Density: + Information: +	[65–71]	BLOMAP, Sparse, Amino Acid Composition, Dipeptide Composition, PseAAC, AAIndexLoc	Substitution and scoring matrix
BLOMAP	incorporates the BLOSUM62 to calculate distances in a high dimensional input space, i.e., the substitution matrix, to a lower dimension, using the Shannon-projection	Density: + Information: +	[65]		Substitution and scoring matrix
Fourier Transformation	to detect underlying patterns in time series, by transforming the time signal to a frequency domain	Density: o Information: +	[73, 74]		Fourier Transformation

+ (good), o (neutral/no declaration), – (bad). For instance, "Density: -" means the encoding results in a high dimensional feature space and "Information: +" reflects a representative mapping from the residue sequence to the numerical vector. "o" denotes encodings, which are difficult to classify, due to missing details in the respective publication or can be considered as neutral. In general, the classification rests upon the authors experience and shall support researchers to quickly grasp suitable encodings. Nevertheless, an encoding which has been rated "-" still might work well for a particular application and should by no means regarded as the final evaluation

Table 2 Summary of structure derived encodings

Encoding	Description	Summary	Used in	Used along with
Quantitative structure-activity relationship (QSAR)	describes amino acids sequences by their chemical properties, molecular characteristics and structure	Density: o Information: +	[78–85]	z-Descriptors
General Structure	protein structure is described by means of their total 3D shape, secondary structure, solvent accessibility, aggregation tendency, contact number, residue depth	Density: + Information: +	[86–88, 97]	
Electrostatic Hull	wraps superimposed shapes of the proteins sub-structure	Density: o Information: +	[17, 89, 90]	Physicochemical Properties
Spheres	incorporates structural variations as consequence of sequential rearrangements	Density: o Information: +	[91]	Physicochemical Properties
Distance Distribution	distribution of euclidean distances between each atom type	Density: o: Information: +	[92]	
Delaunay Triangulation	encodes the complete protein shape by finding the optimal edges between representative atoms	Density: o Information: +	[93, 94]	

+ (good), o (neutral/no declaration), – (bad) (see Table 1 for further details)

Sequence based encodings

Sparse encoding

The first approach that has been used to describe a peptide sequence is sparse encoding (also named binary encoding). In sparse encoding, each amino acid is represented as an one-hot vector of length 20, where each position, except one, is set to 0. Thus, in a vectorized format, the amino acids alanine and valine are encoded as 10000000000000000000 and 00000000000000000001, respectively [15]. For instance, the amino acid sequence GHKARVLAEAMSQVTGSAAVM, the p2 peptide ([16, 17]), is encoded into the matrix A as:

$$A = \begin{array}{c} \\ G \\ H \\ \vdots \\ V \\ M \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccccc} A & R & N & D & C & E & Q & G & H & I & L & K & M & F & P & S & T & W & Y & V \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Since machine learning models require a fixed input dimension, the respective sequence lengths have to be adjusted before encoding. In the present case, this happens either by a multiple sequence alignment or with a pairwise alignment against a reference sequence. The alignments will introduce gaps, hence a further dummy amino acid

Table 3 Summary of alternative encodings (see Table 1 for further details)

Encoding	Description	Summary	Used in	Used along with
Chaos Game Representation (CGR)	a visual encoding of a sequence, generating a fractal	Density: - Information: o	[98–102]	Physicochemical Properties
Linguistic Model	description of AMPs by a grammar	Density: o Information: o	[103]	

has to be added to the matrix. On the one hand, sparse encoding offers the advantage of providing an easy representation of the 20 proteinogenic amino acids (plus one dummy residue for gaps). On the other hand, the resulting input space for subsequent machine learning is inflated and could impose problems, such as the curse of dimensionality [18]. The feature vector dimension will be inflated to 21*max(l), whereby l denotes the length of a given peptide sequence. Nevertheless, sparse encoding is frequently used. For instance, Hirst et al. (1992) used this encoding to train a neural network and to predict secondary structure as well as the function [15]. However, the authors used sliding windows to separate the original sequence into segments such that the impact of spatially close residues is considered. Thus, the dimension of the input vector is 20 (each amino acid) times the window size [15]. Another study combined sparse encoding and a substitution-matrix-based encoding to predict peptide binding affinity to T-cell epitopes using neural networks [19]. The latter encoding increases the generalization ability of the classifier, whereas the sparse encoding does not provide additional information, except simply the amino acid itself [19]. This drawback of sparse encodings has been recognized by others. For instance, as part of a study to predict peptide induced modulation of antigen presenting cells, Nagpal et al. (2018) encoded the N-terminus and the C-terminus as binary vectors and used this encoding along with the overall amino acid composition as features for a support vector machine (SVM) [20]. Usmani et al. (2018) used a similar combination of sparse encoding of both termini and amino acid composition in order to predict antitubercular peptides by means of an ensemble classifier [21]. In addition, they state that sparse encoding has the advantage to keep the sequence order information [21].

Amino acid composition

An approach to overcome the limitations of sparse encoding and hence making the resulting feature space more dense, is the representation of the amino acid sequence as its respective composition. Here, the final feature vector contains at each position the proportion of an amino acid in relation with the sequence length (Fig. 1). For instance, one can divide a peptide into chunks including both termini and calculate the local amino acid composition [22]. The amino acid composition differs from one class to the another and, for instance, cell penetrating peptides require hydrophobic residues at the N-terminus, which could be approximated well by the features gained from the local

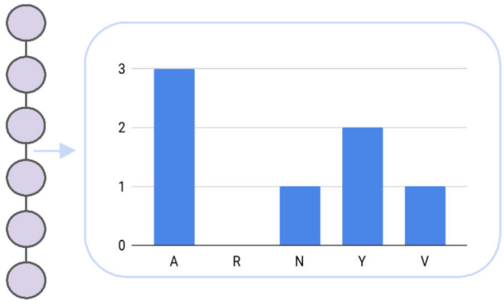


Fig. 1 The single letter amino acid composition counts the occurrence of the respective amino acids

composition [1]. Additional performance has been achieved by introducing a technique called distance frequency, which calculates the distance between amino acids of similar properties and bins the occurrence according to the gap length. Matsudo et al. (2005) used both encodings to predict the subcellular location by means of SVMs [22]. Commonly, amino acid composition is applied to distinguish between different classes of peptides, i.e., antimicrobial and non-antimicrobial peptides [23] or to classify antiviral, antitumor, antibacterial, and antifungal peptides [24]. The former introduces quantitative matrices as a novel descriptor, which encodes the propensity of each amino acid at a certain position. This encoding has been employed in addition to local sparse encoding for analysing as well as predicting antimicrobial peptides in general. In contrast, the latter study applied increment of diversity (ID) to classify unknown peptides to the respective classes. To ensure a well-performing classifier, the ID is not only based on the amino acid composition, but is rather used along with the dipeptide and the pseudo-amino acid composition, which will be introduced hereinafter. Dubchak et al. (1995) proposed an encoding, which describes the composition (C), transition (T) and distribution (D) of similar, hence in terms of physicochemical properties, amino acids along the peptide sequence [25]. C refers to the composition of the respective residues, T denotes the frequency of the transition from one group to another and finally, D reflects the distribution of properties within 0, 25, 50, 75 and 100% of the sequence. The CTD-descriptor has been employed to predict protein folding classes [25].

Pseudo-amino acid composition

Sparse encoding and the amino acid composition do not take into account the sequence order effect. This effect considers the vast amount of possible amino acid combinations as the sequence length increases. That is, for a peptide of length 6, there are already $20^6 = 64{,}000{,}000$ different sequence arrangements. In terms of antimicrobial activity, Cherkasov et al. (2009) pointed out that, albeit having very similar amino acid compositions, some peptides were virtually inactive [26]. Thus, the pseudo-amino acid composition (PseAAC) has been introduced to consider the effect of the sequence order [27]. The PseAAC computes the correlation between different ranges among a pair of amino acids, which leads to a $20 + \lambda$ dimensional vector (Fig. 2a). The first 20

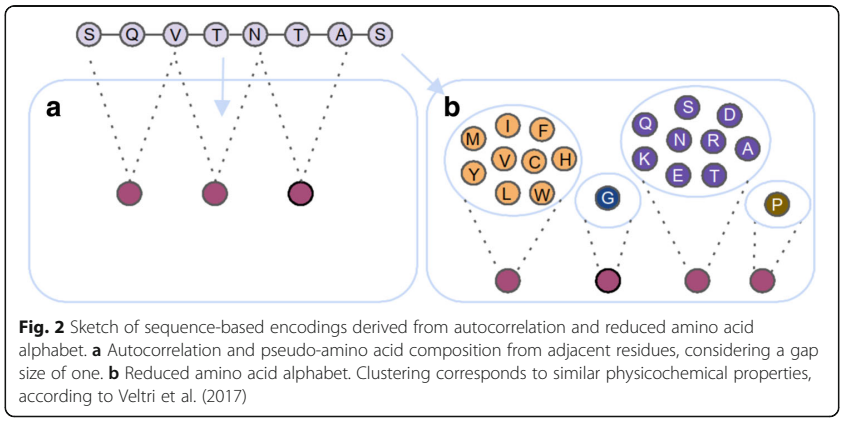


Fig. 2 Sketch of sequence-based encodings derived from autocorrelation and reduced amino acid alphabet. **a** Autocorrelation and pseudo-amino acid composition from adjacent residues, considering a gap size of one. **b** Reduced amino acid alphabet. Clustering corresponds to similar physicochemical properties, according to Veltri et al. (2017)

components are the composition of the 20 natural occurring amino acids, whereas the $20 + 1$ to $20 + \lambda$ components describe the correlation according to the respective sequence order level. For the most contiguous ($\lambda = 1$) and the second-most contiguous ($\lambda = 2$) amino acids, the PseAAC results in a 22-D (dimensional) vector. Thus, for $\lambda = 1$ the sequence order for adjacent amino acids are taken into account. The correlation function incorporates several physicochemical properties, such as the hydrophobicity and amino acid side chain mass. To verify that this method leads to a lower loss of information compared to the usual amino acid composition, several similarity measures have been employed. These include the prediction of subcellular locations of proteins, membrane protein types, as well as their particular locations [27]. To improve prediction accuracy, the PseAAC has been used by several studies, e.g., [28, 29], and [30], in combination with other types of encodings. For instance, in order to predict AMPs and additional efficiencies towards, e.g., cancer cells and HIV, PseAAC was applied in a two-level approach: first, it was used to encode peptide sequences to distinguish between AMPs and non-AMPs and second, to determine additional effects. Both classifications have been conducted by means of fuzzy k-nearest neighbors [28]. Moreover, additional physicochemical properties have been used to enhance the discriminative power of PseAAC [28]. Chen et al. (2016) tried to unveil novel anticancer peptides by enhancing the default dipeptide composition with PseAAC [29]. This approach considers long range interactions between amino acid pairs along with the dipeptide composition. The latter might reflect structural interactions, such as hydrogen bridge bonds between spatial close amino acids to form alpha helices [31]. An extension to the interaction of multiple encodings, including PseAAC, has been conducted by Meher et al. [30]. They used PseAAC in addition to structural and physicochemical encodings in order to distinguish between AMPs and non-AMPs. Again, an SVM was used to conduct the classification [30].

Reduced amino acid alphabet

Sparse encoding, amino acid composition, and PseAAC consider, more or less, the actual amino acid sequence to encode a peptide. Therefore, the encoding might not reflect sequence variations well and this might negatively contribute to the classifier performance. In order to improve generalization, also considering mutations, one could make use of the reduced amino acid alphabet. Here, similar amino acids are grouped together, based on physicochemical, such as hydrophobicity and hydrophilicity [9] or structural properties, e.g., the backbone structure (Fig. 2b) [32]. The reduced amino acid alphabet has been employed in combination with the n-gram model to ease the classification of protein sequences. The n-gram model counts the occurrences of n-mers for an alphabet of size m, leading to a m^n dimensional, sparse representation of the initial sequence (Fig. 3). Nevertheless, despite the preceding alphabet reduction, the increased dimensionality is again a major drawback of the n-gram model. Thus, single value decomposition [33] has been applied to reduce the number of features to efficiently train an artificial neural network (ANN). Finally, the ANN is used to assign the query proteins to the respective protein families [9]. Comparable to the n-gram model, the n-peptide composition leads, in particular for an increasing n, to an inflation of the feature space. Yu et al. (2004) used the n-peptide model to predict the subcellular

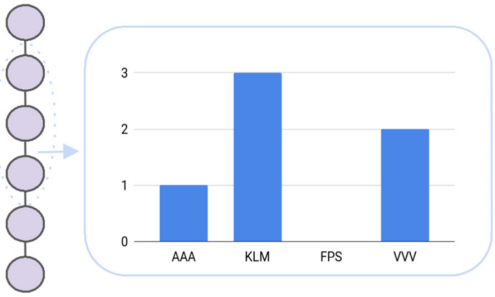


Fig. 3 Similarly to the amino acid composition, the k-mer composition counts the presence of k-mers. In this example k is set to three

location of proteins in Gram-negative bacteria [34]. For this purpose, the dipeptide, amino acid, as well as the partitioned amino acid composition have been leveraged. For the latter, the sequence is split into equal-length segments and these segments are used to train several SVMs. The assignment of the respective subcellular location is then based on a majority vote of all classifiers [34]. Furthermore, the reduction of the amino acid alphabet, based on structural properties, has been used as the initial step to construct more complex features. These complex features consist of compositional, positional, position-shifted, and correlated features, which are combined through several boolean functions, such as matches and/or matchesAtPosition. The ultimate goal of the study was the prediction of AMPs and their selectivity for different kinds of bacteria and to this end, the complex features are further reduced by means of a filter-based feature selection [35, 36]. Another study uses the k-nearest neighbor clustering to aggregate amino acids into five classes using their amino acid index, i.e., amino acids with the respective highest (T), high (H), medium (M), low (L), and lowest (B) values of a particular physicochemical property are clustered together. This encoding (AAIndex-Loc) is extended by the five-level dipeptide composition, which extends the aforementioned clustering by aggregating pairs of amino acids, such as TT, TH, and so forth. Along with these descriptors, Tantoso et al. (2008) employed the amino acid composition, for both termini and the middle part of the peptide, which leads to a dataset of 70 features for an SVM to predict subcellular location [37].

Physicochemical properties

One of the important encodings in AMP prediction, if not the most important one, is the translation of an amino acid to a particular physicochemical property, which have been determined in various wet lab experiments (Fig. 4a). The amino acid index database (AAindex) has been established as a unified source for these descriptors [38]. The AAindex is grouped into three parts, whereby the AAindex1 contains the just mentioned biochemical properties (one for each amino acid) and the AAindex2 aggregates different substitution matrices, such as the PAM250 or the BLOSUM62. The AAindex3 provides protein contact potentials, hence empiric values for spatial close amino acids, such as the Gibbs free energy change, to indicate preferred interactions between residue pairs [39]. The AAindex database, as a consistent source for numerical amino acids

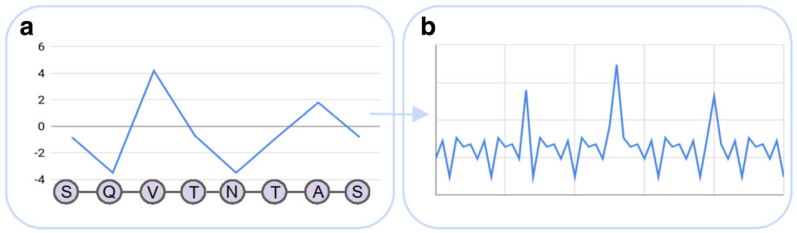


Fig. 4 Sketch of sequence-based encodings derived from physicochemical properties and Fourier transformation. **a** The numerical representations are based on the physicochemical properties of Serine (S), Glutamine (Q), Valine (V), Threonine (T), Asparagine (N) and Alanine (A). **b** Fourier transformation derived from the encoded peptide sequence

indices, has proven its usefulness in several studies. An example is the prediction of transmembrane protein segments [40]. Deber et al. (2001) used, among others, the hydrophobicity scale introduced by Kyte and Doolittle [41], as a reference to their experimental derived values of hydrophobicity [40]. The program annotates α-helical regions in the query sequence, based on the respective hydrophobicity and helix tendency thresholds [40]. So called z-descriptors have been employed as part of the prediction of cell-penetrating peptides [42]. These type of peptides reveal an important property, as they are capable to introduce macromolecules into the cell, which is especially interesting for the pharmaceutical industry [43]. The z-descriptors are derived from the principal components of physicochemical properties by means of partial least squares (PLS) projections [44]. PLS leads to a subset of five final features, capable to describe the 20 proteinogenic as well as 67 additional amino acids. The first three components can be considered as lipophilicity, volume (steric bulk), and polarity, respectively, whereas the fourth and the fifth component are not clearly derivable [44]. These properties are appropriate for the cell-penetrating peptide prediction, due to the intrinsic properties, which are the polarity (positively charged residues are advantageous) as well as the the amphi- and hydrophobicity [42]. However, Hansen et al. (2008) pointed out, that the method benefits from averaging z-descriptors, because that allows to compare sequences with varying length [42]. Nevertheless, to deal with varying protein or peptide sequence lengths, interpolation techniques have been introduced [45]. Sequence interpolation refers to a method, which connects multiple points, that is amino acid indices, via different linear and nonlinear functions. In order to obtain a continuous feature vector, the amino acid sequence is first mapped to the respective physicochemical property, followed by the actual smoothing, employing one of the interpolation functions [45, 46]. Physicochemical representations of peptides have been utilized to classify AMPs and non-AMPs [47]. To this end, Torrent et al. (2011) investigated the different characteristics of antimicrobial peptides, such as the isoelectric point, in-vivo aggregation, and hydrophobicity with respect to their discriminative power [47]. A peptide is described by its different characteristics and the particular averages were fed into an ANN to obtain the class to which the query peptide belongs [47]. In addition, the physicochemical property encoding is employed by various web servers for peptide retrieval, i.e., database queries, as well as for classification. Two examples are AVPpred [48] for antiviral peptide prediction and DBAASP for structure and activity of AMPs [49]. Moreover, this encoding has been used as part of several

other studies to predict antimicrobial effects of synthetic peptides [50] or to find substructures with antimicrobial potency in larger proteins [51]. In order to take into account that some traits of AMPs are dependent on particular parts within the sequence, such as a positively charged N-terminus, further studies elucidated the physicochemical property dependence with respect to different sequence sections. One of these studies divided AMPs into datasets for both termini, calculated the physicochemical representation, and finally uses an SVM for classification on the best performing feature subset [52]. Another study leverages pattern changes of amino acid characteristics along a peptide sequence for the prediction of antimicrobial peptides by means of random forests (RF) [53]. An alternative approach, which leverages hydrophobicity values, is designated as the d-descriptor [54]. This encoding is founded on sequence moments, a two dimensional extension of sequence profiles. The amino acid sequence is squeezed between the y- (N-terminus) and the x-axis (C-terminus) with gradually bending of the single amino acids and subsequent vector summation. The length of the vectors arise from the respective property and the angle results from the amino acids orientation in the 2D space. Finally, the sequence moments are mapped to scalar values, which is named the d-descriptor. Juretić et al. (2009) used the latter in order to estimate the therapeutic index, the ratio of hemolytic and antimicrobial activity [54]. Finally, owing to the high dimensional feature vectors, if one uses all possible amino acid indices, several studies, such as [52], performed statistical analysis in order to reduce the features before the accomplishment of the actual experiments. Other studies used techniques such as PCA to obtain the aforementioned z-descriptors as well as factor analysis in order to describe all amino acids with only five factors [55]. Recently, Boone et al. (2018) proposed a classification method by means of the rough set theory [56]. To this end, physicochemical properties have been used to encode the samples and afterwards the algorithm finds suitable boundaries to differentiate between antimicrobial and non-active peptides [56].

Autocorrelation

An approach to consider physicochemical properties not only for a specific position, but also for amino acids which might be related in higher dimensional protein structure assemblies, can be described by an encoding, which is known as autocorrelation. In general, autocorrelation describes the interdependence between two distant signals in a time series, whereby the distance or the lag, respectively, is predetermined and fixed for a particular computation (Fig. 2a). For amino acid sequences, repeating patterns, i.e., a certain periodicity, might be unveiled [57]. In peptide, or generally in protein science, two algorithms to detect spatial autocorrelation have been employed: the Moron autocorrelation, which considers the local dependence of amino acids [58] as well as the Broto-Moreau autocorrelation, which describes the global relationship of the residues [59]. These formulas yield either positive values, meaning that amino acids with similar physicochemical properties follow each other (positive autocorrelation) or negative values, i.e., amino acids with different physicochemical properties are interconnected (negative autocorrelation). Values near zero point to no or less autocorrelation [60]. One of the earliest applications of autocorrelation was the statistical analysis of protein content [60] and the prediction of α-helices [57]. A noteworthy relationship exists between autocorrelation and PseAAC,

since both take the sequence order effect into account, by measuring the correlation among amino acid pairs. Further advantages of this encoding are the reduction of the feature space as well as the normalization of the sequence length [61]. To this end, this descriptor has been utilized in several studies and facilitated, for instance, the prediction of mutation induced stability alterations of the gene V protein by bayesian-regularized genetic neural networks [61]. Another study dealt with protein-protein interactions and used the autocorrelation descriptor to train the rotation forest algorithm [58]. Furthermore, Kleandrova et al. (2016) used this encoding for the prediction of antimicrobial activity in known peptides as well as for screening of novel, artificial AMPs [59].

Substitution and scoring matrix

Substitution matrices, such as BLOSUM62 or PAM250, represent accepted mutations between amino acid pairs, i.e., sequence alterations with either no or positive impact in terms of the protein function. More specifically, it is the likelihood for a specific mutation within a certain time frame [62]. In contrast, the position-specific scoring matrix (PSSM) describes, based on a initial BLAST alignment, and iterative refinement, how amino acids are evolutionary conserved at a specific position. This results in positive values for a highly conserved residue and negative values for the others. Values near zero indicate weakly conserved residues [63]. Alignments with PSSMs can be regarded as an extension of substitution matrices, since instead of using, e.g., the PAM250, the PSSM is used for the alignment score, which leads to improved substitution probabilities and hence more sensitive alignments [64]. With regard to antimicrobial peptides, this encoding weights functional important residues stronger, such that conclusions for antimicrobial effects can be drawn and hereof facilitates querying peptides with unknown activity. For instance, the BLOMAP-encoding incorporates the BLOSUM62 to calculate distances in a high dimensional input space, i.e., the substitution matrix, to a lower dimension, using the Shannon-projection [65]. Maetschke et al. (2005) demonstrated how this descriptor improves signal peptide cleavage site prediction using, among others, Naïve Bayes (NB) and ANNs [65]. Due to the ambiguity of some BLOSUM50 entries, i.e., same values for amino acids, which in fact differ towards their physicochemical properties, Huang et al. (2005) utilized this substitution matrix in order to extend the sparse encoding [66]. They replaced each non-zero value with the respective BLOSUM50 score, such that the information of a particular amino acid is kept and additional information, derived from the substitution probabilities, is taken into account. The adjusted encoding has been used to predict T-cell epitopes by means of an SVM [66]. Karypis et al. (2006) applied substitution matrices to train SVMs for protein secondary structure prediction [67]. Therefore, k-mers are generated and mapped by means of the PSSM and BLOSUM62 matrices, respectively, to their numerical encoding. A binary SVM has been trained on this input and the results of this classification are used along with the aforementioned encoding for a second classification, which incorporates both [67]. Kumar et al. (2008) employed PSSMs as the encoding for a SVM to predict RNA binding sites in proteins [68]. Another study builds several SVMs using different encoding schemes, such as split-, dipeptide-, and regular amino acid composition together with PSSMs to enable the prediction of malaria parasite mitochondrial proteins [69]. Furthermore, the classification of bacterial virulent

proteins has been facilitated through the usage of sequence order effect conserving descriptors like PseAAC, the PSSM, and the above mentioned AAIndexLoc encoding. Nanni et al. (2012) used SVMs as well as an ensemble classifier approach for the final protein identification [70]. The latter employs a two-stage feature transformation method, which couples PCA and neighborhood preserving embedding, followed by decision trees [70]. In order to reveal DNA-binding proteins, Xu et al. (2015) extended PSSMs to incorporate dipeptide composition, which allows the computation of the probability of simultaneously appearing pairs of same and different amino acids within a certain distance along the peptide sequence [71].

Fourier transformation

Fourier Transformation (FT) can be used to detect underlying patterns in time series by transforming the time signal to a frequency domain (Fig. 4b) [72]. Examples for the application in biomedicine are the detection of the repeated occurring of coding and non-coding regions in DNA sequences and the prediction of cellular locations of proteins [73]. FT has been applied as part of a study to discover peptides with antimicrobial activity [73]. To this end, the residues have been first mapped to physicochemical properties, followed by the actual FT. Afterwards, the similarity between a reference peptide and potential hits has been measured by means of the Euclidean distance between the respective power spectra [73]. Moreover, Yin et al. (2017) proposed an approach to predict protein-protein interactions by means of discrete Fourier transformation (DFT) [74]. They showed, that the detection of coevolution patterns can be carried out without using multiple sequence alignments. Again, hydrophobicity values have been used to encode the amino acid sequences. Afterwards, subsequences have been extracted with a sliding window approach and transformed via DFT. Based on the DFT results, the evolutionary distances between proteins were calculated using the Euclidean metric. Finally, a protein-protein interaction is indicated by means of the Pearson correlation coefficient [74].

Structure based encodings

The secondary structure of a protein or peptide, respectively, is mainly determined by its primary structure, i.e., the order of the amino acids [75]. Moreover, the peptide structure has a strong correlation with antimicrobial activity [76]. Thus, for the prediction of antimicrobial activity, it is reasonable to use sequence-based encodings, but, since the secondary structure cannot be completely derived from the primary structure, it is also conclusive to develop structure-based encodings. In addition, the employing of both descriptors simultaneously, allows the classifier a better generalization and thus improves the overall accuracy [77]. The following section introduces several applications of structure-based encodings.

Quantitative structure-activity relationship

An alternative approach to describe amino acids sequences by their chemical properties has been developed as part of quantitative structure-activity relationship (QSAR) studies. In essence, QSAR refers to the prediction of a particular property or activity by means of its molecular characteristics and structure [78]. This is also the crucial

difference between the description of amino acids by their physicochemical properties and QSAR. The latter focuses solely on molecules, whereas the former encodes the whole residue. In addition, QSAR is mainly applied in chemoinformatics for high-throughput screening, i.e., to find novel active substances in databases using two- and three-dimensional representations of compounds [79]. However, several studies propose QSAR modeling based approaches to predict antimicrobial activity. For instance, one study uses this encoding[1] to imitate the artificial AMP Novispirin G_{10} by similar peptides in order to enhance its potency. Here, molecular modeling was used to calculate 3D structure conformations. The structure was then used to obtain a set of descriptors, such as hydrophobicity, amphipathicity, and electrostatic charges. Finally, a subset of meaningful features have been determined and the activity measurement of the analogs was determined by predicting the amount of inhibited bacterial growth [80]. Moreover, Bhonsle et al. (2007) aimed to find informative 3D physicochemical descriptors in order to predict bioactivity of AMPs [81]. Solvent-accessible surface describing (e.g., fractional charged partial surface area), structural (H-bond acceptor) and spatial (density) descriptors, among others, turned out to be good indicators for antimicrobial activity [81]. Jenssen et al. (2007) investigated, whether there is a set of molecular descriptors, which can be used to optimize antimicrobial activity against *P. aeruginosa* [82]. This set encompasses the aforementioned z-descriptors as well as the contact energy between amino acids, inductive and conventional QSAR descriptors [82]. Similar descriptors have been evaluated in order to design AMPs in silico [83]. Shu et al. (2013) uses PCA to extract the first six principal components from topological and structural characteristics to predict antimicrobial activity of synthetic cationic polypeptides [84]. In contrast, Schneider et al. (2017) utilized molecular descriptors to train self-organizing maps (SOM) [85]. Afterwards, the continuous SOM responses are adjusted by means of lateral inhibition and utilized as input for a deep learning model in order to predict helical AMPs [85].

General structural encodings

Unlike QSAR-based methods, general structural encodings map structure information derived from the whole peptide, to a numerical representation. The peptide structure is described by means of their total 3D shape. This is contrary to QSAR, because instead encoding an amino acid sequence from a molecular viewpoint, the whole peptide structure is considered (Fig. 5a). For instance, Cui et al. (2008) predicted the secretion of proteins into the bloodstream [86]. They used features including physicochemical properties as well as structural information, such as secondary structure, and solvent accessibility. The final prediction has been facilitated by an SVM [86]. Chang et al. (2015) employed conditional random fields (CRF) for probability prediction of critical regions along an AMP sequence [87]. CRFs are an algorithm similar to hidden Markov models, but more variables, such as the surrounding context, can be incorporated. In the present case, several structural descriptors along with physicochemical properties have been used for the prediction. The structure-based encodings encompasses the assignment of predicted secondary structure, conserved protein domains, predicted antimicrobial regions [88] as well as the aggregation tendency [87]. Dybowski et al. (2010) proposed a stacked classifier model to predict the HIV-1 tropism [89]. To this end, the

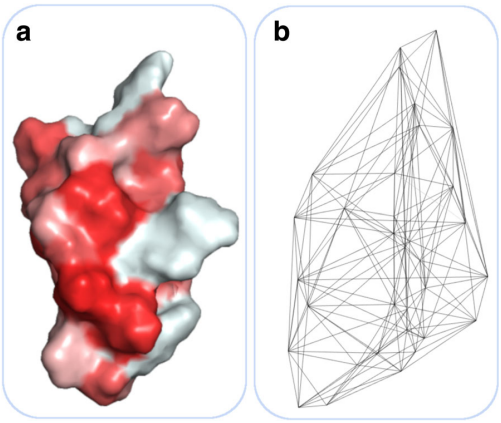


Fig. 5 Exemplary structure-based encodings for antimicrobial peptide Human Defensin 5 (PDB:2LXZ). **a** Solvent accessible surface. Color coding according to hydrophobicity scale (Eisenberg et al., 1984) **b** Delaunay triangulation of the same peptide calculated from Cα-atoms. Bose et al. (2011) used the summed distances between amino acid pairs to encode protein structure

authors trained two independent RFs, whereby the first used hydrophobicity values and the second used the hulls of the electrostatic potentials of the V3 loop, a short peptidic sequence of the viral gp120 protein, as descriptors. The electrostatic hull has been determined in order to acknowledge even subtle differences between different co-receptor tropisms as well as to wrap superimposed shapes of the peptides sub-structure. A third RF combined the output of the other models for the final class assignment [89]. Due to high computational effort during the calculation of the electrostatic potential, Heider et al. (2014) presented an extension of this method [90]. The authors leveraged, that the current model achieves good performance with a constant dielectric value and ionic strength, thus simplifying the calculation of the potential to Coulomb's law. Finally, the electrostatic potential has been calculated based on the cluster centers. The centroids are determined by all points within a certain distance to the C_α-atoms of the V3 loop [90]. As part of another study, the authors increased the prediction power by means of multiple RFs, combined to an ensemble classifier. The respective classifiers used physicochemical as well as structural properties to predict resistance against a novel HIV-1 maturation inhibitor.

The structural encoding is based on the aforementioned electrostatic potential. In addition, a genetic algorithm has been implemented to find an optimal subset of the physicochemical properties [17]. However, Bozek et al. (2013) pointed out, that the structural encoding of the V3 loop exhibits limitations, since only two physicochemical properties has been used for description [91]. To this end, they proposed a novel encoding, which incorporates structural variations as consequence of sequential rearrangements. Thus, based on the template structure, spheres, whose centers are depicted by reference atoms, are used to enclose spatial related residues of different loop variants. Afterwards, the averaged physicochemical properties of all residues within these regions are used to determine HIV-1 co-receptor usage [91]. In contrast, Sander et al. (2007) introduced an distance distribution approach in order to improve co-receptor tropism based on V3 loops [92]. This method calculates the euclidean

distances between each atom type. Afterwards, the respective distances are used to obtain the underlying distribution. Finally, the feature vector is obtained by sampling from this distribution, leading to a final size of each possible combination times samples [92]. Nevertheless, HIV-1 is a very complex organism and hence, several strategies have been tackled in order to combat the virus, such as the aforementioned relation between the V3 loop and tropism as well as between mutations, structure and drug resistance [93]. To this end, another encoding has been developed to describe protein structure based on Delaunay triangulation (Fig. 5b). In essence, the Delaunay triangulation states that, if three points or vertices, respectively, are connected via edges, no further vertex must be located within the circumcircle of these three vertices. This encoding facilitates to encode the complete protein shape by finding the optimal edges between representative points, such as C_α-atoms. Thus, it is able to incorporate information about spatial close residues, which might be lost by a descriptor based on the primary structure only. Finally, the feature vector consists of 210 entries, derived from the adjacency matrix of all amino acid pairs. The respective values are resulting from the averaged distance among these pairs [94]. Albeit this encoding has been mainly employed in the context of computational HIV research, it might work as well for antimicrobial peptides, owing to very good classification results of several studies [95]. To sum up, structural encodings are an appropriate extension to sequence-based encodings since antimicrobial activity is determined by the three-dimensional composition of the residues [96] and in addition, the combination of sequence- and structure-based encodings increases discriminating power [97].

Alternative encodings

There are further encodings, which do not really fit into the proposed categories, i.e., sequence or structural encodings. One of these encodings, which are summarized in Table 3, is the Chaos Game Representation (CGR). In general, the CGR is a visual encoding of a sequence, generating a fractal. The sequence can be obtained, e.g., from random numbers or from biological sequences, such as bases (DNA) and amino acids (proteins). In the case of the former, numbers from 1 to 3 denoting a vertex of a triangle. The algorithm works as follows: firstly, a starting point s is determined and afterwards, one of the numbers is randomly selected as the target vertex t. The next point is located on the half way between s and t. By repeating this procedure, the so called Sierpinski triangle will be generated. The Sierpinski triangle is special about its recursively defined sub-structures, which are also triangles [98]. In the case of the DNA, t is not selected by chance, but rather by the successive base. Here, adenine (A), thymine (T), guanine (G) and cytosine (C) are the labels of a square. After conducting the algorithm, the resulting fractal shows lower order, but still exhibits notably patterns, originated from the underlying sequence. Moreover, points which are close in the CGR do not have to be necessarily adjacent in the sequence, which means that the CGR might introduce novel distance metrics of subsequences [98]. However, with respect to AMPs, CGR has been applied as part of a variety of studies in order to deal with amino acid sequences. As such, Basu et al. (1997) classified similar amino acids to 12 different groups, each representing a target vertex for the CGR algorithm [99]. In addition, the resulting dodecagon has been divided in 24 grids and the amount of points per grid has

been used to predict the affiliation to protein families [99]. A further study reduced the amount of vertices to 8, whereby the grouping happened according to the respective physicochemical properties [100]. Moreover, He et al. (2016) extended the illustration to three dimensions, which results in a cube, rather than a planar octagon [100]. The study investigated how this encoding could be employed for multiple sequence alignments. To this end, the authors introduced a method, which computes the euclidean distance between amino acid pairs of two encoded proteins. Finally, the similarity of two proteins is denoted by the sum of the distances [100]. Recently, one study used CGR in a 10D space, using a hypercube for the prediction of anticancer peptides [101] as well as for protein-protein interactions [102].

Another method, which does not fit into the proposed sections has been introduced by Loose et al. (2006) in order to design novel AMPs [103]. In this study, the authors considered AMPs as a corpus of sentences and the goal was to examine, whether antimicrobial activity is described by a certain grammar. To this end, a linguistic model has been derived from active peptides and successfully employed for the design of AMPs [103].

Models

So far, state of the art encodings have been discussed extensively. The next section will summarize the utilized learning algorithms. Popular models in antimicrobial peptide prediction include decision trees [21, 50, 71] and random forests [17, 53, 104, 105], but also neural networks have been employed in several studies [9, 26, 106]. Moreover, deep learning, as an extension to ordinary neural networks, has been applied frequently and thus a more detailed description, along with a summary in Table 4, is provided in the next section. Support vector machines are a further outstanding model in AMP prediction and were part of several studies [29, 30, 91]. In fact, there are specific kernels designed for amino acid based proteins/peptides sequences, known as string kernels. To shed some light into this topic, the upcoming section will highlight these kernels in more detail. In addition, Table 5 summarizes the presented kernels. However, besides the popular algorithms mentioned above, further methods leveraged partial least squares [82, 83, 107], hidden Markov models [108], logistic regression [109] and Bayesian networks [110]. Furthermore, ensembles of several classifiers have been also successfully implemented, such as in [17] or [21], whereby often one classifier is trained with a particular sequence or structural encoding. As part of an optimized feature set construction, genetic algorithms have been employed, by, e.g., Kernytsky et al. (2009) [111] as well as Veltri et al. (2017) [36]. Moreover, Krause et al. (2018) made use of genetic algorithms to optimize cell-penetrating peptides [43].

Table 4 Different encodings from deep learning models (see Table 1 for details)

Encoding	Description	Summary	Used in	Used along with
ProtVec	amino acid sequences are encoded as a distributed representation of k-mers	Density: + Information: +	[124]	
Voxel	structures of proteins are encoded as voxels	Density: o Information: +	[125, 126]	
Matrix	mimicks images by regarding the respective entries of PSSMs as pixel densities	Density: o Information: +	[127, 129, 130]	PSSM
Autoencoder	extracts representative characteristics in order to reproduce the input as good as possible	Density: + Information: o	[131]	

Table 5 Different types of string kernels (see Table 1 for further details)

Encoding	Description	Summary	Used in	Used along with
Spectrum Kernel	generates all possible subsequences of length k and counts the occurrences of these k-mers	Density: - Information: -	[112]	
Mismatch Kernel	considers a certain distance, hence mismatches, between two k-mers	Density: - Information: o	[114–116]	General Structure
Distant Segment Kernel	allows a gap between two k-mers	Density: - Information: o	[118]	
Local Alignment Kernel	obtained from local alignment scores	Density: + Information: o	[119]	Spectrum Kernel, Mismatch Subsequence Kernel
Subsequence Kernel	measures sequence similarity, gaps within k-mers are taken into account	Density: + Information: o	[119]	Frequency of Amino Acid Pairs
Frequency of Amino Acid Pairs	similar to dipeptide composition	Density: - Information: o	[119]	
String Kernels + Physicochemical Properties	optimization of existing string kernels such that these involve physicochemical properties	Density: + Information: +	[120]	Physicochemical Properties
Generic String Kernel	string kernel with physicochemical properties and penalization of non adjacent segments	Density: + Information: +	[121, 122]	

String kernel

Support vector machines (SVM) are capable to efficiently distinguish between binary input data by projecting the data to a higher input space, using kernel techniques [112]. Moreover, these kernel techniques allow a linear separation of a nonlinear classification problem, which is also known as the kernel trick [113]. One type of these kernels are string kernels, which are employed to measure sequence similarity [112]. In essence, the idea of string kernels implies that strings are mapped to a numerical representation in order to be used as input for an SVM. Thus, it is basically another encoding of an amino acid sequence, i.e., a method to map the string representation of peptide sequences to high dimensional feature vectors. Hence, several studies proposed corresponding methods, such as Leslie et al. (2002), who extended the spectrum kernel, in order to incorporate sequence variations, to the mismatch kernel [112]. The former generates all possible subsequences of length k and counts the occurrences of these k-mers within the query sequences, leading to a similarity metric based on shared k-mers [112]. This encoding is similar to the k-peptide composition, for instance the dipeptide composition (k = 2), which has been introduced earlier. The mismatch kernel on the other hand, considers a certain distance, hence mismatches, between two k-mers and takes into account, that similar sequences might have similar properties. Owing to the nature of spectrum kernels, further investigations revealed important and meaningful motifs. As a case study, the authors predicted homolog proteins [114]. Furthermore, string kernels have been applied to predict tumor suppressors, among others. Here, small molecules are encoded in their 1D, 2D, and 3D representations. In 1D, mismatch kernels have been employed to measure the similarity between the atomic sequences [115]. Another study investigated the performance of combined as well as weighted mismatch and structure derived similarity score kernels [116]. For these

kernels, each entry in the feature vector is obtained from structure alignments between the input peptide and a peptide database [117]. The encoding incorporates the similarity to further peptides, whereby conserved peptides are depicted with higher scores. Boisvert et al. (2008) proposed an extension of the string kernel, which allows a gap between two k-mers [118]. Thus, the distant segment kernel takes into account the co-occurrence of remote sequence segments. The authors used this kernel in order to predict HIV-1 co-receptor tropism and achieved higher levels of accuracy compared to other methods [118]. Moreover, several string kernels have been employed and compared to predict linear B-cell epitopes [119]. These include the already introduced spectrum and mismatch kernel as well as the local alignment kernel, obtained from local alignment scores, and the subsequence kernel, which measures sequence similarity, similar to the mismatch kernel, albeit gaps within k-mers are taken into account. A third kernel measures the frequency of amino acid pairs (see dipeptide composition), which is due to a bias towards certain dipeptides in B-cell epitopes [119]. Toussaint et al. (2010) recognized that dealing with the sequence only might result in a loss of information [120]. For this reason, the aim of their study was the optimization of existing string kernels such that these involve physicochemical properties [120]. This kernel has been used by another study in conjunction with the penalization of non-adjacent segments, which finally has led to the generic string kernel for small molecules [121]. The authors applied this kernel in a subsequent study in order to detect antimicrobial peptides. All possible peptides with a specific length have been generated by means of source-to-sink graphs. In these graphs, all vertices are k-mers and all edges are weighted according to the antimicrobial activity, computed by means of the generic spectrum kernel. Finally, the detection of the most active peptide corresponds to the detection of the longest path within the graph [122].

Deep learning

Machine learning algorithms based on artificial neural networks, especially deep learning models, have the advantage of incorporating automated encoding, i.e., feature generation. In general, the encoding results from several, successive connected layers, which work as filters for particular parts of the input [5]. However, these models require a large number of training examples in order to generalize well. Fortunately, owing to advances in next-generation sequencing technologies, biological sequences, such as peptides and proteins, are publicly available in vast amounts [123]. Several studies made use of that and showed how deep neural networks perform well on biological problems. For instance, Asgari et al. (2015) proposed a method called protein-vectors, which splits a sequence into k-mers to learn the context of these word representations [124]. Here, amino acid sequences are encoded as a distributed representation of k-mers, which were employed for protein family classification or the prediction of disordered proteins. This approach is derived from natural language processing and uses the context, hence the adjacent residues, for the central k-mers ("words") syntactic and semantic description. The realization is carried out through building a sufficient large training corpus of protein sequences ("sentences") by breaking all available sequences into overlapping k-mers. Afterwards, neural networks are used to find optimal, numerical representations, i.e., feature vectors, of the input

sequences by means of the skip-gram model. By using these vectors, the authors showed that this framework encodes physicochemical properties well and high levels of accuracy have been achieved in the family classification task [124]. Jiménez et al. (2017) utilized deep learning to predict protein-binding sites [125]. To this end, the structures of proteins are encoded as three-dimensional objects, whereby a cubic segmentation in so-called voxels, which are 3D pixels, takes place beforehand. The encoding of each of these cubes is based on the contained atoms. In order to incorporate physicochemical properties, the input is further upscaled to 8 property channels [125]. A similar approach has been elaborated by Amidi et al. (2018) to predict enzyme classes [126]. Again, protein structures are encoded as voxels and are used as input for a convolutional neural network (CNN), but in contrast to Jiménez et al., the orientation of the protein has been considered. The authors point out, that the structure orientation in the Protein Data Bank (PDB) does not capture the dynamic of the protein and consequently used the proteins barycenter as origin and the first principal components for the orientation of the coordinate system. Overall, the model achieves good accuracy [126]. Another study uses position-specific scoring matrices (PSSM) as 2D input for CNNs, hence mimicking images by regarding the respective substitution probabilities as pixel densities. The studies goal is the automated partitioning of efflux proteins families [127]. This class of proteins provide an important tool for multi-resistant pathogens, because they allow them to convey molecules out of the cell, thus lowering the overall concentration of antibiotics [128]. Two further publications deal with alignment-free comparison of sequences, using CNNs. Both methods encode the input sequences as two-dimensional one-hot matrices, leveraging the convolutional layers for unveiling of latent features. Seo et al. (2018) employed this approach in order to predict protein families [129]. However, Zheng et al. (2018) extended this approach by training of two identical neural networks (siamese neural networks), which allows to compare sequences with respect to their dissimilarity [130]. These two methods, as well as the earlier introduced ProtVec [124], have in common that they aggregate amino acid sequences of varying lengths to a fixed-length numeric vector of lower dimension. Since this feature reduction keeps intrinsic properties of the proteins, these algorithms might serve as potential encodings for AMPs. Similar to this CNN based dimension reduction are autoencoders. Autoencoders are applied to learn a dense representation of the input, i.e., to extract representative characteristics in order to reproduce the input as good as possible. For instance, Wang et al. (2017) employed stacked autoencoders to predict protein-protein interactions [131].

Databases and packages

Having access to existing data sets is crucial to push computational, antimicrobial peptide prediction further. Thus, several projects aim to enable researchers a public database to active peptides. Consequently, this part introduces established databases and highlights some characteristics of these web services. Although data access is granted, there are still a plenty of possible encodings for testing. Fortunately, there are ready-to-use implementations of many encodings and the subsequent section lists a choice of these handy packages.

Databases

Piotto et al. (2012) presented YADAMP (yet another database of antimicrobial peptides) [132]. The authors collected the data sets, i.e., AMPs, from various, published studies. Potential hits can be limited, e.g., by specifying certain physicochemical properties and/or target organisms. Respective results provide more details with respect to activity and structural properties [132]. CAMP (collection of antimicrobial peptides) obtains AMP sequences and structures from well-known protein databases, such as UniProtKB [133]. Active peptides have been filtered out via keyword search. By providing several links to further web services, CAMP is a comprehensive resource for AMPs as well as active peptides in general [133, 134]. Wang et al. (2016) published the third update for the antimicrobial peptide database (APD3) [135]. Besides its focus on natural occurring AMPs, this database stores various active peptides, e.g., anti-HIV, spermicidal, and for wound healing. A web form lets the user specify custom query parameters, such as physicochemical properties [135]. Pirtskhalava et al. (2016) extended the database of antimicrobial activity and structure of peptides to the second version (DBAASPv.2) [49]. The service provides, among further details, potency values against several pathogens, described by inhibition coefficients. Moreover, the authors conducted molecular modeling for unveiling unknown structures of AMPs [49]. Finally, a comprehensive data repository of antimicrobial peptides (DRAMP) has been set up by Fan et al. (2016) [136]. They included additional features, hence similarity search, sequence alignment, and conserved domain search, besides established tools, which already have been introduced by other [136]. More information about web services for AMP retrieval can be found in two recent studies, published by Porto et al. [137] and Gabere et al. [138].

Packages

As mentioned before, many of the sequence-based encodings have been implemented in user-friendly packages, using, e.g., R[2] or Python.[3] Interpol is an R-package for normalizing peptide sequences to a uniform length, using different interpolation methods and descriptors of the AAindex database [45]. Cao et al. (2013) developed propy, which provides Python access to methods for amino acid composition, autocorrelation and pseudo-amino acid composition (PseAAC), among others [139]. In contrast, protr, implemented by Xiao et al. (2015), provides similar methods for the R programming language [140]. In addition, all methods can be accessed through a public web server. However, the web interface lacks the possibility of passing custom method parameters and is hence only recommended for ad-hoc calculations [140]. Ofer et al. (2015) released ProFET, i.e., protein feature engineering toolkit, a Python-based distribution with a variety of ready-to-use amino acid encodings [141]. Among default encodings, which have been implemented by others, this package offers also reduced amino acid alphabet, autocorrelation, amino acid propensities, as well as transformed CTD features [141]. modlAMP is a Python library specifically developed for antimicrobial peptides. Besides a selective choice of encodings, Müller et al. (2017) added methods for the whole prediction pipeline, i.e., sequence retrieval, visualization, and machine learning algorithms [142]. Moreover, performant model parameters can be obtained automatically via a grid search [142]. In contrast, POSSUM (position-specific scoring matrix-based feature generator for machine learning) is a

toolkit, which facilitates the representation of amino acids with PSSM derived encodings [143]. Wang et al. (2017) published POSSUM as a public web server as well as a Perl/Python-based tool, executable via the command line [143]. PyBioMed is another Python library foremost aiming cheminformaticians, owing to the fact, that many molecular encodings are implemented, e.g., topological descriptors, applicable in QSAR studies [144]. Nevertheless, Dong et al. (2018) rounded out this package with a variety of amino acid encodings and additional tools, such as sequence and structure retrieval [144]. Recently, Chen et al. (2018) published iFeature, which is accessible as a Python package and web server [145]. This tool adds functionality in order to encode amino acids based on AAindex entries as well as structure-based encodings, such as accessible surface area and main-chain-torsional angles. Moreover, algorithms for clustering, feature selection, and dimensionality reduction are available [145].

Encoding selection

It is quite challenging to find a suitable encoding within the variety of possibilities, thus, this section provides recommendations for the selection process. This might be helpful for computational biologists, due to the fact, that, as far as we know, no guidance of an appropriate encoding selection has been published until now. Unfortunately, it is not easy to provide generally applicable processes, which encoding will work for a particular application, thus we follow the approach from Heider et al. (2014) [90] and propose the measurement of diversity as a rule of thumb [146], until more sophisticated techniques have been unveiled. In order to calculate the diversity, it is necessary to train various classifiers on different encoded peptide data sets and combine the outputs. In particular, the diversity is based on the decision of single classifiers with their respective strengths and weaknesses. Thus, we suggest to conduct the encoding selection in such a way, that the ensemble maximizes the disagreement measure D, which is the probability of the disagreement between the classifier i and j, which minimizes the correlation of two classifiers i and j, as well as maintains the overall prediction accuracy [90]. The disagreement measure D is defined as:

$$D_{i,j} = \frac{1}{n} * \sum_{n}^{k=1} \mid o_k^i - o_k^j \mid$$

Here, o^i and o^j refer to the outputs of classifier i and j. Furthermore, we recommend to combine sequence and structure based encodings. For more details we refer to [90]. A comprehensive introduction into the diversity of classifier ensembles can be found in [146].

Conclusions

The amount of effort that has been expended in the last decades, demonstrates how important and essential efficient encodings are for detection of peptides with antimicrobial activity. This is reflected by diverse approaches and methods, which have been proposed in numerous publications. In the current study, we tried to aggregate existing, useful encodings and models, specifically for antimicrobial peptide (AMP) classification for multi-resistant pathogens. But also as part of other protein or peptide studies, respectively, promising encodings have been developed. In particular, sequence- and structure-based encodings have been discussed along with their applications. As

part of sequence representations, major encoding schemes as well as different customizations are introduced. Moreover, structural encodings encompassed molecular as well as general representations and a particular focus was set again on application dependent customizations. Finally, a selection of alternative encodings, beyond sequence- and structure-based encodings, are presented. The second part highlighted employed models as well as string kernels as encodings for support vector machines. Deep learning is a popular machine learning method and requires little or no encoding for the classification process. Nevertheless, exciting applications in protein research can be found in literature and thus, have been covered as well. As mentioned at the beginning, this review summarized encodings specifically for AMPs, however, every machine learning based protein/peptide classification task can be tackled by means of the proposed techniques. Moreover, to enhance research capabilities, several studies already implemented many of the reviewed encodings and published ready-to-use packages in commonly used programming languages. Again, this review collected most popular ones and provides an unified source of these. In order to lower obstacles further, we added a separate section about existing antimicrobial sequence databases. In conclusion, this review provides a common basis of methodologies in theory as well as practical tools to promote AMP research. Due to the fact, that we emphasized on encodings derived from AMP classification tasks, it is not surprising, that a large number of further techniques for amino acid representation exist, which, for obvious reasons, could not covered in this review. Moreover, additional research is required in order to incorporate the structure of AMPs and to examine whether the simultaneous encoding of sequence and structure can increase the prediction performance further. Nevertheless, many studies showed already at this point very good results. The engineering of amino acid encodings supports not only the detection of novel AMPs and consequently the battle against multi-resistant pathogens, but could also impact other major diseases, such as HIV and cancer. Research must be continued in each direction, in order to leverage the full potential of AMPs. To this end, besides the aforementioned simultaneous deployment of sequence- and structure based encodings, we propose further approaches. Delaunay triangulation is a promising encoding for peptide structure. By integrating additional information, e.g., physicochemical properties, to the graph, one could leverage advantages of both. In order to ease the access, this, as well as structure encodings in general, might be provided in a separate library. Moreover, since implementations exist for R and Python and each language provides a unique set of encodings, it is beneficial to develop a package, which provides those, that are not covered by an existing one. Finally, a comparative study is necessary to examine the potential of single encodings on a range of independent, biomedical data sets. Thus, encodings could be revealed, which are preferable for a designated application.

Endnotes

[1]Since QSAR actually refers to the general model, the abbreviation will be used from now on interchangeable with the molecule property encodings.

[2]https://www.r-project.org/

[3]https://www.python.org/

Abbreviations
AMP: Antimicrobial peptide; ANN: Artificial neural network; CGR: Chaos game representation; CNN: Convolutional neural network; CRF: Conditional random fields; DFT: Discrete fourier transformation; FT: Fourier transformation; ID: Increment of diversity; NB: Naïve Bayes; PDB: Protein data bank; PLS: Partial least squares; PseAAC: Pseudo-amino acid composition; PSSM: Position-specific scoring matrix; QSAR: Quantitative structure-activity relationship; RF: Random forest; SOM: Self-organizing maps; SVM: Support vector machine

Acknowledgements
We thank our group members Theodor Sperlea and Franziska Löchel for helpful suggestions as well as enlightening discussions.

Funding
Not applicable.

Availability of data and materials
Not applicable

Authors' contributions
SS developed the concept and wrote the manuscript. DH gave conceptual advice, supervised the study, and revised the final draft. Both authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 21 December 2018 Accepted: 24 February 2019
Published online: 04 March 2019

References
1. Mahlapuu M, Håkansson J, Ringstad L, Björn C. Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. Front Cell Infect Microbiol. 2016;6:194.
2. Roca I, Akova M, Baquero F, Carlet J, Cavaleri M, Coenen S, et al. The global threat of antimicrobial resistance: science for intervention. New Microbes New Infect. 2015;6:22–9.
3. Nellums LB, Thompson H, Holmes A, Castro-Sánchez E, Otter JA, Norredam M, et al. Antimicrobial resistance among migrants in Europe: a systematic review and meta-analysis. Lancet Infect Dis. 2018;18:796–811.
4. Li Y, Xiang Q, Zhang Q, Huang Y, Su Z. Overview on the recent study of antimicrobial peptides: Origins, functions, relative mechanisms and application. Peptides. 2012;37:207–15.
5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
6. Chen C, Seff A, Kornhauser A, Xiao J. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. 2015 IEEE International Conference on Computer Vision (ICCV); 2015. p. 2722–30.
7. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018;173:338–54 e15.
8. Wang Z. APD: the Antimicrobial Peptide Database. Nucleic Acids Res. 2004;32:590D–592.
9. Wu C, Berry M, Shivakumar S, McLarty J. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. Mach Learn. 1995;21:177–93.
10. Cruz J, Ortiz C, Guzmán F, Fernández-Lafuente R, Torres R. Antimicrobial Peptides: Promising Compounds Against Pathogenic Microorganisms. Curr Med Chem. 2014;21:2299–321.
11. Lee EY, Lee MW, Fulan BM, Ferguson AL, Wong GCL. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus. 2017;7:20160153.
12. Guilhelmelli F, Vilela N, Albuquerque P. da S. Derengowski L, Silva-Pereira I, Kyaw CM. Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. Front Microbiol. 2013;4:1–12.
13. Mookherjee N, Hamill P, Gardy J, Blimkie D, Falsafi R, Chikatamarla A, et al. Systems biology evaluation of immune responses induced by human host defence peptide LL-37 in mononuclear cells. Mol Biosyst. 2009;5:483–96.
14. Hancock REW, Haney EF, Gill EE. The immunology of host defence peptides: beyond antimicrobial activity. Nat Rev Immunol. 2016;16:321–34.
15. Hirst JD, Sternberg MJ. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry. 1992;31:7211–8.
16. Heider D, Verheyen J, Hoffmann D. Predicting Bevirimat resistance of HIV-1 from genotype. BMC Bioinformatics. 2010;11:37.
17. Dybowski JN, Riemenschneider M, Hauke S, Pyka M, Verheyen J, Hoffmann D, et al. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. BioData Min. 2011;4:26.

18. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. In: Springer Science & Business Media; 2013.
19. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12:1007–17.
20. Nagpal G, Chaudhary K, Agrawal P, Raghava GPS. Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. J Transl Med. 2018;16:181.
21. Usmani SS, Bhalla S, Raghava GPS. Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. Front Pharmacol. 2018;9:954.
22. Matsuda S, Vert J-P, Saigo H, Ueda N, Toh H, Akutsu T. A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci. 2005;14:2804–13.
23. Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides. BMC Bioinformatics. 2007;8:1–10.
24. Chen W, Luo L. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. J Microbiol Methods. 2009;78:94–6.
25. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A. 1995;92:8700–4.
26. Cherkasov A, Hilpert K, Jenssen H, Fjell CD, Waldbrook M, Mullaly SC, et al. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. ACS Chem Biol. 2009;4:65–74.
27. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001;43:246–55.
28. Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013;436:168–77.
29. Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7:16895–909.
30. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep. 2017;7:42362.
31. Ding H, Feng P-M, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol Biosyst. 2014;10:2229–35.
32. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins. 2000;38:149–64.
33. Das B, Turkoglu I. A novel numerical mapping method based on entropy for digitizing DNA sequences. Neural Comput Appl. 2017;29:207–15.
34. Yu C-S, Lin C-J, Hwang J-K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 2004;13:1402–6.
35. Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the Twentieth International Conference on Machine Learning; 2003.
36. Veltri D, Kamath U, Shehu A. Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. IEEE/ACM Trans Comput Biol Bioinform. 2017;14:300–13.
37. Tantoso E, Li K-B. AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids. 2008;35:345–53.
38. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36:D202–5.
39. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules. 1976;9:945–50.
40. Deber CM, Wang C, Liu LP, Prior AS, Agrawal S, Muskat BL, et al. TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. Protein Sci. 2001;10:212–9.
41. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157:105–32.
42. Hansen M, Kilk K, Langel U. Predicting cell-penetrating peptides. Adv Drug Deliv Rev. 2008;60:572–9.
43. Krause T, Röckendorf N, El-Sourani N, Ramaker K, Henkel M, Hauke S, et al. Breeding Cell Penetrating Peptides: Optimization of Cellular Uptake by a Function-Driven Evolutionary Process. Bioconjug Chem. 2018.
44. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 1998;41:2481–91.
45. Heider D, Hoffmann D. Interpol: An R package for preprocessing of protein sequences. BioData Min. 2011;4:1–6.
46. Heider D, Verheyen J, Hoffmann D. Machine learning on normalized protein sequences. BMC Res Notes. 2011;4:94.
47. Torrent M, Andreu D, Nogués VM, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PLoS One. 2011;6:e16968.
48. Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res. 2012;40:W199–204.
49. Pirtskhalava M, Gabrielian A, Cruz P, Griggs HL, Squires RB, Hurt DE, et al. DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucleic Acids Res. 2016;44:6503.
50. Lira F, Perez PS, Baranauskas JA, Nozawa SR. Prediction of antimicrobial activity of synthetic peptides by a decision tree model. Appl Environ Microbiol. 2013;79:3156–9.
51. Pane K, Durante L, Crescenzi O, Cafaro V, Pizzo E, Varcamonti M, et al. Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: Application to the detection of "cryptic" antimicrobial peptides. J Theor Biol. 2017;419:254–65.
52. Veltri D, Shehu A. Physicochemical Determinants of Antimicrobial Activity. In: Intl Conf on Bioinf and Comp Biol(BICoB); 2013.
53. Bhadra P, Yan J, Li J, Fong S, Siu SWI. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep. 2018;8:1697.
54. Juretić D, Vukicević D, Ilić N, Antcheva N, Tossi A. Computational design of highly selective antimicrobial peptides. J Chem Inf Model. 2009;49:2873–82.
55. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 2005;102:6395–400.

56. Boone K, Camarda K, Spencer P, Tamerler C. Antimicrobial peptide similarity and classification through rough set theory using physicochemical boundaries. BMC Bioinformatics. 2018;19:1–10.
57. Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers. 1988;27:451–77.
58. Xia J-F, Han K, Huang D-S. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. Protein Pept Lett. 2010;17:137–45.
59. Kleandrova VV, Ruso JM. Speck-Planche A, Dias Soeiro Cordeiro MN. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. ACS Comb Sci. 2016;18:490–8.
60. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol. 1968;21:170–201.
61. Fernández L, Caballero J, Abreu JI, Fernández M. Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. Proteins. 2007;67:834–52.
62. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89:10915–9.
63. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci. 1998;23:444–7.
64. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
65. Maetschke S, Towsey M, Bodén M. Blomap: an encoding of amino acids which improves signal peptide cleavage site prediction. In: Proceedings of the 3rd Asia-Pacific Bioinformatics Conference; 2005. p. 141–50.
66. Huang L, Dai Y. A support vector machine approach for prediction of t cell epitopes. In: Proceedings of the 3rd Asia-Pacific Bioinformatics Conference; 2005. p. 319–28.
67. Karypis G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins. 2006;64:575–86.
68. Kumar M, Michael Gromiha M, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins: Struct Funct Bioinf. 2008;71:189–94.
69. Verma R, Varshney GC, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids. 2009;39:101–10.
70. Nanni L, Lumini A, Gupta D, Garg A. Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. IEEE/ACM Trans Comput Biol Bioinform. 2012;9:467–75.
71. Xu R, Zhou J, Wang H, He Y, Wang X, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst Biol. 2015;9(Suppl 1):S10.
72. Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarctions using fully convolutional neural networks. arXiv.org; 2018.
73. Nagarajan V, Kaushik N, Murali B, Zhang C, Lakhera S, Elasri MO, et al. A Fourier transformation based method to mine peptide space for antimicrobial activity. BMC Bioinformatics. 2006;7(Suppl 2):S2.
74. Yin C, Yau SS-T. A coevolution analysis for identifying protein-protein interactions by Fourier transform. PLoS One. 2017; 12:e0174862.
75. Baker D. Protein Structure Prediction and Structural Genomics. Science. 2001;294:93–6.
76. Zasloff M. Antimicrobial peptides of multicellular organisms. Nature. 2002;415:389–95.
77. Löchel HF, Riemenschneider M, Frishman D, Heider D. SCOTCH: subtype A coreceptor tropism classification in HIV-1. Bioinformatics. 2018;34:2575–80.
78. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57:4977–5010.
79. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018;23:1538–46.
80. Taboureau O, Olsen OH, Nielsen JD, Raventos D, Mygind PH, Kristensen H-H. Design of novispirin antimicrobial peptides by quantitative structure-activity relationship. Chem Biol Drug Des. 2006;68:48–57.
81. Bhonsle JB, Venugopal D, Huddler DP, Magill AJ, Hicks RP. Application of 3D-QSAR for Identification of Descriptors Defining Bioactivity of Antimicrobial Peptides. J Med Chem. 2007;50:6545–53.
82. Jenssen H, Lejon T, Hilpert K, Fjell CD, Cherkasov A, Hancock REW. Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward *P. aeruginosa*. Chem Biol Drug Des. 2007;70:134–42.
83. Jenssen H, Fjell CD, Cherkasov A, Hancock REW. QSAR modeling and computer-aided design of antimicrobial peptides. J Pept Sci. 2008;14:110–4.
84. Shu M, Yu R, Zhang Y, Wang J, Yang L, Wang L, et al. Predicting the activity of antimicrobial peptides with amino acid topological information. Med Chem. 2013;9:32–44.
85. Schneider P, Müller AT, Gabernet G, Button AL, Posselt G, Wessler S, et al. Hybrid Network Model for "Deep Learning" of Chemical Data: Application to Antimicrobial Peptides. Mol Inform. 2017;36:1–7.
86. Cui J, Liu Q, Puett D, Xu Y. Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics. 2008;24:2370–5.
87. Chang KY, Lin T-P, Shih L-Y, Wang C-K. Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields. PLoS One. 2015;10:e0119490.
88. Torrent M, Di Tommaso P, Pulido D, Nogués MV, Notredame C, Boix E, et al. AMPA: an automated web server for prediction of protein antimicrobial regions. Bioinformatics. 2012;28:130–1.
89. Dybowski JN, Heider D, Hoffmann D. Prediction of co-receptor usage of HIV-1 from genotype. PLoS Comput Biol. 2010; 6:e1000743.
90. Heider D, Dybowski JN, Wilms C, Hoffmann D. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Min. 2014;7:14.
91. Bozek K, Lengauer T, Sierra S, Kaiser R, Domingues FS. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. PLoS Comput Biol. 2013;9:e1002977.

92. Sander O, Sing T, Sommer I, Low AJ, Cheung PK, Harrigan PR, et al. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLoS Comput Biol. 2007;3:e58.
93. Yu X, Weber I, Harrison R. Sparse Representation for HIV-1 Protease Drug Resistance Prediction. In: Proceedings of the 2013 SIAM International Conference on Data Mining; 2013. p. 342–9.
94. Bose P, Yu X, Harrison RW. Encoding protein structure with functions on graphs. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW); 2011. p. 338–44.
95. Weber IT, Harrison RW. Decoding HIV resistance: from genotype to therapy. Future Med Chem. 2017;9:1529–38.
96. Cardoso MH, Oshiro KGN, Rezende SB, Cândido ES, Franco OL. The Structure/Function Relationship in Antimicrobial Peptides: What Can we Obtain From Structural Data? Adv Protein Chem Struct Biol. 2018;112:359–84.
97. Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou K-C, et al. PREvalL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol. 2018; 443:125–37.
98. Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990;18:2163–70.
99. Basu S, Pan A, Dutta C, Das J. Chaos game representation of proteins. J Mol Graph Model. 1997;15:279–89.
100. He P-A, Xu S, Dai Q, Yao Y. A generalization of CGR representation for analyzing and comparing protein sequences. Int J Quantum Chem. 2016;116:476–82.
101. Ge L, Liu J, Zhang Y, Dehmer M. Identifying anticancer peptides by using a generalized chaos game representation. J Math Biol. 2018:1–23.
102. Jia J, Li X, Qiu W, Xiao X, Chou K-C. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. J Theor Biol. 2019;460:195–203.
103. Loose C, Jensen K, Rigoutsos I, Stephanopoulos G. A linguistic model for the rational design of antimicrobial peptides. Nature. 2006;443:867–9.
104. Maccari G, Di Luca M, Nifosí R, Cardarelli F, Signore G, Boccardi C, et al. Antimicrobial peptides design by evolutionary multiobjective optimization. PLoS Comput Biol. 2013;9:e1003212.
105. Joseph S, Karnik S, Nilawe P, Jayaraman VK, Idicula-Thomas S. ClassAMP: a prediction tool for classification of antimicrobial peptides. IEEE/ACM Trans Comput Biol Bioinform. 2012;9:1535–8.
106. Mooney C, Haslam NJ, Pollastri G, Shields DC. Towards the improved discovery and design of functional peptides: common features of diverse classes permit generalized prediction of bioactivity. PLoS One. 2012;7:e45012.
107. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers. 2005;80:775–86.
108. Polanco C, Samaniego JL. Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models. Acta Biochim Pol. 2009;56:167–76.
109. Randou EG, Veltri D, Shehu A. Binary Response Models for Recognition of Antimicrobial Peptides. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13; 2007. p. 76–85.
110. Barrett R, Jiang S, White AD. Classifying antimicrobial and multifunctional peptides with Bayesian network models. Pept Sci. 2018;110:e24079.
111. Kernytsky A, Rost B. Using genetic algorithms to select most predictive protein features. Proteins. 2009;75:75–88.
112. Leslie C, Eskin E, Noble WS. The spectrum kernel: a string kernel for svm protein classification. Biocomputing. 2002;2001: 564–75.
113. Fjell CD, Hiss JA, Hancock REW, Schneider G. Designing antimicrobial peptides: form follows function. Nat Rev Drug Discov. 2011;11:37–51.
114. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics. 2004;20:467–76.
115. Swamidass SJ, Chen J, Bruand J, Phung P, Ralaivola L, Baldi P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics. 2005;21(Suppl 1):i359–68.
116. Lewis DP, Jebara T, Noble WS. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics. 2006;22:2753–60.
117. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 2002;11:2606–21.
118. Boisvert S, Marchand M, Laviolette F, Corbeil J. HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels. Retrovirology. 2008;5:110.
119. El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. J Mol Recognit. 2008; 21:243–55.
120. Toussaint NC, Widmer C, Kohlbacher O, Rätsch G. Exploiting physico-chemical properties in string kernels. BMC Bioinformatics. 2010;11(Suppl 8):S7.
121. Giguère S, Marchand M, Laviolette F, Drouin A, Corbeil J. Learning a peptide-protein binding affinity predictor with kernel ridge regression. BMC Bioinformatics. 2013;14:82.
122. Giguère S, Laviolette F, Marchand M, Tremblay D, Moineau S, Liang X, et al. Machine learning assisted design of highly active peptides for drug discovery. PLoS Comput Biol. 2015;11:e1004074.
123. Telenti A, Lippert C, Chang P-C, DePristo M. Deep learning of genomic variation and regulatory network data. Hum Mol Genet. 2018;27:R63–71.
124. Asgari E, Mofrad MRK. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One. 2015;10:e0141287.
125. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33:3036–42.
126. Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ. 2018;6:e4750.
127. Taju SW, Nguyen T-T-D, Le N-Q-K, Kusuma RMI, Ou Y-Y. DeepEfflux: a 2D convolutional neural network model for identifying families of efflux proteins in transporters. Bioinformatics. 2018;34:3111–7.
128. Sun J, Deng Z, Yan A. Bacterial multidrug efflux pumps: Mechanisms, physiology and pharmacological exploitations. Biochem Biophys Res Commun. 2014;453:254–67.

129. Seo S, Oh M, Park Y, Kim S. DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics. 2018;34:i254–62.
130. Zheng W, Yang L, Genco RJ, Wactawski-Wende J, Buck M, Sun Y. SENSE: Siamese neural network for sequence embedding and alignment-free comparison. Bioinformatics. 2018:1–9.
131. Wang Y-B, You Z-H, Li X, Jiang T-H, Chen X, Zhou X, et al. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol Biosyst. 2017;13:1336–44.
132. Piotto SP, Sessa L, Concilio S, Iannelli P. YADAMP: yet another database of antimicrobial peptides. Int J Antimicrob Agents. 2012;39:346–51.
133. Waghu FH, Gopi L, Barai RS, Ramteke P, Nizami B, Idicula-Thomas S. CAMP: Collection of sequences and structures of antimicrobial peptides. Nucleic Acids Res. 2014;42:D1154–8.
134. Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Res. 2016;44:D1094–7.
135. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2016;44:D1087–93.
136. Fan L, Sun J, Zhou M, Zhou J, Lao X, Zheng H, et al. DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep. 2016;6:24482.
137. Porto WF, Pires AS, Franco OL. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. Biotechnol Adv. 2017;35:337–49.
138. Gabere MN, Noble WS. Empirical comparison of web-based antimicrobial peptide prediction tools. Bioinformatics. 2017; 33:1921–9.
139. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013;29:960–2.
140. Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;31:1857–9.
141. Ofer D, Linial M. ProFET: Feature engineering captures high-level protein functions. Bioinformatics. 2015;31:3429–36.
142. Müller AT, Gabernet G, Hiss JA, Schneider G. modlAMP: Python for antimicrobial peptides. Bioinformatics. 2017;33:2753–5.
143. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017;33:2756–8.
144. Dong J, Yao Z-J, Zhang L, Luo F, Lin Q, Lu A-P, et al. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Cheminform. 2018;10:16.
145. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34:2499–502.
146. Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. Hoboken: Wiley; 2004.

6.3 A Large-scale Comparative Study on Peptide Encodings for Biomedical Classification

The selection of effective encodings for categorical data is a particular challenge in the context of ML. Since most ML algorithms require numerical and fixed-length input, categories must be translated to continuous values[32]. Common techniques are one-hot, dummy, and ordinal descriptors[189,51]. Regarding a group with n different classes, the one-hot, or binary, encoding maps each member to a binary vector of length n, where one bit is set, respectively. The dummy encoding also represents each class as a binary vector, using $n - 1$ variables. The n-th type is encoded as the zero-vector. In contrast, the ordinal encoder maps the classes to integers from 1 to n. Additionally, more complex encodings have been developed over time, for instance, Helmert n or target encoding[51,189].

These encoding schemes reflect the data as numerical values. However, for biological sequences, the representation is even more complex since information is encoded beyond individual amino acids[251]. For instance, interactions of amino acids, potentially non-adjacent in the primary sequence, stabilize the secondary and tertiary structure[251]. Consequently, amino acid sequences hamper encoding since biological information, hence, the sequence's function, must be retained. In this light, many peptide encodings have been developed that describe more complex associations[116].

An example for translating physicochemical properties of individual amino acids is the amino acid index[98]. This encoding is based on a database providing experimentally derived physical and chemical effects of individual amino acids, such as the hydrophobicity[119,130]. Other encodings condense multiple amino acids to one figure, for instance, by numerically depicting auto-correlation between recurring residues[142]. Moreover, researchers applied encodings for various biomedical ML tasks, including AMPs classification or the prediction of cell-penetrating peptides.

The tertiary structure is a crucial for efficient StBEs. For instance, Löchel *et al.* (2019) encoded the V3 loop sequence of the human immunodeficiency virus (HIV) using the electrostatic hull[146]. However, various studies from multiple biomedical areas applied diverse encoding types[7]. The great variety raises the question of whether specific encodings are superior for particular domains. Consequently, we collected 48 encoding groups, partly parameterized, and 50 datasets from different biomedical disciplines, ultimately resulting in 397,700 encoded datasets.

We demonstrated that the biomedical application and performance are unrelated, and no encoding is superior in a specific domain. Nevertheless, some encodings are more frequently top-ranked. Furthermore, the results revealed a high correlation between parameterized descriptors, specifically concerning adjacent configurations. We also observed that SeBEs are in general superior to StBEs.

The similarity of the classifier output, comprising diversity[128], also reflects the inferiority of StBEs. We observed that the predicted classes between the two categories only show sparse consensus. Moreover, encodings inferred from the same type, hence, with varying parameters, indicated higher similarity. To comply with findable, accessible, interchangeable, and reproducible (FAIR) standards[247], the experiments are implemented as an end-to-end pipeline using Snakemake[126]. The datasets and results are hosted in a public repository, and the results are illustrated in a web-based platform.

We developed the PEPTIDE REACToR, a workflow to evaluate the performance of encoded datasets from various biomedical domains. Although we showed that none of the encodings work particularly well in a specific field, the results enable researchers to select initial encodings. More precisely, the encoding recommendation pinpoints significant steps to select encodings for a biomedical classification task at hand. Moreover, the inferiority of StBEs is insofar surprising since a peptide's structure mainly defines its function[131]. However, StBEs render acceptable performance, which could be due to sequences with known structures.

Additional research is also required to address the effect of hyper-parameter optimization, including model selection. Albeit the workflow reduces the several thousand initially encoded datasets to a few hundred, manual encoding selection is still required. However, the reduced number of datasets paves the way for more sophisticated approaches, for instance, automated ML. In this light, Feurer *et al.* (2015) demonstrated the good performance and the ease of use of automated ML[69]. Research about unsupervised encoding selection is essential to completely omit manual selection for automatic ML in biomedical classification.

Published online 22 May 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab039

A large-scale comparative study on peptide encodings for biomedical classification

Sebastian Spänig, Siba Mohsen, Georges Hattab, Anne-Christin Hauschild and Dominik Heider*

Data Science in Biomedicine, Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg, Germany

Received January 15, 2021; Revised April 13, 2021; Editorial Decision April 19, 2021; Accepted April 26, 2021

ABSTRACT

Owing to the great variety of distinct peptide encodings, working on a biomedical classification task at hand is challenging. Researchers have to determine encodings capable to represent underlying patterns as numerical input for the subsequent machine learning. A general guideline is lacking in the literature, thus, we present here the first large-scale comprehensive study to investigate the performance of a wide range of encodings on multiple datasets from different biomedical domains. For the sake of completeness, we added additional sequence- and structure-based encodings. In particular, we collected 50 biomedical datasets and defined a fixed parameter space for 48 encoding groups, leading to a total of 397 700 encoded datasets. Our results demonstrate that none of the encodings are superior for all biomedical domains. Nevertheless, some encodings often outperform others, thus reducing the initial encoding selection substantially. Our work offers researchers to objectively compare novel encodings to the state of the art. Our findings pave the way for a more sophisticated encoding optimization, for example, as part of automated machine learning pipelines. The work presented here is implemented as a large-scale, end-to-end workflow designed for easy reproducibility and extensibility. All standardized datasets and results are available for download to comply with FAIR standards.

INTRODUCTION

With the increasing popularity of machine learning methods, scientists began to use them for a wide range of biomedical applications. A particular application is the prediction of amino acid (AA) sequence properties, for example, a peptide's antimicrobial efficiency (1), cell-penetrating (2) and cell-entry (3) properties, or the classification of T-cell epitopes (4). However, the mode of action of a peptide sequence depends on a variety of biochemical factors, which cannot be reflected by the order of the AAs alone (1). Moreover, many machine learning models require a numerical input with a fixed dimension (5). To this end, many descriptors, i.e. sequence-based encodings (SeBEs) have been developed, aiming to compute adequate numerical representations of the primary structure. In short, SeBEs are algorithms mapping the AAs to numerical values, but also incorporate interactions of non-adjacent residues, for instance, by autocorrelation techniques (6,7). SeBEs have been successfully employed in numerous studies, for example, for the applications mentioned above, but also to predict antiviral (8) or anticancer peptides (9). In addition, tools such as *iFeature* (6) or *BioSeq-Analysis2.0* (10), which allow easy access to SeBEs, have paved the way for a wide range of biomedical applications.

However, the function of a peptide is not only defined by its primary structure, but biological meaning will be also encoded in higher dimensions, i.e. the peptide's secondary or tertiary structure. Consequently, structure-based encodings (StBEs) augment SeBEs to maximize the information gain. StBEs can be divided into two further groups: encodings derived from the secondary structure and those derived from the tertiary structure. The former includes encodings describing, for example, the α-helix composition (6), based on an *ab initio* secondary structure prediction (11). For the latter, Bose *et al.* (2011) utilized the Delaunay triangulation to encode protein structures as numerical feature vectors (12). The aim of the study was to predict protein structure properties and the results showed, that this StBE is capable to preserve tertiary structure information for machine learning purposes (12). Furthermore, Löchel *et al.* (2018) demonstrated, that using the electrostatic hull of V3-loop of the gp120 protein, substantially improved the prediction of co-receptor tropism of the human immunodeficiency virus 1 (13). A comprehensive introduction to encodings, specifically dealing with the prediction of antimicrobial peptides, can be found in our recent review (7).

*To whom correspondence should be addressed. Tel: +49 6421 2821579; Email: dominik.heider@uni-marburg.de

Nevertheless, several major challenges remain. First of all, there is no guideline or clear recommendation which encodings work well for specific biomedical applications, facing researchers with the effort of matching the right encoding for the task. Second, even if one or more encodings have been determined, researchers are very likely challenged with parameterized ones, further increasing the hyperparameter search space and thus, actually aggravating the encoding exploration. Third, many studies confirmed that combining different encodings to ensemble classifiers, effects the prediction performance positively (14,15). Specifically, Dybowski *et al.* (2011) employed stacked generalization on the predictions of SeBE- and StBE-based classifiers and thus, improved the resistance prediction to Bevirimat, an antiretroviral drug class (16). Consequently, applying ensemble learning techniques enlarges the hyperparameter search space further and a structured exploration becomes more and more difficult.

For this reason, we present here, to the best of our knowledge, the first large-scale comprehensive study on state of the art peptide encodings on a wide range of datasets from a wide range of biomedical domains. Our study closes the gap between the availability of a great variety of encodings and the important question whether one of them is best suited for a specific domain application or task. This study builds upon our recent review on peptide encodings (7), which allows us to add additional, literature-known sequence- and structure-based encodings. The goal of the study is to provide researchers, faced with a biomedical classification task at hand, general guidelines, which encodings are likely to be superior on a certain biomedical classification task. Thus, we investigated the two major encoding types, namely SeBEs and StEBs, in total leading to 48 encoding groups. Moreover, we collected 50 datasets from multiple domains, including antimicrobial, -viral and -cancer as well as cell-penetrating peptides as already mentioned above, but also from further fields, such as HIV drug resistance prediction. By further taking the parameterization of some of the encoding groups into account, we generated altogether hundreds of thousands of encoded datasets.

To meet this unique challenge we have developed the PEPTIDE REACToR, a platform bundling manifold analyses to examine characteristics of the encoded datasets (see Figure 1). The workflow is designed for high parallelization, enabling an efficient evaluation, even in the case of additional encodings and datasets in the future. Surprisingly, our results point out, that no particular encoding can be recommended in general. However, there are encodings that show increased performance across multiple datasets, hence, biomedical domains. Contrary, our method reveals many inferior encoding groups, questioning the necessity of computing them at all. Thus, our findings pave the way for automated machine learning approaches, in that the hyperparameter space is drastically reduced and relevant techniques become computationally feasible. According to the FAIR data principles (findability, accessibility, interoperability and reusability) (17), the results can be interactively accessed at https://peptidereactor.mathematik.uni-marburg.de/ and all datasets can be downloaded at a central location. The source code as well as the datasets are available at https://github.com/spaenigs/peptidereactor.

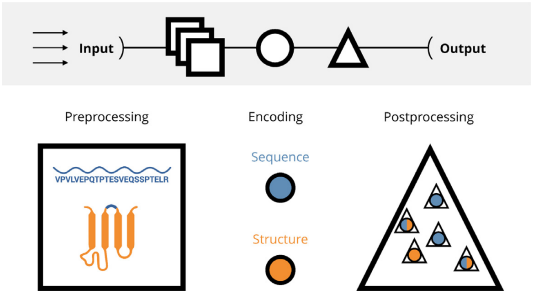


Figure 1. The general principle of the PEPTIDE REACToR. The emphasis is put on a high-throughput processing of an arbitrary amount of input datasets (arrows), followed by the preprocessing, encoding, and postprocessing, generating the final output (top). The preprocessing includes sanitizing of the input sequences, the filtering and the tertiary structure approximation (squares). Afterwards, the sequences as well as the accompanied structures are used for the encoding (circles). The postprocessing involves the machine learning and the actual benchmarking including the visual preparation of the analyses (triangles).

MATERIALS AND METHODS

We collected 50 datasets from a wide range of biomedical applications. Furthermore, building upon our recent encoding review (7), we aggregated in total 48 encodings and developed a high-throughput approach facilitating a parallelized encoding and the subsequent comparison of the encoded datasets. Every task is part of a large-scale, end-to-end workflow and will be executed automatically. An overview of the workflow can be found in Figure 1. We used Python v3.7.4 (https://www.python.org/) and R v3.5.2 (https://www.r-project.org/) throughout the analysis. The pipeline itself as well as the algorithms in particular have been implemented as a modular Snakemake v5.19.0 (18) pipeline. Moreover, we used Scikit-learn v0.23.1 for the machine learning algorithms and validation metrics (19).

The following sections describe the applied methodology by keeping the actual order of the workflow. Thus, the dataset collection will be presented at first. The subsequent section introduces the tertiary structure approximation, since it is crucial before the actual encodings and their properties are presented. Some of the encodings are parameterized, thus, leading to thousands of encoded datasets. Therefore, the next section sheds light on the algorithmic details of the encoded datasets filtering. Finally, the actual benchmark methodology will be presented and the method section concludes with the result visualization description. Refer to Figure 1 for a visual summary.

Datasets collection

We collected 50 different datasets comprising peptides and small proteins from various biomedical domains. These include immunomodulatory and cell-penetrating peptides, but also peptides specifically targeting cancer, fungi, microbes, tuberculosis and viruses. Moreover, we added datasets from HIV research specifically covering resistance prediction against different drug classes and protease cleavage site prediction. A further application refers to the detec-

tion of neuropeptides as well as a- and b-cell epitopes. More attributes, for example, origin, size, etc. can be found in the Supplementary Table S1. Detailed dataset descriptions are specified in Supplementary Table S3.

The datasets were composed for manifold reasons. They reflect a broad field of action, including infectious diseases, for example, HIV, antimicrobial resistance of multi-drug resistant bacteria, and others, to elevate the significance of the results. If possible, we used several datasets per domain to also reflect the sequence diversity. In order to cope with the high-dimensionality of the present study, we limited the benchmark to two-class problems.

Moreover, the datasets have been applied largely as they are, in order to stay as close as possible to the original usage. That is, the class ratio of the datasets at hand ranges from well balanced (e.g. *ace_vaxinpad*) to very imbalanced ones (*hiv_v3*) (see Figure 2). This affects also the size of the datasets, which ranges from small ones (e.g. *amp_gonzales*) to relatively large datasets (e.g. *amp_iamp2l*). Refer to Supplementary Tables S1 and S3 for more details. Too large datasets have been excluded from the study or, if present, the validation dataset were used instead.

All in all, the datasets are composed of 53 041 sequences ranging from 3 to 255 amino acids. The mean sequence length is 55.04 (±67.58) with a median length of 26 amino acids. Refer to Supplementary Table S2 for a comprehensive descriptive evaluation on the datasets used in this study. In particular, let D_i be the i-th dataset from a biomedical application, i.e. composed of a set of n amino acid sequences s of length k, denoted as

$$D_i = \{s_1, s_2, \ldots, s_{n-1}, s_n\} \tag{1}$$

and

$$s_i = \{a_1, a_2, \ldots, a_{k-1}, a_k\} \tag{2}$$

with a_i being one of the 20 natural amino acids.

Tertiary structure approximation

Two categories of encodings have been investigated: sequence- and structure-based encodings (SeBEs and StBEs, respectively). While the former are derived from the primary structure, i.e. the amino acid sequence, the computation of the latter bears on the secondary, if not the tertiary structure of a peptide or protein, respectively. Even though algorithms exist for the prediction of secondary structure properties, for example, SPINE X (20), or the prediction of the tertiary structure, for instance, RaptorX (21), they are often computationally expensive, above all, if one aims to predict hundreds of structures simultaneously.

However, for a large-scale approach, this is not practical, thus, we developed in addition an algorithm, which approximates the tertiary structure for later usage by StBEs. While PSI-BLAST (22) is capable of finding more distant relative sequences, it often suffers from a long run time for long sequences. Thus, we applied BLAST (23) v2.9.0 instead. In order to set up a database, we downloaded all available structures (as of May 2020) from the Protein Data Bank (24) (PDB, http://www.rcsb.org/), extracted all sequences into a FASTA file using Biopython v1.7.4 (25,26) and used it as input for the *makeblastdb* command. By doing so, we ensure, that the database contains only sequences with a known structure.

For a sequence s_i, the structure approximation works as follows: first, an initial BLAST run tries to find the query sequence within a PDB entry. For the best match, i.e. the match with the lowest e-value, the respective PDB file will be fetched. The algorithm clips the matching part from the structure and returns the i-th tertiary structure approximation for a query sequence s_i. Any s_i, for which no structure has been found, is omitted in the later encoding step.

Encodings

Spänig and Heider (2019) conducted an extensive literature search and collected a wide range of SeBEs and StBEs (7). We employed the Python package *iFeature*, which already provides many encodings (6). Moreover, we also added the frequency matrix chaos game representation (FCGR), an adoption of the original CGR, recently developed by our group (27). However, as part of this study, we contribute the implementation of 10 additional encodings to the scientific community, i.e. encodings, which have been used successfully in the literature, but where an actual implementation is lacking. For a comprehensive list of all encodings, refer to Supplementary Table S4. Supplementary Note S1 provides the algorithmic details on the additional encodings, for the remainders, refer to (6) or (7). In addition, we employed MUSCLE v3.8.1551 (28) in case an encoding, for instance, the binary encoding, requires a multiple sequence alignment beforehand. In particular, an encoding is a function f, mapping an amino acid sequence s_i to an numerical vector $\hat{s}_i$:

$$f : s_i \rightarrow \hat{s}_i, \hat{s}_i \in \mathbb{Q}^{\mathbb{N}} \tag{3}$$

Filtering

Since some of the encodings are parameterized, thus, leading in total to thousands of encoded datasets, an important part of the pipeline is the filtering of the d encoded datasets $\{\hat{D}_1, \ldots, \hat{D}_d\}$, hence to reduce the extent of d before the actual benchmark. For the purpose of a benchmark, we covered the input parameter space for all encodings as extensive as possible, thus we generated in total d encoded datasets:

$$d = \sum_{i}^{48} |\overrightarrow{x_1(i)} \times \cdots \times \overrightarrow{x_n(i)}| \tag{4}$$

Whereby $\times$ denotes the Cartesian product and n refers to the n-th parameter set for the i-th encoding group. Specifically, the amino acid index-based encodings are highly related owing to an intrinsic correlation of certain amino acid indices. Moreover, parameterized encodings take the window length $\vec{w}$ of size k for autocorrelation-based encodings, or correlation types $\vec{c}$ of size l, tuple sizes $\vec{t}$ of size m, and gap length parameters $\vec{g}$ of size n for the pseudo K-tuple reduced amino acids composition (PseKRAAC) encoding leading to $|\vec{w}| + |\vec{c} \times \vec{t} \times \vec{g}|$ encoded datasets for these encodings groups alone. Refer to Supplementary Table S4 for the comprehensive list on parameterized encodings and which parameter space have been covered in par-

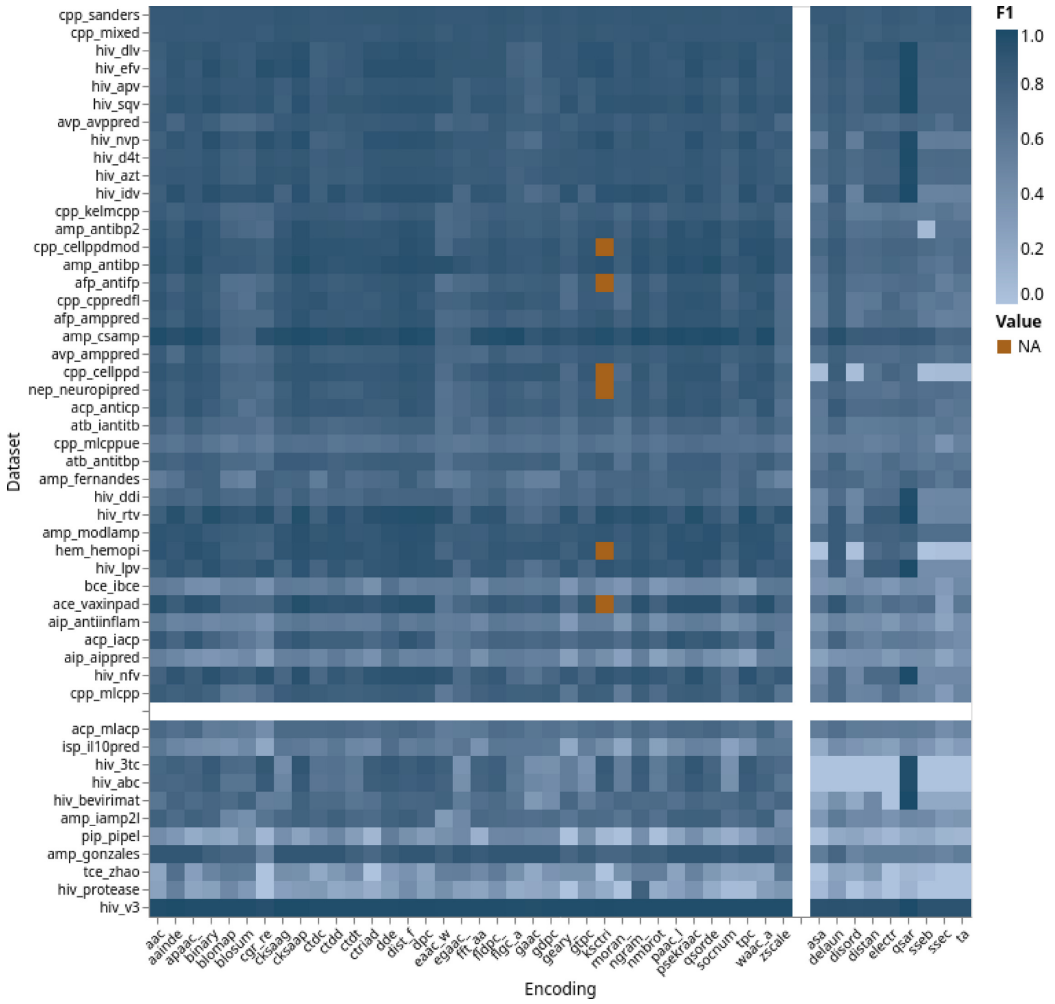


Figure 2. Encoding groups performance, sorted by class imbalance and encoding type. Color coding corresponds to the maximum F1-score of the bootstrapped medians for a group. The *x*-axis is organized by sequence- and structure-based encodings. The *y*-axis is sorted by class imbalance (cut-off 0.35). Groups are separated by white bars. An interactive version of this plot can be found at https://peptidereactor.mathematik.uni-marburg.de/.

ticular. Supplementary Note S4 provides a detailed description of the filter algorithm for the amino acid index as well as PseKRAAC encodings.

Benchmark

The essential part of this project is the high-throughput evaluation of all encodings across multiple biomedical datasets. To this end, several advanced processing as well as analysis steps are conducted, which are introduced more detailed hereinafter.

Model training. In order to standardize the analysis, we used the Random Forest classifier (RFC) (29) with default parameter settings as the default machine learning model. RFCs already perform good without hyper-parameter optimization, which is contrary to, for example, Support Vector Machines, which achieve far greater performance with optimized hyper-parameters compared to the defaults (30). That is, RFCs are more stable, allowing us to neglect the influence of hyper-parameter optimization on the encoding performance. Moreover, we chose this classifier since it exhibits a variety of advantages compared to other prediction models. It internally picks the most predictive features out of a set of multiple decision trees, that is, it has a built-in feature selection method. Moreover, the final prediction is based on the trees built from the selected features; hence, it is also an ensemble algorithm. In addition, the feature im-

portance can be calculated and the RFC is also capable of reducing overfitting (29).

Cross-validation. In order to generalize the model performance, we applied a repeated, stratified k-fold cross-validation (CV). In particular, for each validation round, an encoded dataset $\hat{D}_i$ is splitted into $k = 5$ folds and each CV is repeated 10 times. For each fold, the intermediate results, i.e. the vectors of the predicted classes $\vec{t}$, the probabilities $\vec{p}$ and the actual classes $\vec{y}$ are stored in the matrices $\boldsymbol{R_p}$ and $\boldsymbol{R_t}$, whereby $p \in \{\vec{t}, \vec{p}\}$ and $t \in \{\vec{y}\}$. $\boldsymbol{R_p}$, analogous to $\boldsymbol{R_t}$, is denoted as shown in Equation 5, with $p_{fold_k, pred_n}$ being the n-th predicted probability or class of the k-th fold. In addition, the number of rows in both matrices corresponds to the repetitions as well as folds of the CV, hence 50 in the present case.

$$\boldsymbol{R_p} = \begin{bmatrix} p_{fold_1, pred_1} & \cdots & p_{fold_1, pred_n} \\ \vdots & \ddots & \vdots \\ p_{fold_k, pred_1} & \cdots & p_{fold_k, pred_n} \end{bmatrix} \tag{5}$$

Note, that the overall CV is conducted two times: one time for each $\hat{D}_i$ without any restrictions and a second time for two groups of encodings, for example, for SeBEs and StBEs. As mentioned above, it might be the case, that a tertiary structure approximation failed. Consequently, a dataset $\hat{D}_i$, based on a StBE, might lack certain sequences, but the two-group CV needs to ensure equal records in both $\hat{D}_i \in$ SeBEs and $\hat{D}_k \in$ StBEs. Thus, in the case of a two-group CV, we compute the intersection of the record labels and remove the additional rows from $\hat{D}_i$ prior to the actual CV.

Performance metrics. In order to evaluate the performance of the encodings with a single measure, we calculated the following metrics: F1-score, Matthews Correlation Coefficient (MCC), Precision, Recall (Sensitivity) and Specificity. Each of these measures has particular properties, allowing them to highlight the advantages or disadvantages of specific encodings concerning the task. Refer to Supplementary Note S2 for the respective formulas. All metrics are computed on the k-th split of the k-th row from $\boldsymbol{R_p}$ and $\boldsymbol{R_t}$.

Similarity. The similarity of classifiers, that is, the similarity of the predictions of unknown test examples from the respective classifiers, trained with the encoded datasets $\hat{D}_i$ and $\hat{D}_j$, could stress specific strengths and weaknesses of an encoding. To this end, we implemented two similarity measurements, namely the Phi coefficient (31) (see Supplementary Note S2) and the disagreement measure D (31,32), with the respective output of the i-th classifier ${o^i}_k$ and of the j-th classifier ${o^j}_k$, denoted as:

$$D_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \left| {o^i}_k - {o^j}_k \right| \tag{6}$$

Analogous to the performance metrics, we computed the particular similarity for the k-th CV split on the k-th row of the i-th and the j-th classifier outputs ${\boldsymbol{R^i}}_{\boldsymbol{p}}$ and ${\boldsymbol{R^j}}_{\boldsymbol{p}}$, respectively. Finally, the overall similarity is the average across all splits. The two-group CV is the basis for the similarity measures since the output of the classifiers i and j, need to be traceable to the same sequences.

Critical difference. There are several statistical tests for evaluating machine learning models trained on multiple datasets. Depending on the classification task at hand, Santafé *et al.* (2015) provided an overview of the recommended procedure (33). In the present case, we considered the models trained on many encoded datasets as the statistical comparison of several classifiers trained on several datasets. In particular, we assume, that using the RFC models trained on k encoded datasets instead of k algorithms fulfills the criteria for the Friedman statistic ${\chi^2}_F$, meaning the models are related, i.e. paired, and each fold is independent of each other:

$${\chi^2}_F = \frac{12N}{k(k+1)} \left[\sum_j {R^2}_j - \frac{k(k+1)^2}{4} \right] \tag{7}$$

with the the Iman and Davenport correction:

$$F_F = \frac{(N-1){\chi^2}_F}{N(k-1) - {\chi^2}_F} \tag{8}$$

in order to verify, whether one of the models outperforms another. That is, to reject the null hypothesis, which states, that there is no difference between the classifiers. In particular, the Friedman test compares the ranks ${r^j}_i$ of the j-th model validated on the i-th fold. The average rank is denoted as $R_j = \frac{1}{N} \sum_i {r^j}_i$ calculated on N folds and k trained classifiers using $k - 1$ degrees of freedom. Moreover, F_F is F-distributed with $k - 1$ and $(k - 1)$ $(N - 1)$ degrees of freedom (34).

The alternative hypothesis states, that there is a statistically significant difference across the models. In the case of acceptance, the post-hoc analysis using the Nemenyi test unveils the significantly different models. Hence, the critical difference CD, denoted as

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{(6N)}} \tag{9}$$

is computed using the critical value q_α, which is based on the Studentized range statistic with $k(N-1)$ degrees of freedom and a significance level of $\alpha = 0.05$. Two classifiers perform significantly different, if $|R_j - R_j| \geq CD$ (34).

The statistical tests are implemented as part of the R-package *scmamp* v0.2.55 (35).

Encoding correlation. As already pointed out in a previous section, many encoded datasets are either intrinsically correlated, for instance, the AAI-based encodings or derived from the same encoding group, but with slightly different parameters, for example, the window size. Ultimately, we are dealing with high-dimensional, potentially very similar datasets of varying dimensions. In order to measure the degree of correlation, we utilized the adjusted RV-coefficient,

which has been developed for these particular case (36):

$$RV_{adj}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} r^2{}_{adj}(x_i, y_j)}{\sqrt{\sum_{i,j=1}^{p} r^2{}_{adj}(x_i, x_j)\sum_{i,j=1}^{q} r^2{}_{adj}(y_i, y_j)}} \tag{10}$$

with $r^2{}_{adj}(x, y)$ being the adjusted Pearson correlation coefficient (see Supplementary Note S4, Equation S10) between two feature vectors, denoted as:

$$r^2{}_{adj}(x, y) = 1 - \frac{n-1}{n-2}(1 - r^2(x, y)) \tag{11}$$

Moreover, $\boldsymbol{X}$ and $\boldsymbol{Y}$ refer to the encoded datasets $\hat{D}_i$ with p and $\hat{D}_j$ with q features as well as n encoded sequences, respectively. The i-th feature vector from $\boldsymbol{X}$ is denoted as x_i and the j-th feature vector from $\boldsymbol{Y}$ is denoted as y_j. Indahl *et al.* (2015) implemented the adjusted RV-coefficient as part of the *MatrixCorrelation* R-package, which we utilized in version 0.9.4 (37). Since an all versus all calculation is computationally expensive, we determined the RV_{adj} only for the top 50 encodings, based on the F1-score.

Encodings across multiple domains

For the comparison of encodings across multiple datasets, i.e. biomedical domains, we merged the encodings into groups (see Supplementary Table S4) and considered the best-performing encoding (average F1-score of the CV results) as the group representative. Based on these, we ranked the encoding groups across all datasets in order to uncover domain-specific patterns. Moreover, we clustered the datasets and encoding groups by means of the hierarchical clustering using the UPGMA (Unweighted Pair Group Method with Arithmetic mean) method with the euclidean distance (38). We used the implementation provided by the SciPy package (39).

Moreover, for each dataset $\hat{D}_i$, encoded via the amino acid composition encoding, we applied t-SNE with default settings on the sequences of the positive class as well as for both classes. Thus, each s_i^+ and s_i is embedded in the same two-dimensional space, allowing insights specifically regarding the sequence similarity within various biomedical domains and the diversity of the datasets on the sequence level.

Data visualizations

The results are visually depicted and summarized by means of *Altair* v4.1.0 statistical visualization library (40). In particular, we plotted the results for analyzing two kinds of categories (single datasets and summary graphics for all encoded datasets). We followed the 10 simple rules on how to colorize biological data visualizations and applied them in our workflow (41). Note, that in general the choice of the top encodings is made due to the corresponding F1-score. Refer to the Supplementary Note S3 for more details. Finally, the visualizations are aggregated into an interactive report, which can be found at https://peptidereactor.mathematik.uni-marburg.de/.

RESULTS

Workflow

The PEPTIDE REACToR features high-throughput capabilities and a modular design, allowing the processing of an arbitrary amount of encodings and datasets. Novel encodings and additional datasets can be investigated, making it sustainable and future-ready. The benchmark is set up as a high-throughput, large-scale Snakemake (18) workflow. In particular, it is implemented with three important goals in mind: first, efficient use of the available computing power, second, a high parallelization and third, make it findable (F), accessible (A), interchangeable (I) and reusable (R), according to the FAIR data principles (17). However, as the different preprocessing, encoding, as well as benchmark tasks are very diverse and the implementation as one large workflow is cumbersome, the workflow has been designed in a way, that multiple meta nodes, responsible for a specific task or algorithm, are aggregated to a meta workflow. Each meta node is a Snakemake pipeline itself, exposing a defined application programming interface (API), thus, making them interchangeable and reusable. For an easy setup and high reusability, the meta workflow is executed within a Docker v19.03.2 (https://www.docker.com/) environment using Conda v4.8.3 (https://docs.conda.io/en/latest/) for package management.

Performance

In general, the performance of the SeBE groups are superior to the StBE groups (see Figure 2). As an exception, the *qsar* encoding works better on some of the *hiv* datasets. We also observed an increased performance on datasets with relatively balanced class sizes, i.e. the more imbalanced a dataset, the poorer the performance. The *hiv_v3* dataset is an exception. Albeit its striking imbalance, i.e. 200 versus over 1000 sequences for positive and negative class, respectively, the performance of all encodings is good. In addition, we were not able to observe specific encoding groups that are more powerful on certain biomedical classification tasks (see Supplementary Figure S1). The performance does not seem to follow a specific pattern. For instance, all encoding groups showed average performance on the *cpp_mlcppue* dataset, although the classification of the remaining *cpp* datasets was clearly better.

Ranks. Three groups stood out: the *cksaap*, the *distance_frequency* and the *qsar*-based encodings (see Figure 3). Encodings within these groups were more often among the top 3, compared to encodings from the remaining ones. In contrast, the majority of the encoding groups, in particular StBE groups, were rarely among the best.

Clustering. An automated clustering confirmed our findings mentioned above. One can observe two major clusters for the encoding groups and datasets, respectively (see Figure 4). The encoding ones include mainly the SeBE and StBE groups. The former can be further distinguished in three sub-clusters, ranging from (i) the *qsar* to the *ctdd*, (ii) the *ctdt* to the *fldpc_*, as well as (iii) the *egaac_* to the *moran_* encoding groups, although no real pattern

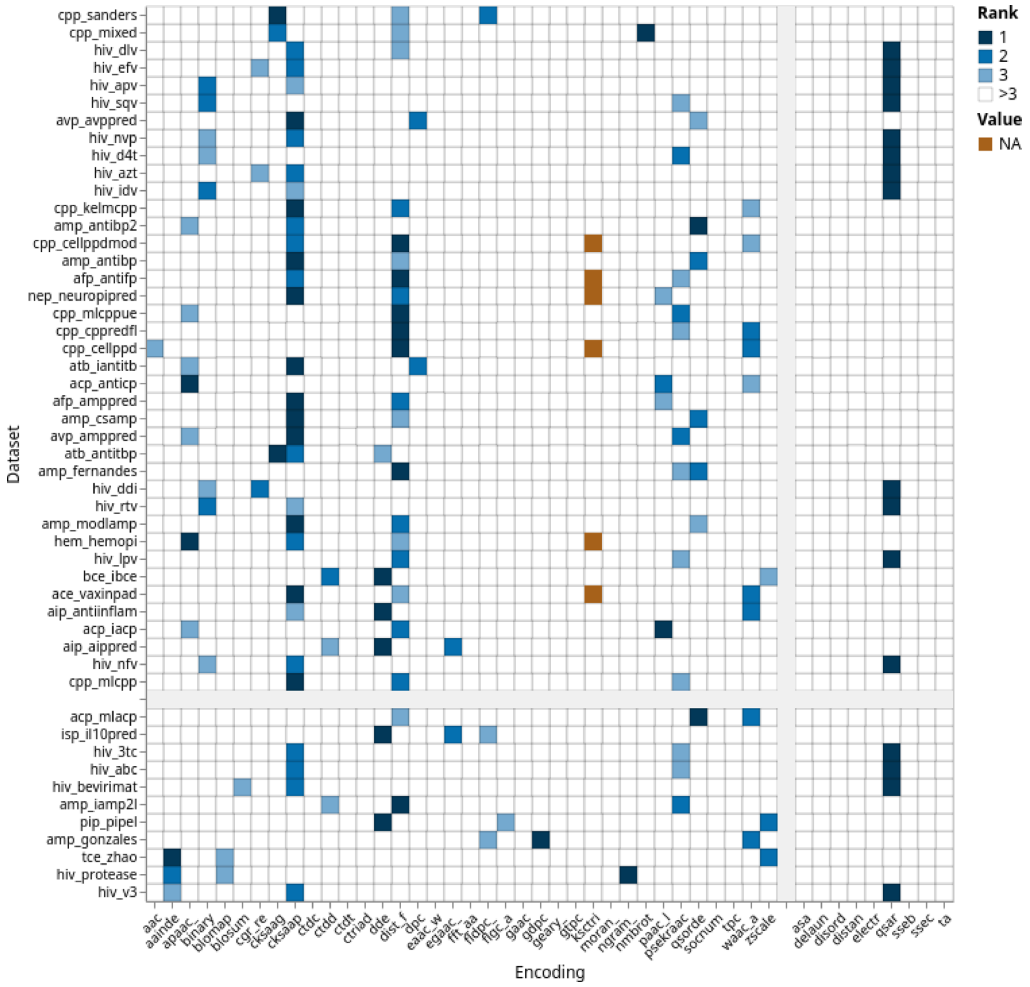


Figure 3. Ranked encoding groups performance, sorted by class imbalance and encoding type. Color coding corresponds to the ranks of encodings across datasets. The x-axis is organized by sequence- and structure-based encodings and the y-axis is sorted by class imbalance (cut-off 0.35). Groups are separated by gray bars. An interactive version of this plot can be found at https://peptidereactor.mathematik.uni-marburg.de/.

emerges within these. An exception are encodings based on the dipeptide composition, namely the *dde*, *dpc* and the *fldpc_* encoding, as these are all within the second cluster. However, the *gdpc* encoding can be found in the first cluster.

Regarding the dataset clusters, the larger of the two can be divided again into three parts, namely (i) from the *hiv_bevirimat* to the *hiv_ddi*, (ii) the *cpp_cellppdmod* to the *atb_antitbp*, and finally (iii) from the *cpp_kelmcpp* to the *hiv_sqv* datasets. Albeit the latter includes predominantly *hiv* related datasets, in general no actual patterns can be observed within the groups. In addition, a two-dimensional embedding of the sequences of the positive class explains some of the dataset clusters (see Supplementary Figure S2). One example is the grouping of the *hiv_nfv*, *hiv_rtv* and *hiv_idv* datasets. The sequences of these datasets form similar, compact clusters.

Median performance. A closer examination of the encodings reveals groups where the range spanned between the worst and the best encoding is noticeable, meaning the best encodings show similar performance compared to the top encodings across all groups and vice versa (see Supplementary Figure S4). In addition, the StBEs show in general worse performance compared to the SeBEs. This can be verified by considering the metrics in detail (see Supplementary Figures S5 and S6). StBEs are mainly located more to the right, i.e. showing a smaller value of the respective metric. However, by comparing adjacent encodings in Supplementary Figures S5 and S6, we found no significant dif-

Figure 4. Encoding groups performance, clustered by biomedical domain and encoding group. Color coding corresponds to the max F1-score of a group. The x-axis is arranged by clustering datasets, i.e. the biomedical application. The y-axis is organized by clustering sequence- and structure-based encodings. An interactive version of this plot can be found at https://peptidereactor.mathematik.uni-marburg.de/.

ferences (42). Furthermore, some of the outliers explain the gap between the best and the worst encodings, mentioned above. Overall, encodings from the same group are frequently among the best encodings, i.e. if two encodings are derived from the same group, but with different parameters, the performance is similar. By considering the receiver-operation characteristic (ROC)- and the Precision-Recall (PR)-curve areas of the overall top 6 encodings as well as the top 3 SeBEs and the top 3 StBEs, the observations mentioned above can be further endorsed (see Supplementary Figure S7).

Similarity

The similarity of the classifier outputs based on the Phi correlation indicates that encodings within groups and similar performing ones reveal a higher correlation (see Supplementary Figure S8). This can be verified by specifically considering SeBEs versus StBEs, which show in general a lower similarity. Furthermore, the diversity of the predictions, i.e. the disagreement measure of the classifier outputs, underpins these observations, since similar encodings as well as similar outputs leading to a lower diversity, hence greater similarity (see Supplementary Figure S8).

Class separation. With this respect, considering not only the diversity but also the probabilities predicted by a particular encoding combination, one can observe that the clustering quality, i.e. the classification capability of two encodings, measured by the Davis-Bouldin score (DBS), is often dependent on the diversity. In particular, by combining a well-performing SeBE and StBE, which show higher diversity compared to the best group-independent encodings, an increased DBS, hence better class separation, can be observed for the former (see Supplementary Figure S9).

However, this is not always the case (see Supplementary Figure S10). Albeit the encoding diversity and the DBS of

the clusters are related, the DBS seems to increase only until a particular diversity, meaning, that a too diverse classifier output negatively affects the class separation furthermore.

Critical difference. The observations made by the similarity measurements are further statistically revised by the critical difference of the respective classifier outputs. The critical difference unveils a great variety of encodings, which are not significantly different (see Supplementary Figure S11). Like above, this can be specifically observed for encodings from one group, which is in accordance with the previous experiments. However, the *psekraac_* and the *ngram_* groups are an exception (see Supplementary Figure S11). In addition, encodings, which surpass the critical threshold by several orders of magnitude, are less present (see Supplementary Figure S11).

Dataset correlation. Finally, the measured correlations, solely based on the encoded datasets, verify our observations made throughout the analyses (see Figure 5). The results illustrate foremost that encodings, originating from the same group, are clustered in separate branches. In addition, considering specifically StBEs, also here a clustering in an own sub-branch can be observed. This is in agreement with our findings from above, i.e. similar encodings are jointly clustered and thus, their predictions are also often related significantly.

Encoding recommendation

Based on our results elaborated above, we are not able to determine encodings, which can be specifically recommended for a particular application. However, following our findings a general guideline can be provided:

1. Some of the encoding groups are often among the top 3. Refer to Figure 3 for an overview and to which this applies in particular. Encodings from these groups are in general superior and should be preferably applied.
2. SeBEs are faster to compute and show in general a higher performance; thus, they should be preferred over the StBEs (see Figure 2 and Supplementary S13). However, combining SeBEs and StBEs to an ensemble classifier could outperform single SeBEs (see (7) and Supplementary Figure S9).
3. The dataset size should be also considered (see Supplementary Figure S12), i.e. we recommend for larger ones to carefully deliberate the choice of encodings. Contrary, for smaller datasets all encodings can be computed without hesitation.
4. A few encodings show better performance on imbalanced datasets. Refer to the Figure 3 for an overview and to which encodings/datasets combination this applies to.
5. Consider the size parameter for autocorrelation-based encodings (*cksaagp*, *cksaap*, *socnumber*, *qsorder*, *nmbroto*, *moran_*, *ksctriad*, *geary_*, *eaac*, *apaac_*, *paac*, *egaac_*, and *psekraac*). Shorter sequences require a smaller, for example, window size and vice versa.
6. Select solely one particular encoding from a parameterized encoding group. Encodings from the same group often show a similar performance (see Supplementary Figures S5, S8, and S11). This is due to highly correlated encoded datasets (see Supplementary Figure S5).
7. Use ensemble methods and aggregate different encodings to a meta learner in order to improve the performance.
8. For encodings that are seemingly relevant for a specific task, but fail in practice, extend the encoding choice iteratively, i.e. be less stringent with respect to the points mentioned above, in order to find encodings with improved performance.

DISCUSSION

We presented here, to the best of our knowledge, the first large-scale comprehensive study on peptide encodings. In particular, we aggregated numerous sequence- and structure-based encodings (SeBEs and StBEs, respectively) as well as datasets from a wide range of biomedical domains. Albeit proteins and peptides may exhibit multifunctionality (43), we limited our case study to two-class classification tasks. Hence, we can exclude that an insufficient size of the respective classes affects the prediction negatively, ultimately decreasing the complexity of this work and allowing for more robust conclusions.

The choice of the Random Forest classifier (RFC) as the default machine learning model also reduces the complexity. A hyper-parameter optimization (HPO) is less important as it would be for other models (30). In addition, the built-in feature selection discards irrelevant features, thus RFCs standardize the pre-condition for all encodings. This also reflects applied machine learning, where feature selection is a standard measure and encodings would be ultimately assessed based on their representative feature subset. Nevertheless, HPO (including the choice of the classifier) has the possibility to impact the encoding performance slightly. In order to cope with the computational feasibility we omitted an in-depth HPO. However, further research is necessary to address the impact of HPO on the encoding performance.

All in all, our study closes the gap between a broad range of peptide encodings and the challenge which to use on a specific biomedical dataset. We observed that no particular encoding group shows superior performance within a biomedical domain, i.e. no general pattern emerged from the respective encoding performance. However, insights are hereinafter discussed in more detail.

Performance

The encoding performance depends on two main characteristics. First, the class imbalance and second, the type, i.e. SeBE or StBE. While the former is not surprising, as it needs more sophisticated measures for coping, the second is potentially due to the initial tertiary structure approximation. Thus, in many cases, the structure is probably unrelated with the *in vivo* one. In contrast, for the database, we used only sequences with a known structure deposited at the PDB. This could be the reason, why the general performance of StBEs is lower compared to SeBEs, but the predictions are still satisfying. We suspect that disordered regions also affect the prediction negatively, since no conformational information can be derived from it.

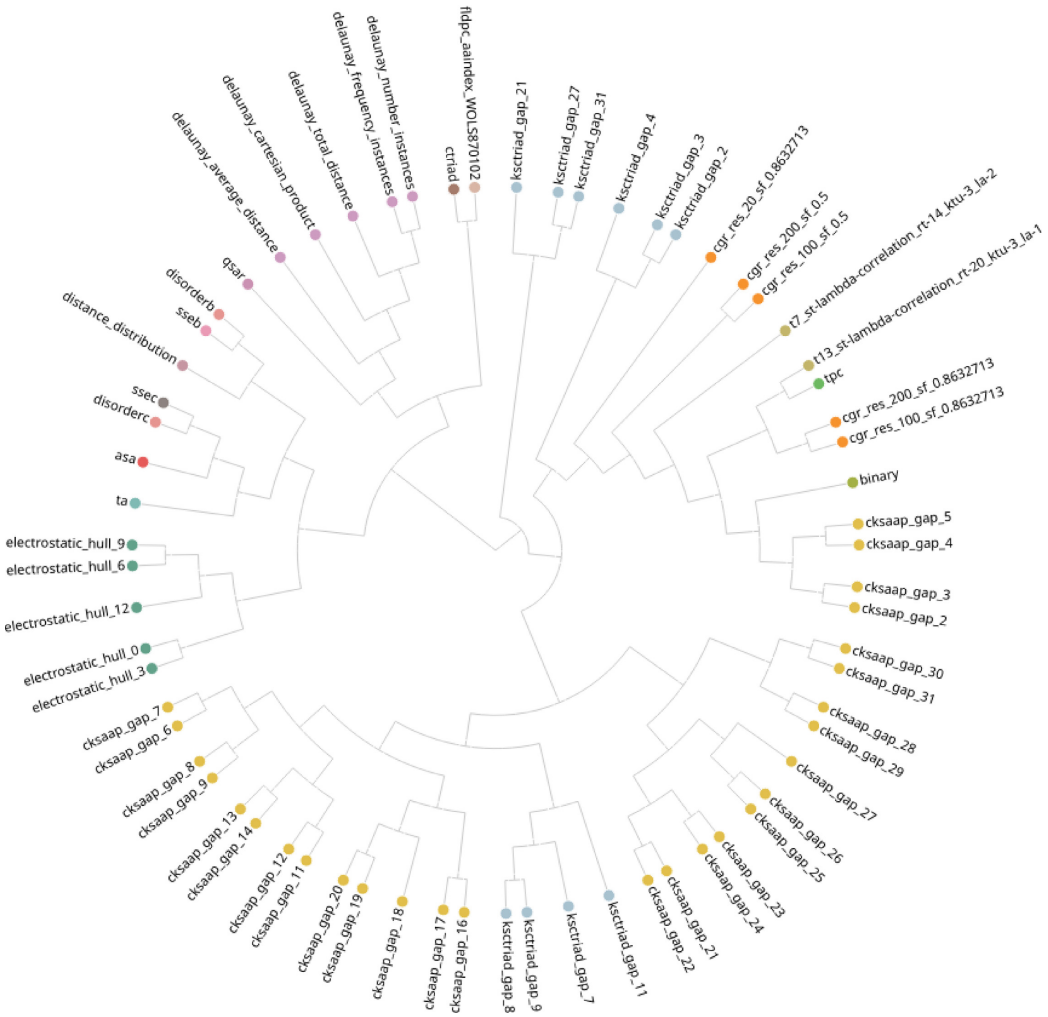


Figure 5. Top 50 datasets correlation. Correlation of top 50 encoded datasets based on the adjusted RV-coefficient. Color coding corresponds to the encoding group. The graphic shows the example of the *hiv_ddi* dataset. The online version of this plot can be found at https://peptidereactor.mathematik.uni-marburg.de/.

A further reason for a convincing prediction is the sequence similarity within and across the positive and the negative class. Regarding the former, the *qsar* encodings extraordinary performance for some of the *hiv* datasets (see Figure 3) could be due to overfitting, owing to very similar sequences (see Supplementary Figure S2). With this respect, the *hiv_v3* dataset verifies this hypothesis further, as it contains very similar sequences and almost all encodings demonstrate very high performance on this dataset. In addition, the sequence embedding, shown in Supplementary Figure S2, provides a further, visual explanation. Finally, a low inter class similarity affects the class separation positively, which can also be observed by considering solely the performance of the amino composition encoding (see Supplementary Figure S3).

Albeit no real pattern emerges on the performance within biomedical domains one can still observe slightly similar results on these datasets (see Supplementary Figure S1), presumably owing to redundant sequences. We collected the datasets as they are and many studies build upon each other, which explains overlapping sequences in some cases.

As mentioned above, the class imbalance as well as the encoding type contribute mainly to the encoding performance. This explains also the result of the clustering, i.e. two major clusters for datasets and encoding groups. An exception refers to the *ksctriad* encoding, which clusters adjacent to

StBEs, likely due to missing values (NA, see Figure 4) since for too short sequences this encoding type cannot be calculated. In addition, despite similar performance, it is not possible to draw conclusions on a similar function of the encodings. Far more datasets of the same biomedical application would have been necessary.

The similar performance of within-group encodings can be explained by adjacent parameter configurations, for example, a slightly larger gap length or window size (see Supplementary Figure S5), probably leading to only a marginal information change. Moreover, this observation supports our conducted *psekraac_* filtering, since it is likely that many of these encodings would perform similar, which in turn question the necessity of computing all of them.

Similarity

The parameter configuration space for encodings emerging from the same group could also explain the similarity of the classifier outputs. That is, adjacent parameters provide no further or new insights for the machine learning model. This is in accordance with Kuncheva *et al.* (2003), who stated that diversity is a crucial condition for effective ensemble learning by mutually compensating weaknesses of single models (31). Certainly, this would not be possible if the classifier output is too similar. This is also the reason to consider SeBEs and StBEs, which show continuously low similarity (see Supplementary Figure S8) but also satisfying performance (see Figure 4). However, we observed, that the diversity cannot be arbitrarily high, since a greater diversity does not necessarily imply an improved class separation (see supplementary Figure S9).

The general trend, i.e. encodings from the same group show similar performance and lead to similar predictions can be verified by the statistical assessment, ultimately revealing a great variety of non-significant differences (see Supplementary Figure S11). The dataset correlation supports these observation impressively (see Supplementary Figure S5). The exceptions are the *pskraac_* and the *ngram_* encoding group, which is due to different sub-types, intrinsically generating different, within-group encodings.

Time versus performance

The total computing time depends on the dataset size, i.e. the more sequences, the longer the required computation (see Supplementary Figure S12). A more detailed look at the total amount of sequences per dataset indicates that the computation time depends on the dataset size (see Supplementary Figure S12). However, the mean sequence length does not necessarily lead to an increased calculation time (see Supplementary Figure S12).

Moreover, some of the encodings impact the duration crucially, above all the StBEs (see Supplementary Figure S13). One can observe, that the majority of the SeBEs require less computation time and demonstrate at the same time a higher performance. We added the elapsed time required for the tertiary structure approximation to the total computation time of StBEs; thus, the calculation of the latter is in general prolonged. In addition, the tertiary structure approximation and the associated *electrostatic_hull* encoding as well as the *cgr* and *fldpc_* encoding, and finally the *psekraac_* filtering are main contributors to the total run time (see Supplementary Figure S13).

Encoding recommendation

The recommendations serve as a general guideline, i.e. researchers have to decide case-wise, which encodings to use in particular. Some of the encodings seem to be redundant and usage is not reasonable at the first glance. However, using ensemble methods could compensate for weaknesses of single encodings, thus, even those encodings are applicable. This is also a matter of the dataset size and available resources. Moreover, although some encodings seem to work on imbalanced datasets, more research is necessary to draw meaningful conclusions.

CONCLUSION

Our study marks the first comprehensive benchmark on various peptide encodings and we demonstrated, that in general, the performance of all encodings is similar and more or less independent from the biomedical task at hand. This allows us to reduce the vast number of encodings dramatically, paving the way for more sophisticated optimization methods in the future. A potential application refers to automated ensemble classifier configuration or to extend established automated machine learning methods like *auto-sklearn* (44). With this respect, a challenge remains the continuous search space, which could be tackled with pre-computed diversity measures to transform categorical hyperparameters (encodings) into numerical ones. Additional research is also necessary to verify whether and how StBEs can exhaust their full potential as part of ensemble classifiers. However, datasets with many sequences aligning to disordered regions can decrease the usability of StBEs clearly.

Our reproducible, parallelized pipeline conducts different analyses in order to get an expressive picture of the encoding performance across multiple biomedical domains. The results are aggregated across multiple biomedical domains and revamped as part of a great variety of interactive visualizations. All standardized datasets are available for download to comply with FAIR standards. The PEPTIDE REACToR allows researchers not only comparison at one glance, but also provides the state of the art for future encoding benchmarks, bundled in a single platform. With this respect, an extension is conceivable in order to allow researchers to upload their own (private) datasets.

DATA AVAILABILITY

The results can be interactively accessed at https://peptidereactor.mathematik.uni-marburg.de/. The source code is available at https://github.com/spaenigs/peptidereactor. Due to the large size, intermediate data as well as intermediate results are available upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB online.

ACKNOWLEDGEMENTS

We especially thank Martin Braun from the de.NBI cloud location Berlin for his technical support. An early version of this workflow relied on resources from the MaRC2 high-performance cluster of the University of Marburg. Here, we would like to thank in particular René Sitt for his technical guidance. Furthermore, we are grateful for the contributions of Markus Flicke, Christopher Zapp, and Roman Martin.

Author contributions: S.S. and D.H. developed the concept. S.S. designed and performed the experiments as well as gathered, curated and analyzed the data. S.M. implemented the Delaunay Triangulation and Distance Frequency encoding. G.H. supervised the data visualization aspect and created the logo as well as the overview figure. S.S., A.C.H. and D.H. interpreted the results. S.S. wrote the manuscript. A.C.H. and D.H. supervised the study. All authors read and approved the final manuscript.

FUNDING

Bundesministerium für Wirtschaft und Energie [16KN0742325]; Bundesministerium für Bildung und Forschung [031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537B, 031A537C, 031A538A, 031A537D, 031A537A].

Conflict of interest statement. None declared.

REFERENCES

1. Fjell,C.D., Hiss,J.A., Hancock,R.E. and Schneider,G. (2012) Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug. Discov.*, **11**, 37–51.
2. Sanders,W.S., Johnston,C.I., Bridges,S.M., Burgess,S.C. and Willeford,K.O. (2011) Prediction of cell penetrating peptides by support vector machines. *PLoS Comput. Biol.*, **7**, e1002101.
3. Heider,D., Dybowski,J.N., Wilms,C. and Hoffmann,D. (2014) BioData mining a simple structure-based model for the prediction of HIV-1 co-receptor tropism. *BioData Min.*, **7**, 14.
4. Zhao,Y., Pinilla,C., Valmori,D., Martin,R. and Simon,R. (2003) Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, **19**, 1978–1984.
5. Wu,C., Whitson,G., Mclarty,J., Adisorn,E. and Chang,T.-C. (1992) Protein classification artificial neural system. *Protein Sci.*, **1**, 667–677.
6. Chen,Z., Zhao,P., Li,F., Leier,A., Marquez-Lago,T.T., Wang,Y., Webb,G.I., Smith,A.I., Daly,R.J., Chou,K.C. *et al.* (2018) IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**, 2499–2502.
7. Spänig,S. and Heider,D. (2019) Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Min.*, **12**, 7.
8. Thakur,N., Qureshi,A. and Kumar,M. (2012) AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.*, **40**, W199–W204.
9. Manavalan,B., Basith,S., Shin,T.H., Choi,S., Kim,M.O. and Lee,G. (2017) MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, **8**, 77121–77136.
10. Liu,B., Gao,X. and Zhang,H. (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127.
11. Drozdetskiy,A., Cole,C., Procter,J. and Barton,G.J. (2015) JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.
12. Bose,P. and Harrison,R.W. (2011) Encoding protein structure with functions on graphs. In: *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. Atlanta, GA, pp. 338–344.
13. Löchel,H.F., Riemenschneider,M., Frishman,D. and Heider,D. (2018) SCOTCH : Subtype A Coreceptor Tropism Classification in HIV-1. *Bioinformatics*, **34**, 2575–2580.
14. Nagpal,G., Usmani,S.S., Dhanda,S.K., Kaur,H., Singh,S., Sharma,M. and Raghava,G.P. (2017) Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci. Rep.-UK*, **7**, 42851.
15. Manavalan,B., Govindaraj,R.G., Shin,T.H., Kim,M.O. and Lee,G. (2018) iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.*, **9**, 1695.
16. Dybowski,J.N., Riemenschneider,M., Hauke,S., Pyka,M., Verheyen,J., Hoffmann,D. and Heider,D. (2011) Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Min.*, **4**, 26.
17. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci.Data*, **3**, 160018.
18. Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
19. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2015) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
20. Faraggi,E., Zhang,T., Yang,Y., Kurgan,L. and Zhou,Y. (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.*, **33**, 259–267.
21. Peng,J. and Xu,J. (2011) Raptorx: exploiting structure information for protein alignment by statistical inference. *Proteins: Struct. Funct. Bioinform.*, **79**, 161–171.
22. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
24. Berman,H.M., Westbrook,J.D., Feng,Z., Gilliland,G.L., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
25. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
26. Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
27. Löchel,H.F., Eger,D., Sperlea,T. and Heider,D. (2020) Deep learning on chaos game representation for proteins. *Bioinformatics (England)*, **36**, 272–279.
28. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
29. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
30. Caruana,R. and Niculescu-Mizil,A. (2006) An empirical comparison of supervised learning algorithms. In: *ICML '06: Proceedings of the 23rd international conference on Machine learning*. Pittsburgh, pp. 161–168.
31. Kuncheva,L.I. (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, **51**, 181–207.
32. Skalak,D.B. (1996) The sources of increased accuracy for two proposed boosting algorithms. In: *Proc. American Association for Arti Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*. Portland, pp. 120–125.
33. Santafe,G., Inza,I. and Lozano,J.A. (2015) Dealing with the evaluation of supervised classification algorithms. *Artif. Intell. Rev.*, **44**, 467–508.
34. Demšar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.
35. Calvo,B. and Santafé,G. (2016) scmamp: statistical comparison of multiple algorithms in multiple problems. *R J.*, **8**, 248–256.

36. Mayer,C.D., Lorent,J. and Horgan,G.W. (2011) Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. *Stat. Appl. Genet. Mol. Biol.*, **10**, doi:10.2202/1544-6115.1540.
37. Indahl,U.G., Næs,T. and Liland,K.H. (2018) A similarity index for comparing coupled matrices. *J. Chemometr.*, **32**, e3049.
38. Bouguettaya,A., Yu,Q., Liu,X., Zhou,X. and Song,A. (2015) Efficient agglomerative hierarchical clustering. *Expert. Syst. Appl.*, **42**, 2785–2797.
39. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
40. VanderPlas,J., Granger,B., Heer,J., Moritz,D., Wongsuphasawat,K., Satyanarayan,A., Lees,E., Timofeev,I., Welsh,B. and Sievert,S. (2018) Altair: Interactive Statistical Visualizations for Python. *J. Open Source Software*, **3**, 1057.
41. Hattab,G., Rhyne,T.M. and Heider,D. (2020) Ten simple rules to colorize biological data visualization. *PLoS Comput. Biol.*, **16**, e1008259.
42. Krzywinski,M. and Altman,N. (2014) Visualizing samples with box plots. *Nat. Methods*, **11**, 119–120.
43. Diener,C., Garza Ramos Martínez,G., Moreno Blas,D., Castillo González,D.A., Corzo,G., Castro-Obregon,S. and Del Rio,G. (2016) Effective Design of Multifunctional Peptides by Combining Compatible Functions. *PLoS Comput. Biol.*, **12**, e1004786.
44. Feurer,M., Klein,A., Jost,K.E., Springenberg,T., Blum,M. and Hutter,F. (2015) Efficient and robust automated machine learning. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Montréal, pp. 2962–2970.

6.4 Unsupervised Encoding Selection through Ensemble Pruning for Biomedical Classification

Numerous studies dealt with the classification of peptide properties, for instance, immunosuppressive[89], cell-penetrating[154], and antimicrobial efficiency[165], or drug resistance prediction[146]. Researchers must select ML classifiers, for instance, the Naïve Bayes classifier (NBC), Logistic Regression classifier (LRC), or Decision Tree classifier (DTC). Nevertheless, multiple studies demonstrated that merging base classifiers can improve the performance[75,86,216]. Encodings are of significant relevance since classifiers require numerical and fixed-length input[32]. However, many encodings expand hyper-parameter optimization (HPO). Specifically, researchers are faced with a complex workflow, including encoding selection, optimal model configuration, and evaluation of ensemble fusion methods.

Considering the 20 natural amino acids, millions of active peptides are possible. Thus, automated approaches are required to screen the vast space of possible combinations. In addition, pipelines should conduct encoding selection and HPO. To this end, Chen *et al.* (2021) introduced an ML tool for nucleic acid and amino acid sequences[38]. The program provides a graphical user interface to assemble classification workflows; however, it only offers manual encoding selection[38]. Features for full automation, comprising unsupervised sampling of encodings and ensembles, are lacking.

In the present study, we considered models trained on individual encodings as base classifiers. We utilized convex hull and Pareto frontier pruning[129] which can process even hundreds of encodings. In addition, we investigated the multi-verse optimizer (MVO) as a further ensemble generator[2]. The output of the base models, NBC, LRC, and DTC, is fused by majority vote, averaging, and stacked generalization. Furthermore, the Random Forest classifier (RFC), effectively an ensemble model, is committed as another base model to examine its robustness[25].

The realized workflow features unsupervised encoding selection and ensemble configuration. The tool is extensible by other base and ensemble methods. All results are collected and visually depicted, allowing researchers to determine good ensembles readily. The study demonstrated that the base model and encoding choice affect performance primarily. In contrast, the fusion method is seemingly irrelevant.

The Pareto frontier pruning is an efficient strategy for unsupervised encoding selection and ensemble configuration. In contrast, the MVO produces inferior ensembles, and the optimization suffers from a long run time. Ensembles using the RFC as a base model revealed only marginal performance gain compared to the individual model. Ensembles employing other base models are occasionally superior.

Currently, the pipeline utilizes straightforward ensemble methods. Thus, more research is necessary to incorporate other ensembles, for instance, boosting, which requires the adap-

tation of the involved base models[266].

In conclusion, the study enables unsupervised encoding selection and ensemble construction. We leveraged ensemble pruning to deal with potentially hundreds of encoded datasets. Researchers can readily assess the various encodings, models, and fusion methods by providing statistics and visualizations. Our work is a significant step towards automated biomedical classification and bridges the gap between many peptide encodings and diverse ML models.

METHODOLOGY

Unsupervised encoding selection through ensemble pruning for biomedical classification

Sebastian Spänig, Alexander Michel and Dominik Heider*

*Correspondence:
dominik.heider@uni-marburg.de
Data Science in Biomedicine, Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany
Full list of author information is available at the end of the article

Abstract

Background

Owing to the rising levels of multi-resistant pathogens, antimicrobial peptides, an alternative strategy to classic antibiotics, got more attention. A crucial part is thereby the costly identification and validation. With the ever-growing amount of annotated peptides, researchers employed artificial intelligence to circumvent the cumbersome, wet-lab-based identification and automate the detection of promising candidates. However, the prediction of a peptide's function is not limited to antimicrobial efficiency. To date, multiple studies successfully classified additional properties, e.g., antiviral or cell-penetrating effects. In this light, ensemble classifiers are employed to utilize the advantages of peptide encodings; hence, further improving the prediction. Although we recently presented a workflow to significantly diminish the initial encoding choice, an entire unsupervised encoding selection, considering various machine learning models, is still lacking.

Results

We developed a workflow, automatically selecting encodings and generating classifier ensembles by employing sophisticated pruning methods. We observed that the Pareto frontier pruning is a good method to create encoding ensembles for the datasets at hand. In addition, encodings combined with the Decision Tree classifier as the base model are often superior. However, our results also demonstrate that none of the ensemble building techniques is outstanding for all datasets.

Conclusion

The workflow conducts multiple pruning methods to evaluate ensemble classifiers composed from a wide range of peptide encodings and base models. Consequently, researchers can use the workflow for unsupervised encoding selection and ensemble creation. Ultimately, the extensible workflow can be used as a plugin for the PEPTIDE REACToR, further establishing it as a versatile tool in the domain.

Keywords: Biomedical classification; Antimicrobial peptides; Encodings; Machine Learning; Ensemble Learning

Background

Multi-resistant pathogens are a major threat for modern society [1]. In the last decades, a rising number of bacterial species developed mechanisms to elude efficiency to widely used antibiotics [1]. The importance of developing and implement-

ing alternative strategies is further underpinned by a recent study, which detected a certain baseline resistance in European freshwater lakes [2]. The study confirmed resistance specifically against four critical drug classes in human and veterinary health in freshwater, which is typically considered as a pathogen-free environment [2]. Moreover, already concerning levels of antibiotic resistance in Indian and Chinese lakes emphasize the requirement of alternative biocides [3, 4]. One promising approach to replace or even support common antibiotics refers to the deployment of peptides with antimicrobial efficiency [5]. However, identifying and validating active peptides requires intensive, hence, costly and time-consuming wet-lab work. Thus, in the pre-artificial intelligence (AI) era, the manual classification and verification of antimicrobial peptides (AMPs) engaged researchers. Although the *in vitro* confirmation of activity is still necessary, the application of AI, i.e., in particular machine learning (ML) algorithms, simplifies the identification process drastically and pushes specific AMPs to the second or third phase of clinical trials [6]. In addition, online databases provide access to thousands of annotated sequences and pave the way for AI application in peptide design and classification [7]. For instance, Chung *et al.* (2019) developed a method, which demonstrated good performance on classifying AMPs using a two-step approach, which first predicts efficiency, and afterward the precise target activity [8]. Another study employed a variational autoencoder to encode AMPs, mapped the probability of being active to a latent space, and predicted novel AMPs [9]. Fingerhut *et al.* (2020) introduced an algorithm to detect AMPs from genomic data [10]. For more information on computational approaches for AMP classification, we refer to the recent review of Aronica *et al.* (2021) [11].

However, the prediction of amino acid sequence features is not limited to AMPs. In the literature, one can find various applications, e.g., in oncology for predicting anticancer peptides [12], in pharmacology for the discovery and application of cell-penetrating peptides as transporters for molecules [13], or in immunotherapy, for classifying of pro- or antiinflammatory peptides [14, 15]. Other applications include antiviral peptides [16], or peptides with hemolytic [17] or neuro transmitting activity [18].

Unequivocally, the success of ML methods for the prediction of AMPs was enabled by the development and advances of peptide encodings. Encodings are algorithms mapping the amino acid sequences of different lengths to numerical vectors of an equal length, hence, fulfilling the requirement of many ML algorithms [19]. Moreover, peptides or proteins can be described by their primary structure, i.e., the amino acid sequence, and the aggregation in higher dimensions, denoted as the secondary or tertiary structure. Encodings derived from the primary structure are known as sequence-, and encodings describing a higher-order folding are structure-based encodings. To date, a large number of sequence- and structure-based encodings have been introduced and employed in various studies [19]. A significant amount of encodings has been recently acknowledged by another study, specifically benchmarking these by considering multiple biomedical applications [20]. It turned out that most encodings show acceptable performance, partly also beyond single biomedical domains [20]. In addition, Spänig *et al.* (2021) developed a workflow, which can dramatically reduce the number of initial encodings [20]. However, encoding selection is still challenging, and user-friendly approaches are required.

Furthermore, hyperparameter optimization is additionally aggravated by the model choice. Albeit Support Vector Machines (SVM) and Random Forests (RF) are widely employed in peptide classification [11], the variety of models used in a broad range of studies is large. For instance, Khatun *et al.* (2020) utilized several ML algorithms, including Naïve Bayes, AdaBoost, and a fusion-based ensemble for the prediction of proinflammatory peptides [21]. The fusion-based model outperformed the other ML models significantly for this task [21]. Plisson *et al.* (2020) employed Decision Trees (DT) and Gradient Boosting (GB), among others, to classify non-hemolytic peptides and demonstrated that the GB ensemble has superior performance [22]. In contrast, Timmons *et al.* (2020) used Artificial Neural Networks to characterize therapeutic peptides with hemolytic activity [23]. Singh *et al.* (2021) compared several base classifiers, e.g., Linear Discriminant Analysis and ensemble methods, e.g., GB and Extra Trees to detect AMPs [24]. They demonstrated that the GB performed best [24]. These studies clearly show that ensemble classifiers typically show superior performance than single classifiers, which is based on the fact that they can compensate for weaknesses of single encodings and base classifiers [25].

Recently, Chen *et al.* (2021) introduced a comprehensive tool, which allows less programming experienced researchers to simply select encodings and base or ensemble classifiers through a graphical user interface, allowing easy access to the underlying algorithms [26]. Nevertheless, the approach assumes that the user selects proper settings for the parameterized encodings, which has been previously shown to affect the classification process significantly [20]. Moreover, the encoding selection is independent of the classifier settings, meaning that the tool can set up the classifier automatically; however, the encoding selection is not part of it. Thus, it remains a challenge to pick good encodings and classifiers for a biomedical classification task at hand. To this end, we assessed unsupervised encoding selection and the performance and diversity of multiple ensemble methods. We added different overproduce-and-select techniques for ensemble pruning, facilitating an automatic ensemble generation. In addition, we utilized Decision Trees, Logistic Regression, and Naïve Bayes as base classifiers, owing to their prevalence in the field of biomedical classification due to their explainability [11, 19, 27].

Besides demonstrating the benefit of an unsupervised encoding selection, we also examined how the RF performs as a base and ensemble classifier, i.e., whether the RF, an ensemble method per se, is performance-wise already saturated or whether a subsequent fusion can improve the final predictions. Fusion of RFs has been shown in other studies to improve overall performance, e.g., for HIV tropism predictions [28, 29]. All in all, we complement our recent large-scale study on peptide encodings [20] with an automatic encoding selection and a performance analysis of multiple base and ensemble classifiers. Ultimately, the present research bridges the gap between many peptide encodings and available machine learning models.

Results

We developed an end-to-end workflow, which automatically generates and assesses classifier ensembles using different pruning methods and a variety of encoded datasets from multiple biomedical domains (see Table 4). Data scientists can easily

Table 1 The table shows the performance comparison (including RF) of classifier ensembles derived from different pruning methods and the single best classifier. Numbers refer to the mean performance of a 100-fold Monte Carlo cross-validation. Standard deviation (SD) is added in brackets. Mean and SD are rounded to 2 decimal places. The top base/ensemble classifier combination is always used (see Fig. 2). Classifier ensembles are significantly better than the single best classifiers. In particular, except for one case, the Pareto frontier pruning (pfront) generates the best ensembles. Significance levels are as follows: ** $p \leq 0.001$, * $p \leq 0.01$, and . $p \leq 0.05$.

	best	chull	mvo	pfront	rand	rand_single_best	single_best
acp_mlacp	0.73 (±0.06)	0.73 (±0.06)	0.7 (±0.07)	0.74** (±0.06)	0.69 (±0.06)	0.68 (±0.07)	0.69 (±0.07)
aip_antiinflam	0.5** (±0.04)	0.5 (±0.04)	0.45 (±0.04)	0.5 (±0.04)	0.48 (±0.04)	0.47 (±0.04)	0.47 (±0.04)
amp_antibp2	0.88 (±0.02)	0.89 (±0.02)	0.88 (±0.02)	0.9** (±0.02)	0.87 (±0.02)	0.84 (±0.02)	0.87 (±0.03)
atb_antitbp	0.75 (±0.07)	0.76 (±0.08)	0.72 (±0.08)	0.79** (±0.07)	0.7 (±0.06)	0.66 (±0.07)	0.68 (±0.07)
avp_amppred	0.79 (±0.03)	0.8 (±0.03)	0.77 (±0.02)	0.81** (±0.03)	0.79 (±0.03)	0.76 (±0.03)	0.76 (±0.03)
cpp_mlcpp	0.77 (±0.03)	0.78 (±0.03)	0.78 (±0.03)	0.79** (±0.03)	0.76 (±0.03)	0.74 (±0.03)	0.75 (±0.03)
hem_hemopi	0.88 (±0.03)	0.89 (±0.03)	0.87 (±0.03)	0.89** (±0.03)	0.88 (±0.03)	0.86 (±0.03)	0.87 (±0.03)
isp_il10pred	0.59 (±0.05)	0.59 (±0.05)	0.6 (±0.06)	0.6** (±0.05)	0.57 (±0.05)	0.58 (±0.04)	0.58 (±0.04)
nep_neuropipred	0.79 (±0.03)	0.81 (±0.02)	0.81 (±0.04)	0.81** (±0.03)	0.81 (±0.03)	0.76 (±0.03)	0.78 (±0.03)
pip_pipel	0.5 (±0.04)	0.52 (±0.04)	0.5 (±0.05)	0.53** (±0.04)	0.47 (±0.04)	0.41 (±0.04)	0.49 (±0.03)

Table 2 The table shows the performance comparison (excluding RF) of classifier ensembles derived from different pruning methods and the single best classifier. See Table 1 for more details.

	best	chull	mvo	pfront	rand	rand_single_best	single_best
acp_mlacp	0.72 (±0.06)	0.73 (±0.06)	0.69 (±0.04)	0.74** (±0.06)	0.69 (±0.06)	0.66 (±0.07)	0.67 (±0.07)
aip_antiinflam	0.47 (±0.04)	0.48 (±0.04)	0.44 (±0.05)	0.48** (±0.04)	0.41 (±0.04)	0.36 (±0.04)	0.44 (±0.04)
amp_antibp2	0.88 (±0.02)	0.88 (±0.02)	0.87 (±0.02)	0.89** (±0.02)	0.86 (±0.02)	0.84 (±0.02)	0.87 (±0.03)
atb_antitbp	0.73 (±0.06)	0.76 (±0.08)	0.67 (±0.04)	0.79** (±0.07)	0.68 (±0.07)	0.65 (±0.08)	0.68 (±0.07)
avp_amppred	0.76 (±0.04)	0.77 (±0.04)	0.73 (±0.02)	0.81** (±0.03)	0.74 (±0.04)	0.7 (±0.04)	0.73 (±0.03)
cpp_mlcpp	0.74 (±0.03)	0.75 (±0.03)	0.73 (±0.02)	0.78** (±0.03)	0.74 (±0.03)	0.71 (±0.03)	0.71 (±0.03)
hem_hemopi	0.87 (±0.03)	0.89 (±0.03)	0.87 (±0.03)	0.89** (±0.03)	0.86 (±0.03)	0.86 (±0.03)	0.86 (±0.03)
isp_il10pred	0.59 (±0.05)	0.57 (±0.05)	0.59 (±0.08)	0.6** (±0.05)	0.57 (±0.05)	0.58 (±0.04)	0.58 (±0.04)
nep_neuropipred	0.79 (±0.03)	0.79 (±0.03)	0.79 (±0.02)	0.8** (±0.03)	0.74 (±0.03)	0.65 (±0.04)	0.78 (±0.03)
pip_pipel	0.48 (±0.04)	0.45 (±0.04)	0.47 (±0.05)	0.48** (±0.04)	0.45 (±0.03)	0.38 (±0.03)	0.38 (±0.03)

extend the workflow with different base and ensemble classifiers, pruning methods, encodings, and datasets. The results can be reviewed using the provided data visualizations, and the performance is further revised using multiple statistics. We demonstrate that the Pareto frontier pruning is a valuable technique to generate efficient classifier ensembles. However, the utilized base classifiers show comparable performance, with the Decision Tree classifier being the model of choice for most datasets. We address the results in more detail in the following. We use the example of the avp_amppred dataset throughout the manuscript. The results for the remaining datasets can be found in the supplement. Moreover, the code is publicly available at https://github.com/spaenigs/ensemble-performance. Note that the workflow produces interactive versions of all charts.

Pruning methods

All pruning methods generate ensembles, i.e., combined encodings, superior to the single best classifier, i.e., individual encodings (see Tables 1 and 2). In the case of the Pareto frontier (pfront) pruning, which is predominantly ranked among the best pruning methods, we observe a significant ($p \leq 0.001$) performance improvement compared to the single best classifier. We also observed that the pfront pruning generates larger ensembles than the convex hull (chull) pruning, which can be visually verified in Fig. 1 (red line). Notably, including the Random Forest (RF) classifier (see Table 1, pfront) does not, or very slightly, affect the ensemble performance without RF (see Table 2), although the single best classifier performance is better with the RF included (see Table 1). Consequently, the RF increases the overall performance of the ensembles generated by the best encodings pruning. Finally, the multi-verse optimization (MVO) suffers from high computational demand, i.e., a long pruning time, and in general, an inferior performance compared to the other techniques.

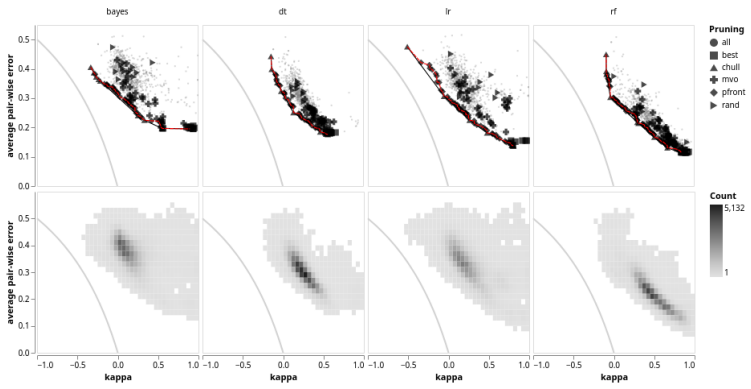


Figure 1 The kappa-error diagram depicts pairs of base classifiers (gray dots) using the kappa diversity on the x-axis and the average error on the y-axis (top row). Each classifier pair denotes two particular encodings. The black line indicates the convex hull, and the red line the Pareto frontier. Ensemble types are represented as symbols. The top row shows the example of the first fold. The bottom row groups the results of all folds in a 2D histogram. The darker the color, the more classifier pairs are binned in one group. The gray line in all panels depicts the theoretical boundary (refer to the method section for more details). The plot shows the example of the avp_amppred dataset.

Ensemble classifiers

The ensemble performance mainly depends on the pruning and the choice of the base classifiers; hence, the collection of individual encodings. Thus, the performance differences among the single best (single_best) and best random (rand) pruning are insignificant, which is in contrast to the remaining methods (see Fig. 2). Furthermore, no significant difference can be observed for ensembles with the same base classifiers, e.g., the RF or Decision Tree (DT). Thus, the fusion method impacts the overall performance slightly. However, various base classifiers result in significantly different ensembles, i.e., employing, for instance, the RF, generates significantly different ensembles compared to the application of other base classifiers (see Fig. 2).

Moreover, it is noticeable that the Naïve Bayes (NB) and the Logistic Regression (LR) classifiers result in ensembles with higher variance (see Fig. 1). In contrast, the area covered by RF and DT models is more compact. Therefore, the variables, i.e., diversity and the pairwise error, are revised by a multivariate analysis of variance (MANOVA), which revealed a significant difference ($p < 0.001$). A separate examination of the variables utilizing variance analysis (ANOVA) followed by a post-hoc analysis using Tukey's HSD, demonstrates that all variables are significantly different ($p < 0.001$). Finally, we conducted an ANOVA on the particular area values, which disproves the initial observation, i.e., all areas are significantly different ($p < 0.001$). However, considering the average values for all datasets, the DT and RF are commonly ranked as the base classifiers with low variance (see Table 3).

Single classifiers

In general, the performance of the base classifiers, i.e., single encodings, is lower compared to the classifier ensembles (see Fig. 3). We also observed that the ran-

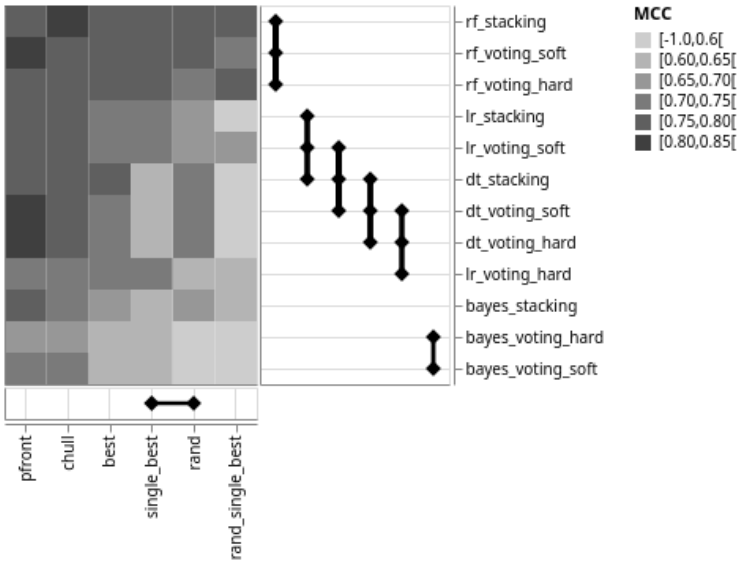


Figure 2 The XCD chart shows the difference between pruning methods (x-axis) and ensembles (y-axis). Entities connected with a bold line are not significant different. The higher the performance, the darker the color. The plot shows the example of the avp_amppred dataset.

domly selected model (rand_single_best) is inferior to the best model (single_best). In addition, we noticed that the RF is relatively saturated, i.e., using the RF as a single classifier and as a base model for ensembles does not have a significant effect on performance improvement. The low-performance variance is in line with the observation that weak models benefit most from ensemble learning; however, RFs are ensemble models [30, 31]. In contrast, the performance of other single classifiers revealed more distinct differences to the ensembles (see Fig. 3).

Data visualization

We leveraged two standard visualization techniques, which we adapted and extended for our particular application. First, we enhanced the kappa-error diagram [25] for

Table 3 The table lists the average area (±SD) covered by the base classifiers across the 100-fold Monte Carlo cross-validation. The lowest area per dataset is highlighted in bold. The DT classifier has the lowest area for most of the datasets, i.e., the predictions are more stable. Refer to Fig. 1 (bottom) for the example showing the avp_amppred dataset.

	bayes	dt	lr	rf
acp_mlacp	3.15 (±0.073)	2.6 (±0.08)	2.66 (±0.046)	**2.55** (±0.081)
aip_antiinflam	2.75 (±0.066)	2.27 (±0.045)	2.41 (±0.033)	**2.14** (±0.045)
amp_antibp2	3.01 (±0.077)	**2.5** (±0.056)	3.07 (±0.122)	2.63 (±0.061)
atb_antitbp	3.18 (±0.124)	2.82 (±0.059)	3.16 (±0.094)	**2.73** (±0.069)
avp_amppred	2.96 (±0.054)	**2.38** (±0.05)	3.15 (±0.081)	2.42 (±0.054)
cpp_mlcpp-complete	2.9 (±0.086)	**2.37** (±0.073)	2.48 (±0.049)	2.42 (±0.079)
hem_hemopi	3.2 (±0.06)	**2.74** (±0.135)	3.07 (±0.076)	2.79 (±0.122)
isp_il10pred	2.96 (±0.059)	2.45 (±0.046)	**2.38** (±0.031)	2.39 (±0.047)
nep_neuropipred	3.23 (±0.132)	**2.61** (±0.081)	3.18 (±0.309)	2.68 (±0.079)
pip_pipel	3.18 (±0.053)	**2.19** (±0.034)	2.23 (±0.024)	2.3 (±0.069)

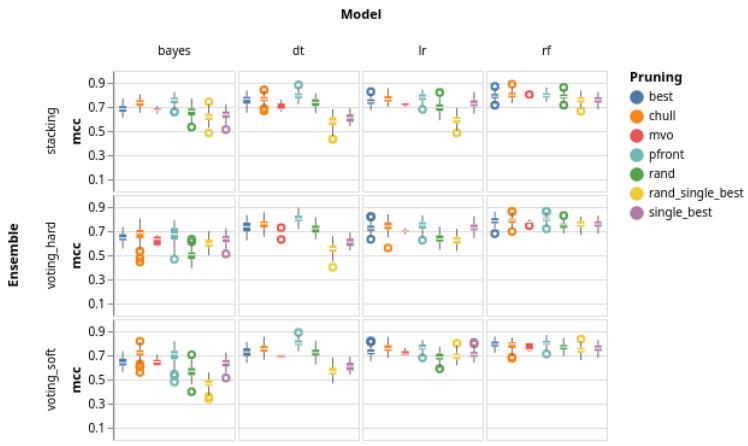


Figure 3 The box plot shows the Matthews correlation coefficients (MCC) distribution using a 100-fold Monte Carlo cross-validation. The columns include the base, i.e., diverse encodings, and the rows depict the ensemble classifiers. Colors refer to the pruning method. Note that the distributions of the best single (single_best) and best random single models (rand_single_best) are independent of the ensemble type. Variations within rand_single_best are due to the random selection. Moreover, MVO pruning (mvo) has only been conducted on the first five folds (see Discussion). The plot shows the example of the avp_amppred dataset.

the presentation of multiple folds, i.e., 100 in the current study, by aggregating the cross-validation results into a two-dimensional histogram (see Fig. 1). The color code allows the viewer to spot the peak at one glance. Hence, the tendency of ensembles to use a specific base classifier. Moreover, considering the distribution of the variables, one can make conclusions about the robustness.

Second, we extended the critical difference (CD) chart [32] with a categorical heatmap displaying the actual performance. The extension enables viewers to statistically compare classifiers and review the individual encoding performance, i.e., Matthews correlation coefficient in the present case, at one glance. In addition, the thickness of the vertical and horizontal rules is directly related to the critical difference, i.e., the thicker the rule, the closer the classifiers to the critical difference. Thus, the rule thickness provides an additional visual channel to access the CD.

Discussion

We developed a workflow for unsupervised encoding selection and performance assessment of multiple ensembles and base classifiers. Thus, we implemented and compared several algorithms to facilitate ensemble pruning, including convex hull, Pareto frontier pruning, and multi-verse optimization (MVO). Our results demonstrate that the crucial factors are the base classifiers and the individual encodings. The ensemble technique was not relevant, i.e., we could not observe performance variations using one of hard or soft voting or stacking. In general, applying the Decision Tree (DT) as a base classifier yielded good performance across all datasets. The Pareto frontier pruning selected suitable encodings throughout the experiments.

However, since we used one encoding per base classifier, we restricted the employed ensemble methods, i.e., majority voting, averaging, and stacking, which do

not modify the base classifiers. These ensemble types are in contrast to others, e.g., boosting, where weights are adapted for misclassified training instances in base classifiers [33]. More research is necessary to investigate how performance and more sophisticated ensemble methods are associated. The employed ensemble types are also the reason for the kappa-error point cloud shape solely depending on the base classifiers. Consequently, computing the kappa-error diagram for all ensemble methods was unnecessary. Our encoding/classifier approach is also contrary to other studies, e.g., [12], [14], or [16], which concatenated several encoded datasets to one final dataset (hybrid model) and applied feature selection before training. In the present study, we solely scaled the datasets to standardize the feature range; nevertheless, used the encoded datasets largely unprocessed, potentially affecting the final performance.

As mentioned above, we employed several methods for ensemble pruning comprising best single and random encodings for reference. In general, utilizing the Pareto frontier pruning generates good ensembles; however, requiring the calculation of the Cartesian product of all base classifiers; thus, encodings. Although only the (lower) triangular matrix is necessary, the computation is still CPU-intensive. Furthermore, considering the performance gain compared to the single best encodings, the diversity contribution is only small, but more research is required in this direction [34]. The results of the MVO also acknowledge the impact of diversity. One can observe that the MVO generates inferior ensembles (see Fig. 3).

Regarding Fig. 1, which depicts preferable classifier pairs towards the lower-left corner, one can readily recognize the inferiority of the MVO. The classifier pairs are distributed across the kappa-error area, i.e., the MVO screens the entire solution space and adds weak classifiers to the final ensemble. Nevertheless, since we limited the maximum number of generations to 15, we cannot rule out that more generations would yield better results. Moreover, due to high resource consumption, we limited the MVO to 5 folds, which might hamper comparison.

Moreover, the Random Forest (RF) deployment as a single classifier reveals good performance, which is expected since it is already an ensemble algorithm per se. With this respect, the other base classifiers are less accurate (see Fig. 3). However, it could be demonstrated that RFs as base classifiers, i.e., using different encoded datasets per model, slightly improves the performance. This further highlights the importance of different encodings, hence the projection of different biological aspects, for the classification process.

The implemented methods demonstrate usability on a broad range of datasets from various biomedical domains. With this respect, we incorporated the MVO owing to its excellent and promising performance on several benchmark datasets [35]. The comprehensive Monte Carlo cross-validation copes with the variance, ultimately increasing the robustness of the results. In addition, the Pareto frontier and convex hull pruning consider simultaneously the performance and the diversity of encodings and base classifiers; hence, compensating their strength and weaknesses and revealing their potential not only for ensembles [36], but also in particular for biomedical classification. Our proposed extension to the critical difference chart allows the viewer at one glance to grasp significant, i.e., critical, performance differences of encodings, models, and pruning methods jointly with the actual performance (see Fig. 2).

Conclusions

In summary, we employed two overproduce-and-select methods, namely Pareto frontier and convex hull pruning, as well as the multi-verse optimizer for exhaustively searching the encoding/base classifier space. We employed Logistic Regression, Decision Trees, Naïve Bayes, and Random Forest as base models and majority vote, averaging, and stacked generalization for the fusion. The experiments and visualizations enable the comparison of the respective components; however, further research is necessary to examine other ensemble classifiers, e.g., boosting. All in all, we propose an extensible workflow for automated encoding selection through diverse ensemble pruning methods. Researchers can utilize our workflow to augment the recently published PEPTIDE REACToR [20] with an unsupervised encoding selection, ultimately easing the access for non-technical users.

Methods

We developed a high-throughput workflow using Snakemake v6.5.1 [37], Python v3.9.1, and R v4.1.0. For the machine learning algorithms, we employed scikit-learn v0.24.2 [38]. The peptide datasets are taken from the PEPTIDE REACToR [20]. Finally, only encoded datasets with the final sequence- and structure-based encodings were used for the subsequent analyses.

Note that there are two approaches to harness multiple encodings in a single model, namely the fusion and the hybrid model [21]. Fusion models train one encoding per base classifier and fuse the output for the final prediction. Contrary, hybrid models use the concatenated features of multiple encodings for single model training. The concatenation approach is particularly problematic for entropy-based models such as DT or RF due to the bias in variable selection. Thus, in the present study, we implemented the fusion design, i.e., each ensemble consists of an arbitrary amount of base classifiers using one particular encoding, respectively. Finally, the employed datasets from a wide range of biomedical domains ensure broad applicability and the robustness of our results.

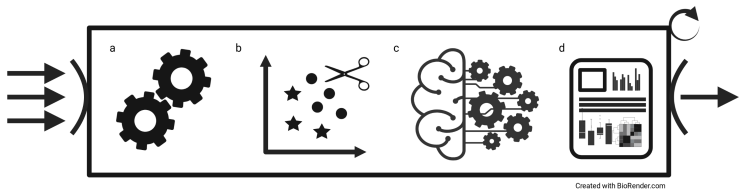


Figure 4 Overview of the workflow. (**a**) For each fold of the MCCV, the preprocessing is conducted, i.e., the indices of the train/test splits are determined, and the data is scaled. (**b**) The pruning methods, e.g., Pareto frontier and MVO, select the current fold's encodings and the number of base classifiers. (**c**) Different ensembles with various base classifiers are trained and validated on the test data. (**d**) The results are collected, statistically validated, and illustrated. The workflow accepts an arbitrary number of datasets as input (arrows). Refer to the method section for more details.

The workflow conducts the following steps. First, indices are determined to ensure equal samples for the comprehensive cross-validation, and the indices for all folds are calculated. Second, we standardized the encoded datasets using a min-max

Table 4 Employed datasets in this study. The function refers to the positive class, i.e., sequences of class + possess the respective function. The stated MCC refers to the performance reported in the original study. See the references or [20] for more details.

Name	Function	MCC	Size (+,-)	Ref.
acp_mlacp	Anti-cancer	0.698	581 (185,396)	[12]
aip_antiinflam	Anti-inflammatory	0.45	2124 (863,1261)	[14]
amp_antibp2	Anti-microbial	0.84	1975 (981,994)	[40]
atb_antitbp	Anti-tubercular	0.52	492 (246,246)	[41]
avp_amppred	Anti-viral	0.8	1476 (738,738)	[16]
cpp_mlcpp	Cell-penetrating	0.793	1901 (737,1164)	[13]
hem_hemopi	Hemolytic	0.52	1013 (522,461)	[17]
isp_il10pred	Immunosuppressive	0.59	1242 (394,848)	[42]
nep_neuropipred	Neuropeptides	0.67	1750 (875,875)	[18]
pip_pipel	Pro-inflammatory	0.454	3228 (833,2395)	[15]

normalization between 0 and 1. Afterward, we trained and assessed models for all encoded datasets and ensemble types using a 100-fold Monte Carlo cross-validation. We selected the best single and the random best encoding per dataset to compare the results to single encodings. Finally, we statistically assessed and visualized the results (see Fig. 4). Significant steps are described in more detail below. We will use the following definitions throughout the manuscript: the original unprocessed dataset is denoted as the dataset. One dataset can be encoded in manifold ways, which we refer to as encoded datasets. Encodings specify particular encoding algorithms.

Note that we used Matthews correlation coefficient (MCC) throughout the study to handle the imbalance in the datasets [39]:

$$\mathrm{MCC} = \frac{a \times d - c \times b}{\sqrt{(a+c)(a+b)(d+c)(d+b)}}. \tag{1}$$

a is the number of true positives, d is the number of true negatives, b is the number of false negatives, and c is the number of false positives.

Datasets

For a comprehensive analysis on peptide encodings, Spänig *et al.* (2021) gathered a variety of datasets from multiple biomedical domains [20]. We specifically selected datasets with low to medium classification performance from this collection, i.e., a reported MCC of 0.63 ± 0.15 on the independent test set; additionally, covering diverse biomedical applications. Moreover, we excluded datasets for which accurate models have been published to investigate the potential effects of different classifiers and ensembles. We limited our study to ten datasets to cope with the computational complexity. The dataset size ranges from 492 to 3,228 sequences with an average of $1,580.8 \pm 812.1$ sequences. The datasets comprise 15,782 sequences with a mean length of 21.17 ± 13.23 amino acids. 6,404 sequences belong to the positive and 9,378 to the negative class. The average sequence length is 22.47 ± 15.88 and 20.29 ± 10.97, respectively. Duplicated sequences have been removed. Refer to Table 4 for more details.

Monte Carlo cross-validation

We applied the Monte Carlo cross-validation (MCCV) [43]. The MCCV improves the generalization and diminishes the variance of the results, i.e., results are more robust, hence comparable. In addition, we ensured that the n-th fold is identical across all experiments leading to improved comparability across all base classifiers and ensembles. Each fold is composed of one split using 80 % of the data for model training and another utilizing the remaining 20 % for testing. In contrast to k-fold cross-validation, MCCV follows a sampling with replacement strategy, i.e., splits can contain identical samples multiple times. However, duplicate samples do not occur in the train, and the test split [43].

Base classifiers

We used the following base classifiers for our experiments: Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. Each classifier will be briefly described hereinafter. We used the implementations provided by the scikit-learn library [38].

Naïve Bayes

The Naïve Bayes (NB) classifier (naively) assumes conditional independence of the feature vectors and applies the Bayes theorem for prediction [25]. Model training is enabled via a probability density function (PDF) and the prior probability of a given class. For simplicity, we assume a Gaussian distribution of the features. Hence, we applied the Gaussian NB using

$$p(x|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2}$$

as the PDF, whereby σ denotes the standard deviation and μ the mean of features x given a class y [44].

Logistic Regression

The binary Logistic Regression (LR) is another probability-based classifier, i.e., it derives the probability of a class y given a feature vector x [45]. The LR predicts probabilities between 0 and 1 using the logistic function denoted as

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3}$$

and the maximum likelihood function to estimate the coefficients β, i.e., to train the model [45].

Decision Tree

The Decision Tree (DT) classifier, precisely the CART (Classification And Regression Trees) implementation, is a tree-based model, i.e., a tree structure is generated during training [46]. Each node is based on the most discriminating feature [25]. New splits are created based on the impurity of the remaining data, i.e., if a split is pure enough, a leaf node is added. Otherwise, intermediate nodes are created [25].

For prediction, the tree is traced until a leaf node, which states the final class. In particular, we used the Gini impurity, denoted as

$$i(t) = 1 - \sum_j P_j^2, \tag{4}$$

where $j \in \{0, 1\}$ for binary classification and P is the probability of class j at a node t [25].

Random Forest
The Random Forest (RF) classifier is an ensemble learning technique, which trains multiple DTs on random samples, i.e., bagging, of the input data [47]. For the final classification, the majority vote of the trees is used [47]. Note that we use the RF as a base learner, which allows comparing the performance with DTs and the actual ensembles techniques in general (see below).

Classifier ensembles

To combine the output specifically of the base classifiers introduced above, we employed the following ensemble methods: majority vote (hard voting), averaging (soft voting), and stacked generalization (stacking). In the present study, each base classifier is trained on one encoded dataset, meaning if for one dataset n encodings are selected, the size of one ensemble is n. We adapted the implementations of the scikit-learn library [38], such that not only one dataset but several encoded datasets can be used for training. For instance, if one passes n encoded datasets, the ensemble consists of n base classifiers trained on one particular encoded dataset, respectively.

Majority voting The majority voting ensemble (hard voting) combines the output by ultimately assigning the class, which has been predicted by the majority of the single base classifiers. We employed the customized version of scikit-learn's VotingClassifier class with hard voting enabled.

Averaging The averaging method (soft voting) computes the means of the predicted class probabilities per base classifier. The maximum value determines the final class. We used the adjusted VotingClassifier with voting set to soft.

Stacked generalization The stacking approach utilizes the output of the base classifiers to train a meta-model, i.e., the predicted class probabilities of the base classifiers are used as features [48]. We adapted the StackingClassifier from the scikit-learn package and employed Logistic Regression as the meta-model.

Ensemble pruning

Selecting the correct number of base classifiers in an ensemble is challenging. Thus, Kuncheva (2014) suggests several approaches to determine the ensemble size [25]. For instance, sequential forward selection, adding one classifier successively, in case the additional model improves the ensemble performance [25]. However, in the

present case, we are dealing with potentially hundreds of encoded datasets, for which this particular technique is not practical. To this end, we used two selection methods, namely convex hull and Pareto frontier pruning, circumventing the limitations mentioned above [25].

Moreover, we implemented the multi-verse optimization algorithm as an automatic encoding selection technique [49]. Finally, we employed best and random encodings selection as a baseline reference. The pruning methods are described more precisely in the following.

Kappa-error diagram

The kappa-error diagram, introduced by Margineantu and Dietterich (1997), is the basis for the convex hull and Pareto frontier pruning [50]. The graph represents pairs of classifiers by their average error and diversity, as shown in Fig. 1. The diversity measures the agreement of classifier outputs, i.e., the better the agreement of the classifier predictions, the less the diversity [25]. Specifically, the kappa diversity is denoted as

$$\kappa = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}. \tag{5}$$

The κ statistic ranges from -1 to 1, whereby $\kappa = 1$ denotes perfect agreement, $\kappa = 0$ random, and $\kappa < 1$ worse than random consensus [50]. The error is calculated using

$$e = 1 - \frac{a+d}{a+b+c+d}, \tag{6}$$

with the subtrahend being the accuracy. However, Kuncheva (2013) pointed out that diversity concerning the average error can not be arbitrarily low [36]. In fact, desirable classifier pairs approximate the lower-left corner (see Fig. 1), i.e., approximating a theoretical boundary, which is defined in Eq. 7 [36].

$$\kappa_{min} = \begin{cases} 1 - \frac{1}{1-e}, & \text{if } 0 < e \leq 0.5 \\ 1 - \frac{1}{e}, & \text{if } 0.5 < e < 1 \end{cases} \tag{7}$$

Note that the classifier pairs are composed using the lower triangular matrix of the Cartesian product. Afterward, the pruning methods select a subset of pairs, also likely include duplicated base classifiers. Thus, all pruning methods ensure that the final ensemble only uses unique classifiers. Hence, base classifiers are trained on individual encoded datasets.

Convex hull

The kappa-error diagram depicts a set of points, i.e., pairs of base classifiers, in a two-dimensional space. The kappa diversity is the first, and the pairwise average

error is the second dimension. We employed the Quickhull algorithm to calculate the convex hull [51]. Hence, the smallest convex set that contains the classifier pairs [51]. Thus, no further classifier pairs exist beyond the convex hull. We utilized the implementation of the Quickhull algorithm provided by the SciPy package in the ConvexHull module [52].

Since we are only interested in the partial convex hull, that is, pairs approaching the theoretical boundary defined in Eq. 7 and depicted in Fig. 1, we adapted the *pareto_n* algorithm from Kuncheva (2014), which returns only classifier pairs fulfilling the criteria [25].

Pareto frontier

The Pareto optimality describes the compromise of multiple properties towards optimizing a single objective [53]. For instance, a pair of classifiers is Pareto optimal if improving the diversity is impossible without simultaneously impairing the average pairwise error. Analog to the partial convex hull introduced earlier, Pareto optimal classifier pairs approach the theoretical boundary as stated in Eq. 7, ultimately defining the Pareto frontier. Again, we used the *pareto_n* algorithm adapted from Kuncheva (2014) to obtain all classifier pairs determining the Pareto frontier (see Fig. 1).

Multi-verse optimization

The multi-verse optimization (MVO) algorithm is inspired by the alternative cosmological model stating that several big bangs created multiple, parallel existing universes, which are connected by black and white holes and wormholes [35]. In terms of an optimization algorithm, black and white holes are used to explore the search space and wormholes to refine solutions [35]. Moreover, the inflation rate, i.e., the fitness, of universes is used for the emergence of new holes; thus, to cope with local minima [49]. For more details, refer to Mirjalili *et al.* (2016) and Al-Madi *et al.* (2019) [35, 49]. We implemented the binary MVO following [49] using Python. Each solution candidate is represented as a binary vector, where each position denotes the path to an encoded dataset, that is, the i-th bit set means that the i-th encoding is included in the final ensemble (see Fig. 1). We examined different generations, i.e., 100, 80, 50, 25, and 15. However, we observed that performance depends mainly on the initialization and count of the universes. Specifically, the performance gain from the 15th generation is minor but requires much time. Thus, we set the optimization to a maximum of 15 generations with 32 universes each. Due to its resource intensity, we executed the MVO only for the first five folds (see section Monte Carlo cross-validation).

Best encodings

A further pruning method uses only the best classifier pairs. In particular, based on the kappa-error diagram, the algorithm selects 15 classifier pairs with the lowest pairwise average error (see Fig. 1).

Random encodings

The last pruning method selects 15 random classifier pairs from the kappa-error diagram. Note that the selection is only performed one time. That is, the pairs are the same across all folds.

Statistics

We examined the areas covered by the respective base classifiers (see Fig. 1). To this end, we calculated the area for each fold. The area is described by multiple variables, i.e., the kappa diversity and the average pairwise error. Thus, we applied the multivariate analysis of variance (MANOVA) to verify if the areas differ significantly. If this is the case, we subsequently employed an analysis of variance (ANOVA) to investigate the effect of the diversity and the average error separated. For post-hoc assessment, Tukey's HSD has been applied. We used the tests provided by the R standard library. α was set to 0.05, i.e., p values ≤ 0.05 are considered as significant.

In addition, we employed the Friedman test with the Iman and Davenport correction for the statistical comparison of multiple single and ensemble classifiers [54]. In the case at least one model is significantly different, we used the Nemenyi test for post-hoc analysis [54]. Refer also to Spänig *et al.* (2021) for more details [20]. The tests were provided by the scmamp R package v0.2.55 [32].

Finally, we examined if the best ensemble has a significant improvement over the best single classifier using Student's t-test for repeated measures, i.e., paired samples. Again, α was defined as 0.05.

Data visualization

All plots are realized using Altair v4.1.0 [55] and described in more detail hereinafter.

Kappa-error diagram

The kappa-error diagram, suggested by Margineantu and Dietterich (1997) [50], shows the result of a single split in the top row and a two-dimensional histogram aggregating all folds in the bottom row (see Fig. 1). The columns show the base classifiers. Note that the kappa-error shape depends only on the base classifiers (see Discussion). The top row also visualizes the partial convex hull (black line) and the Pareto frontier (red line). Symbols refer to the pruning method. Each dot is a classifier pair trained on two encoded datasets. Note that we display only 1000 dots per panel (top row). Moreover, we set the bin size to 40 for the binned heatmap with darker colors depicting more values (bottom row).

XCD chart

The extended critical difference (XCD) chart (Fig. 2) is based on the critical difference chart introduced by Calvo and Santafé (2016) [32]. Classifier groups not surpassing the critical difference (CD) are connected with black lines. The line thickness depicts the actual CD, meaning groups associated with thicker lines are closer to CD. The XCD charts present two classifier groups. The x-axis includes pruning types, and the y-axis the actual ensembles and the corresponding base classifier. The main area contains a categorical heatmap showing Matthews correlation coefficient (MCC) in 0.05 steps. The darker, the higher the MCC. The MCC is the median MCC of the respective group combination and corresponds to the median from Fig. 3. Note that for the computation of the CD, we concatenated the MCCs of all cross-validation runs, e.g., 12 * 100 MCCs for pfront, and 6 * 100 MCCs for bayes_voting_soft.

Funding
This work was financially supported by the BMWi in the project MoDiPro-ISOB (16KN0742325). This work was also supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Availability of data and materials
The source code can be found at https://github.com/spaenigs/ensemble-performance. All datasets are available at https://github.com/spaenigs/peptidereactor.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
SS and DH developed the concept. SS designed and performed the experiments and analyzed the data. SS and DH interpreted the results. AM implemented the MVO algorithm. SS wrote the manuscript. DH supervised the study and revised the manuscript. All authors read and approved the final manuscript.

Author details
Data Science in Biomedicine, Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany.

References
1. Holmes, A.H., Moore, L.S.P., Sundsfjord, A., Steinbakk, M., Regmi, S., Karkey, A., Guerin, P.J., Piddock, L.J.V.: Understanding the mechanisms and drivers of antimicrobial resistance. The Lancet **387**(10014), 176–187 (2016). doi:10.1016/S0140-6736(15)00473-0
2. Spänig, S., Eick, L., Nuy, J.K., Beisser, D., Ip, M., Heider, D., Boenigk, J.: A multi-omics study on quantifying antimicrobial resistance in European freshwater lakes. Environment International **157**, 106821 (2021). doi:10.1016/j.envint.2021.106821
3. Kakkar, M., Walia, K., Vong, S., Chatterjee, P., Sharma, A.: Antibiotic resistance and its containment in India. BMJ (Online) **358**, 25–30 (2017). doi:10.1136/bmj.j2687
4. Qu, J., Huang, Y., Lv, X.: Crisis of antimicrobial resistance in China: Now and the future. Frontiers in Microbiology **10**(SEP) (2019). doi:10.3389/fmicb.2019.02240
5. Lazzaro, B.P., Zasloff, M., Rolff, J.: Antimicrobial peptides: Application informed by evolution. Science **368**(6490) (2020). doi:10.1126/science.aau5480
6. Magana, M., Pushpanathan, M., Santos, A.L., Leanse, L., Fernandez, M., Ioannidis, A., Giulianotti, M.A., Apidianakis, Y., Bradfute, S., Ferguson, A.L., Cherkasov, A., Seleem, M.N., Pinilla, C., de la Fuente-Nunez, C., Lazaridis, T., Dai, T., Houghten, R.A., Hancock, R.E.W., Tegos, G.P.: The value of antimicrobial peptides in the age of resistance. The Lancet Infectious Diseases **20**(9), 216–230 (2020). doi:10.1016/S1473-3099(20)30327-3
7. Waghu, F.H., Idicula-Thomas, S.: Collection of antimicrobial peptides database and its derivatives: Applications and beyond. Protein Science **29**(1), 36–42 (2020). doi:10.1002/pro.3714
8. Chung, C.R., Kuo, T.R., Wu, L.C., Lee, T.Y., Horng, J.T.: Characterization and identification of antimicrobial peptides with different functional activities. Briefings in Bioinformatics **21**(3), 1098–1114 (2020). doi:10.1093/bib/bbz043
9. Dean, S.N., Walper, S.A.: Variational autoencoder for generation of antimicrobial peptides. ACS Omega **5**(33), 20746–20754 (2020). doi:10.1021/acsomega.0c00442
10. Fingerhut, L.C.H.W., Miller, D.J., Strugnell, J.M., Daly, N.L., Cooke, I.R.: ampir: an R package for fast genome-wide prediction of antimicrobial peptides. Bioinformatics **36**(21), 5262–5263 (2020). doi:10.1093/bioinformatics/btaa653/5873588
11. Aronica, P.G.A., Reid, L.M., Desai, N., Li, J., Fox, S.J., Yadahalli, S., Essex, J.W., Verma, C.S.: Computational Methods and Tools in Antimicrobial Peptide Research. Journal of Chemical Information and Modeling, 1–00175 (2021). doi:10.1021/acs.jcim.1c00175
12. Manavalan, B., Basith, S., Shin, T.H., Choi, S., Kim, M.O., Lee, G.: MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget **8**(44), 77121–77136 (2017)
13. Manavalan, B., Subramaniyam, S., Shin, T.H., Kim, M.O., Lee, G.: Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. Journal of Proteome Research **17**(8), 2715–2726 (2018). doi:10.1021/acs.jproteome.8b00148
14. Gupta, S., Sharma, A.K., Shastri, V., Madhu, M.K., Sharma, V.K.: Prediction of anti-inflammatory proteins/peptides: An insilico approach. Journal of Translational Medicine **15**(1) (2017). doi:10.1186/s12967-016-1103-6
15. Manavalan, B., Shin, T.H., Kim, M.O., Lee, G.: PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. Frontiers in immunology **9**, 1783 (2018). doi:10.3389/fimmu.2018.01783
16. Meher, P.K., Sahu, T.K., Saini, V., Rao, A.R.: Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Scientific Reports **7** (2017). doi:10.1038/srep42362

17. Chaudhary, K., Kumar, R., Singh, S., Tuknait, A., Gautam, A., Mathur, D., Anand, P., Varshney, G.C., Raghava, G.P.S.: A web server and mobile app for computing hemolytic potency of peptides. Scientific Reports **6** (2016). doi:10.1038/srep22843
18. Agrawal, P., Kumar, S., Singh, A., Raghava, G.P.S., Singh, I.K.: NeuroPIpred: a tool to predict, design and scan insect neuropeptides. Scientific Reports **9**(1) (2019). doi:10.1038/s41598-019-41538-x
19. Spänig, S., Heider, D.: Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. BioData Mining **12**(1), 1–29 (2019). doi:10.1186/s13040-019-0196-x
20. Spänig, S., Mohsen, S., Hattab, G., Hauschild, A.-C., Heider, D.: A large-scale comparative study on peptide encodings for biomedical classification. NAR Genomics and Bioinformatics **3**(2) (2021). doi:10.1093/nargab/lqab039
21. Khatun, M.S., Hasan, M.M., Shoombuatong, W., Kurata, H.: ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. Journal of Computer-Aided Molecular Design **34**(12), 1229–1236 (2020). doi:10.1007/s10822-020-00343-9
22. Plisson, F., Ramírez-Sánchez, O., Martínez-Hernández, C.: Machine learning-guided discovery and design of non-hemolytic peptides. Scientific Reports **10**(1) (2020). doi:10.1038/s41598-020-73644-6
23. Timmons, P.B., Hewage, C.M.: HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. Scientific Reports **10**(1) (2020). doi:10.1038/s41598-020-67701-3
24. Singh, O., Hsu, W.-L., Su, E.C.-Y.: Co-AMPpred for in silico-aided predictions of antimicrobial peptides by integrating composition-based features. BMC Bioinformatics **22**(1), 389 (2021). doi:10.1186/s12859-021-04305-2
25. Kuncheva, L.I.: Combining Pattern Classifiers, (2014). doi:10.1002/9781118914564
26. Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R.J., Webb, G.I., Zhao, Q., Kurgan, L., Song, J.: iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Research (2021). doi:10.1093/nar/gkab122
27. Schwarz, J., Heider, D.: GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making. Bioinformatics **35**(14), 2458–2465 (2019). doi:10.1093/bioinformatics/bty984
28. Heider, D., Dybowski, J.N., Wilms, C., Hoffmann, D.: A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Mining **7**(1) (2014)
29. Löchel, H.F., Riemenschneider, M., Frishman, D., Heider, D.: SCOTCH: subtype a coreceptor tropism classification in HIV-1. Bioinformatics **34**(15), 2575–2580 (2018)
30. Kuncheva, L.I., Jain, L.C.: Designing classifier fusion systems by genetic algorithms. In: IEEE Transactions on Evolutionary Computation, vol. 4 (2000)
31. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, A.F.T.S.: Data mining in the life science swith random forest: A walk in the park or lost in the jungle? Briefings in Bioinformatics **14**, 315–326 (2013). doi:10.1093/bib/bbs034
32. Calvo, B., Santafé, G.: scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems. The R Journal **8**(1), 248–256 (2016)
33. Zhu, J., Zou, H., Rosset, S., Hastie, T.: Multi-class AdaBoost. Statistics and Its Interface **2**, 349–360 (2009)
34. Kuncheva, L.I.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning **51**, 181–207 (2003)
35. Mirjalili, S., Mirjalili, S.M., Hatamlou, A.: Multi-Verse Optimizer: a nature-inspired algorithm for global optimization. Neural Computing and Applications **27**(2), 495–513 (2016). doi:10.1007/s00521-015-1870-7
36. Kuncheva, L.I.: A bound on kappa-error diagrams for analysis of classifier ensembles. IEEE Transactions on Knowledge and Data Engineering **25**(3), 494–501 (2013). doi:10.1109/TKDE.2011.234
37. Köster, J., Rahmann, S.: Snakemake-a scalable bioinformatics workflow engine. Bioinformatics **28**(19), 2520–2522 (2012). doi:10.1093/bioinformatics/bts480
38. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passps, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2015). doi:10.1145/2786984.2786995
39. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1) (2020). doi:10.1186/s12864-019-6413-7
40. Su, X., Xu, J., Yin, Y., Quan, X., Zhang, H.: Antimicrobial peptide identification using multi-scale convolutional network. BMC Bioinformatics **20**(1) (2019). doi:10.1186/s12859-019-3327-y
41. Usmani, S.S., Bhalla, S., Raghava, G.P.S.: Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. Frontiers in Pharmacology **9**(AUG), 1–11 (2018). doi:10.3389/fphar.2018.00954
42. Nagpal, G., Usmani, S.S., Dhanda, S.K., Kaur, H., Singh, S., Sharma, M., Raghava, G.P.S.: Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. Scientific Reports **7** (2017). doi:10.1038/srep42851
43. Xu, Q.-S., Liang, Y.-Z.: Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems **56**, 1–11 (2000)
44. Ren, J., Lee, S.D., Chen, X., Kao, B., Cheng, R., Cheung, D.: Naive bayes classification of uncertain data. In: Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 944–949 (2009). doi:10.1109/ICDM.2009.90
45. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, 8th edn. Springer, New York (2017)
46. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification And Regression Trees. Routledge, London (1984). doi:10.1201/9781315139470
47. Breiman, L.: Random Forests. Machine Learning **45**, 5–32 (2001)
48. Wolpert, D.H.: Stacked Generalization. Neural Networks **5**, 241–259 (1992)
49. Al-Madi, N., Faris, H., Mirjalili, S.: Binary multi-verse optimization algorithm for global optimization and

discrete problems. International Journal of Machine Learning and Cybernetics **10**(12), 3445–3465 (2019). doi:10.1007/s13042-019-00931-8

50. Margineantu, D.D., Dietterich, T.G.: Pruning Adaptive Boosting. In: ICML (1997)
51. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The Quickhull Algorithm for Convex Hull. ACM Transactions on Mathematical Software **22**(4), 469–483 (1996)
52. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). doi:10.1038/s41592-019-0686-2
53. Messac, A., Ismail-Yahaya, A., Mattson, C.A.: The normalized normal constraint method for generating the Pareto frontier. Structural and Multidisciplinary Optimization **25**(2), 86–98 (2003). doi:10.1007/s00158-002-0276-1
54. Santafe, G., Inza, I., Lozano, J.A.: Dealing with the evaluation of supervised classification algorithms. Artificial Intelligence Review **44**(4), 467–508 (2015). doi:10.1007/s10462-015-9433-y
55. VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., Sievert, S.: Altair: Interactive Statistical Visualizations for Python. Journal of Open Source Software **3**(32), 1057 (2018). doi:10.21105/joss.01057

6.5 Multivalent Binding Kinetics Resolved by Fluorescence Proximity Sensing

The interaction of proteins with proteins, peptides, or molecules is key for many biological processes[61]. Specifically, protein-protein interaction (PPI) is critical in immune response since antibodies interact with epitopes, hence, a unique amino acid pattern on the antigen[15]. Molecules with strong binding potential with proteins are denoted as ligands and are defined by mutual biophysical attraction[61]. According to Du *et al.* (2016), three PPI mechanisms are known, namely “lock-and-key”, “induced fit”, and “conformational selection”[61]. However, computational prediction remains challenging since PPI also requires well-considered negative and positive training examples[91]. Moreover, low binding affinity aggravates reliable predictions[48]. In this light, Cunningham *et al.* (2020) developed an ML pipeline to predict low-affinity protein-peptide interactions in signaling cascades[48]. According to the authors, such a framework enables computational PPI studies on a large scale[48]. Various further methods have been developed for computational PPI prediction[143].

Experimentally groundwork; thus, generating sufficient large datasets is crucial for ML and vice versa. High-throughput procedures allow mutual reinforcement. Artificial intelligence benefits from experimental data and laboratory protocols leverage computational approaches[139]. In this light, Xue *et al.* (2017) stressed ML for improving the design of peptide arrays, which are utilized to screen the activity of dozens of peptides in parallel[256]. The authors applied a Random Forest regressor (RFR) to predict the signal-to-noise ratio of mass spectrometry data[256]. Based on the results, the authors could ultimately improve peptide array configuration[256].

In the present study, we conducted fluorescence proximity sensing (FPS) to quantify the binding affinities of different peptide architectures experimentally. Since FPS indicates protein-ligand interaction using a fluorescence marker bound to the solid phase, modification of the reactants is obsolete. We employed the results as input for an ML model to predict the respective binding rates.

In particular, the binding affinity of three architectures has been examined, including dimeric, tetrameric, and octameric peptides with varying amino acid compositions. The results underpinned the potential of FPS as a crucial technology for high-throughput affinity measuring. We leveraged the PEPTIDE REACTOR to compare various encodings; thus, addressing the ML representation of the modified peptides[223]. We demonstrated that using the best encoding and a binary description of the architecture leads to efficient models. Specifically, the evaluation revealed good accordance between the experimental and predicted binding kinetics, stressing the future potential of computer-aided peptide design.

Multivalent Binding Resolved by Fluorescence Proximity Sensing

Clemens Schulte[1]; Alice Soldà[2]; Sebastian Spänig[3], Nathan Adams[4], Ivana Bekić[4], Werner Streicher[4]; Dominik Heider[3], Ralf Strasser[2]; Hans Michael Maric[1]*
*Correspondence to Hans.Maric@uni-wuerzburg.de

[1]Rudolf Virchow Center; Center for Integrative and Translational Bioimaging; University of Wuerzburg; Josef-Schneider-Str. 2, Germany, 97080 Wuerzburg, Germany

[2]Dynamic Biosensors GmbH Germany, Lochhamer Strasse 15, 82152 Martinsried/Planegg, Germany

[3]Department of Bioinformatics, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany

[4]Nanotemper Technologies GmbH, Flößergasse 4, 81369 Munich, Germany

Summary

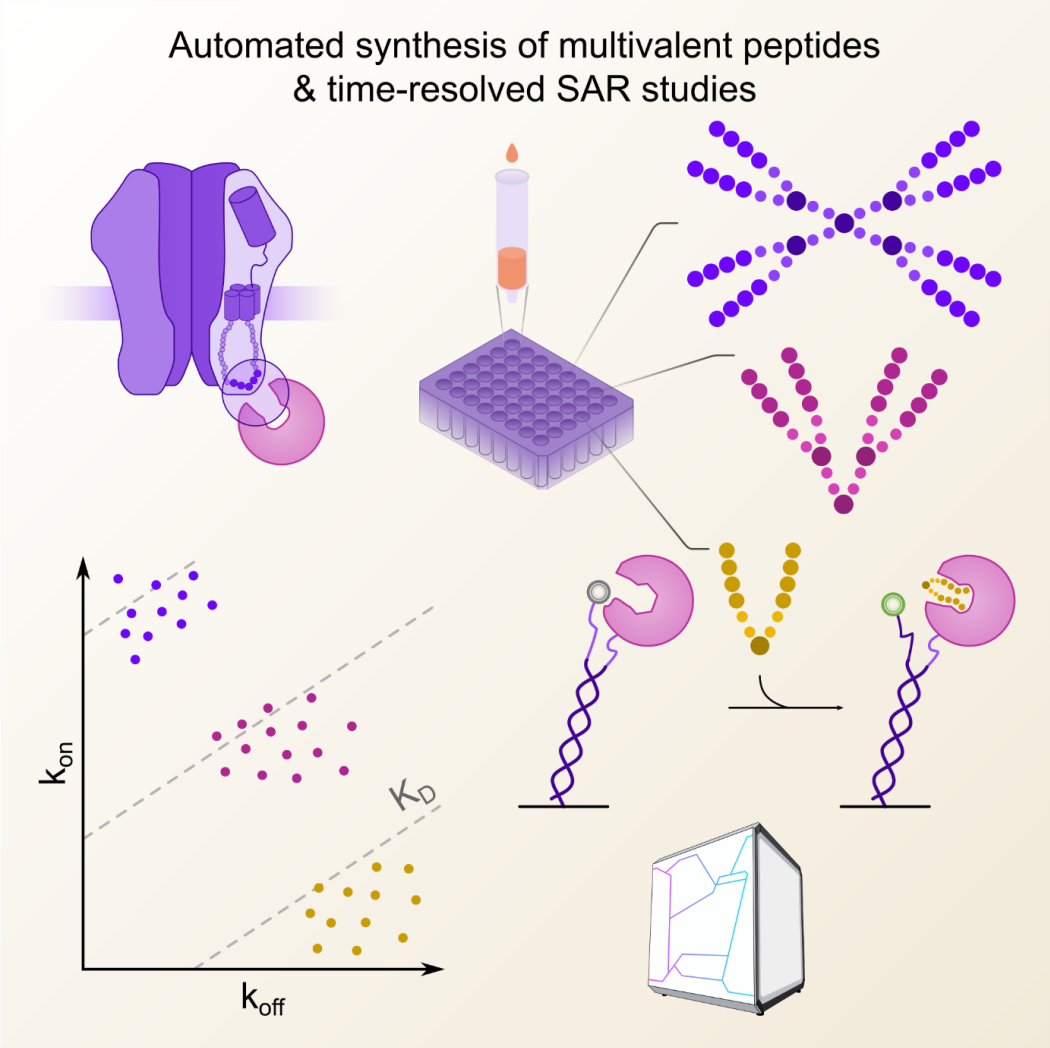
Multivalent protein interactors are an attractive modality for probing protein function and exploring novel pharmaceutical strategies. The throughput and precision of state-of-the-art methodologies and workflows for the effective development of multivalent binders is currently limited by surface immobilization, fluorescent labelling and sample consumption.

Using the gephyrin protein, the master regulator of the inhibitory synapse, as benchmark, we exemplify the application of Fluorescence proximity sensing (FPS) for the systematic kinetic and thermodynamic optimization of multivalent peptide architectures. High throughput synthesis of +100 peptides with varying combinatorial dimeric, tetrameric, and octameric architectures combined with direct FPS measurements resolved on-rates, off-rates, and dissociation constants with high accuracy and low sample consumption compared to three complementary technologies. The dataset and its machine learning-based analysis deciphered the relationship of specific architectural features and binding kinetics and thereby identified binders with unprecedented protein inhibition capacity thus, highlighting the value of FPS for the rational engineering of multivalent inhibitors.

Keywords (10/10)

Protein-Protein Interaction, High-throughput, Kinetics, Peptide, TRIC, ITC, BLI, FPS, SwitchSense, Avidity

Graphical Abstract



Introduction

Protein-protein interactions (PPIs) are of fundamental importance for cellular function and dysfunction (Pawson and Nash, 2003) with up to 40% of all PPIs involving short, linear motifs located in intrinsically disordered protein regions (London et al., 2013). Targeting and probing such PPIs contributes significantly to our understanding of physiology, pathology and ultimately the identification of novel pharmacological strategies (Lu et al., 2020). The development of affine and selective PPI modulators is facilitated by biophysical technologies that enable the determination of binding parameters of large binder libraries with minimal sample requirements. In particular, multimeric or branched peptides (Brunetti et al., 2018, Lee et al., 2005, Demmer et al., 2011) provide superior binding specificities and affinities due to avidity (Erlendsson and Teilum, 2021, Errington et al., 2019) by exploiting protein homo-oligomerization (Kitov and Bundle, 2003) observed for more than half of all proteins (Marianayagam et al., 2004), offering enormous potential for the design of multivalent drugs, including novel drug modalities such as trivalent PROTACs (Imaide et al., 2021), They commonly exhibit slower off-rates, and thus enhanced residence times, compared to their monovalent counterparts (Wooldridge et al., 2009). Despite the availability of robust theoretical mechanistic frameworks (Errington et al., 2019), the accurate prediction of multivalent binding dynamics based on biophysical properties of the interactors alone remains challenging. This is especially true for systems where higher valencies and complex, heterogeneous topologies occur or where structural information is incomplete. *Vice versa*, systematic experimental structure-activity relationship studies remain scarce, mainly due to laborious workflows, commonly relying on sequential synthesis, multimerization and labelling, necessitating multiple re-purification steps of the often comparably large compounds.

The engineering of multivalent architectures benefits from kinetic methodologies, such as surface plasmon resonance (SPR) or biolayer interferometry (BLI) (Patching, 2014, Sultana and Lee, 2015, Walport et al., 2021). However, such surface-based techniques are vulnerable to artefacts that result from the comparably high affinities and slow off-rates of multivalent binders, causing re-binding and, depending on immobilisation density, interference from neighbouring proteins through crosslinking (Errington et al., 2019).

Here we demonstrate the use of Fluorescence Proximity Sensing (FPS) as an alternative approach to study multivalent peptide-protein interactions in high-throughput (HT) and its value for effectively decoding higher order multivalent structure-activity relationships and thereby facilitating the guided engineering of such interactions.

Results and Discussion

FPS detects the binding of molecules in real-time through changes in the dye's local environment (Häußermann et al., 2019). FPS, based on SwitchSENSE technology, relies on a biochip with covalently attached single stranded anchor DNA for target protein immobilization at a distance of approximately 30 nm (Knezevic et al., 2012), thereby potentially precluding re-binding and avidity effects. The peptide (analyte) binding is reported by a fluorescent reporter close to the immobilized protein of interest (ligand) (Figure 1 A) and consequently independent of unspecific binding of the analyte to the chip surface. Importantly, FPS neither requires direct fluorescent labelling of the ligand nor the analytes, thereby avoiding disturbance of their functional integrity or other dye-mediated artefacts. In contrast to other recently reported kinetic methods (Stein et al., 2021), FPS allows for the analysis of slow ($<10^{-4}$ s^{-1}) off-rates and fast on-rates ($>10^{6}$ $M^{-1}s^{-1}$).

While the workflow is designed to be applicable to any multivalent system where combinatorial display is feasible, we here use the neuronal scaffolding protein gephyrin (Tyagarajan and Fritschy, 2014) (geph) and its structurally resolved (Maric et al., 2014) interactor, the glycine receptor (GlyR) β subunit. PPIs within receptor protein complexes (Rosenbaum et al., 2020) and specifically scaffolds of the neuronal synapses are explored with multivalent chemical probes (Maric et al., 2017, Sainlos et al., 2009, Bach et al., 2012) and studied in pharmacological context (ClinicalTrials.gov, NCT04689035) (Schulte and Maric, 2021). Dimeric, tetrameric and octameric binders were synthesized using an accessible and broadly applicable strategy by combining binding sequences with Polyethylene glycol (PEG) linkers and L-Lysine cores (Nomizu et al., 1993) as branching points (Figure 1 A). Using varied geph binding sequences, PEG linkers of variable length and up to three branching points, we synthesized a total of +100 unique multimeric compounds (Supplementary Table 1), differing over one magnitude in molecular weight.

For the FPS measurements, the otherwise unlabelled receptor binding geph E-domain (gephE) was coupled to the ligand strand while a fluorescent reporter was attached to the dye strand (Figure 1 B). Among six tested dyes, the fluorescence change was highest for the green dye B (Dynamic Biosensors GmbH, DE) (Supplementary Figure 1) which was therefore used in all subsequent FPS measurements. The functionality of this setup was demonstrated by recapitulating the structurally resolved geph-binding site of GlyRβ (^{398}FSIVG^{402}) (Maric et al., 2014) using a 1 µM library of unmodified, overlapping dimeric peptides with an offset of one amino acid (Figure 1 C).

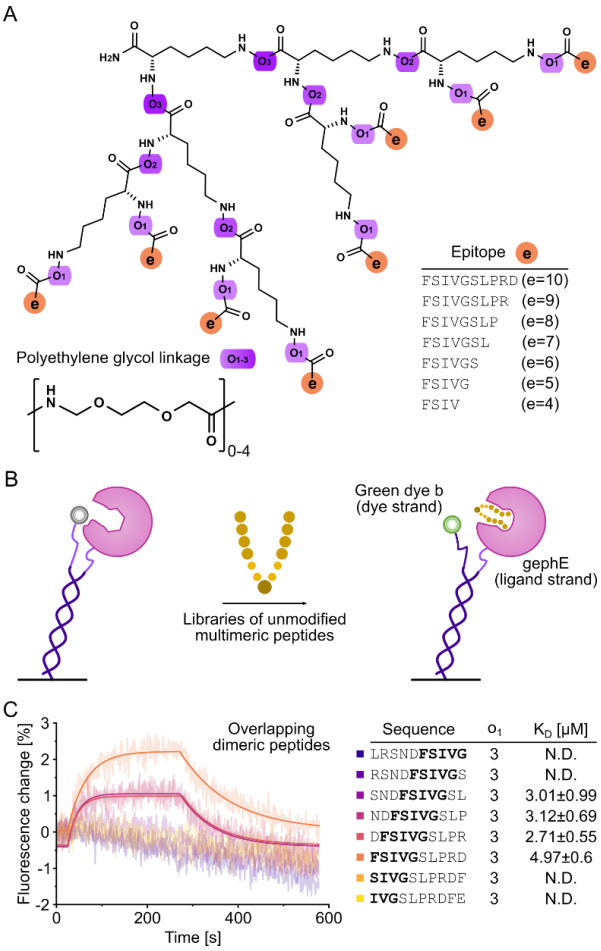


Figure 1: Multivalent peptide architectures, FPS setup and PPI mapping. (A) Architecture of multimeric geph-binding peptides. An (Fmoc)-L-Lys(Fmoc) building block facilitated multimerization of geph-binding epitopes (e), linked together by PEG moieties (o_{1-3}). (B) schematic representation of FPS measurements. The receptor-binding domain of the neuronal scaffolding protein gephyrin (gephE) is immobilized on the ligand strand via an NHS coupling. The binding of unmodified, multimeric peptides during the association phase is detected by a change in fluorescence of green dye b. (C) Real-time affinity determination of overlapping, dimeric GlyR β derived peptides in FPS. Peptides were used at a concentration of 1 µM. Note that only peptides with a centred FSIVG core binding motif exhibited a measurable affinity.

Comparison of FPS with ITC, BLI and TRIC

Next, we assessed the reliability of apparent K_D values determined in FPS by comparing this setup with commonly used immobilization- and in-solution-based PPI-quantification methods. Namely, real-time binding quantification using biolayer interferometry (BLI), HT temperature related intensity change (TRIC) quantification as well as precise calorimetric measurements (ITC) (Figure 2 A). Compared to ITC measurements, which can be considered the gold standard as they quantify directly and label-free in solution (Figure 2A), HT quasi label-free TRIC measurements (Figure 2 B) recapitulate the same trend. The only exception being compound e=8, o_1=0, o_2=4, which was outside of the dynamic range. The BLI measurements (Figure 2 C) necessitated loading densities and ligand concentrations that did for effective dissociation of tetramers (Supplementary Figure 2) and octamers (Supplementary Figure 3). Thus, affinities could not be derived from single curves but were instead assessed through

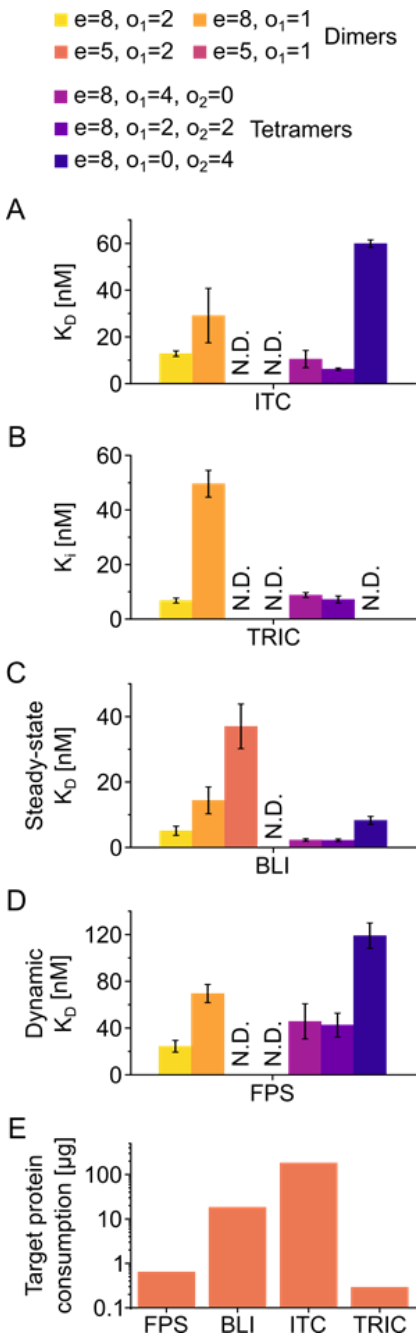


Figure 2: Comparison of apparent affinities of dimeric and tetrameric peptides in ITC, TRIC, BLI and FPS and target protein consumption. Apparent binding affinities of seven benchmark peptides (four dimeric and three tetrameric) were measured using ITC (A), TRIC (B), BLI (C) and FPS (D). For complete measurements, see supplementary figure 6, 7, 8, and 9 respectively. (E) Amount of target protein consumed for affinity determination of one peptide in the four methods tested (FPS: one sensor chip functionalization, BLI: functionalization of eight sensors, ITC: one run with 16

injections, TRIC: 16-point dose response in displacement assay setup).

steady-state BLI measurements using multiple peptide concentrations. The determined K_D values only partly recapitulated the affinities determined in ITC, possibly due to avidity effects such as re-binding.

Along the same line, BLI overestimated the affinity of the tetramers and further enabled the measurement of e=5, o=2, a lower affinity dimer. Conversely, the on- and off-rates of dimeric peptides were resolvable in BLI (Supplementary Figure 4). However, a poor signal-to-noise ratio (SNR) was observed for small dimeric peptides (Supplementary Figure 4 C and D). In stark contrast, FPS enabled measurements of dimeric, and tetrameric compounds independent of compound size (Figure 2 D). The resolved binding hierarchy is in line with ITC and TRIC, similar so the apparent dynamic range.

Next, we compared the protein sample consumption of the four different biophysical PPI quantification methods (Figure 2 E). In terms of target protein consumption by weight, FPS performed second best among the methods employed, consuming 28.5-fold less protein than BLI measurements for sensor functionalization (0.64 µg for one FPS sensor chip versus 18.25 µg for 8 BLI biosensors), 285-fold less than ITC (182.4 µg for one run) and 2.2-fold more than TRIC (0.29 µg for a 16-point dose response) (Figure 2 B).

To facilitate the determination of kinetic binding parameters of hundreds of peptides with a short turnover, we explored the possibility to directly couple FPS to low µM scaled solid-phase peptide synthesis. Consequently, we determined the intra-synthesis reproducibility of real-time affinity measurements of multimeric, unmodified peptides in FPS. K_D values and kinetic parameters could be determined with low deviation using independently synthesized dimers and tetramers (Supplementary Figure 5), indicating that the combined setup allows for reproducible and precise kinetic interactions studies.

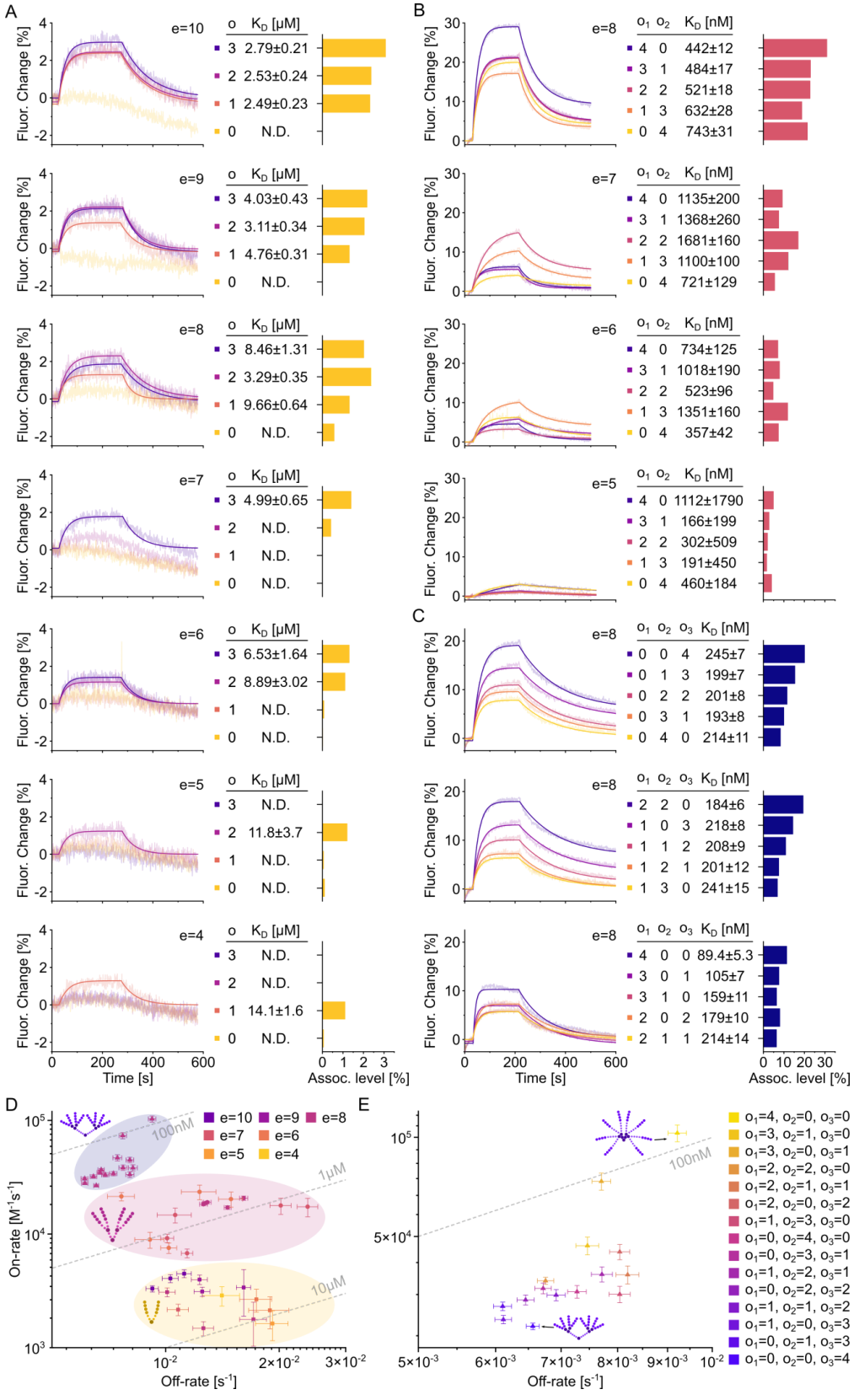
HT determination of protein affinities and kinetics using FPS

Next, we used the established FPS setup to resolve the relationship between multimeric peptide architecture and binding kinetics. Specifically, an array of dimeric, tetrameric, and octameric compounds was subjected to FPS measurements at a fixed concentration of 1 µM to achieve sufficient signal amplitude for weaker binders (Figure 3). In addition to the on- and off-rates determined from functions fit to the obtained curves, association levels, at which the measured curves plateaued, were determined for each peptide. Overall, a prominent gain in affinity could be observed from dimers (Figure 3 A, low µM) to tetramers (Figure 3 B, high nM) and finally octamers (Figure 3 C, mid/low nM). Indeed, plotting of the obtained on-rates against the off-rates for each compound in a rate-map (Figure 3 D) reveals that multimer

affinity primarily depends on the valency. This is in line with previous studies (Errington et al., 2019) which demonstrated that increased valency also increases the ability to create additional binding conformations within the configurational network. The second most important factor is the length of the epitope. This trend recapitulates the changes in binding strength that have been observed for the respective monovalent counterparts (Maric et al 2015). In the here studied multivalent system, the observed affinity gain is primarily driven by on-rate effects which vary over two magnitudes, while the off-rates vary only 5-fold across all tested species. Together, these data confirm the importance of the binding affinity of the single binding epitopes for higher valency systems, demonstrating the importance of on-rate effects.

FPS correlates multivalent topology and binding dynamics

To resolve how topological multimeric features determine on- and off-rates, our measurements included a series of compounds identical in epitope length and number but systematically varied scaffold arrangement. Plotting the obtained on-rates against the off-rates for each compound as a rate-map, together with color-coding of the topological adjustments visualizes a clear trend (Figure 3 E). The octamer with the lowest affinity is characterized by a multivalent architecture that enables flexible movement of the two sides of the multimer but sterically restricted movement of the epitopes themselves within the two tetramers. Vice versa, the multimeric architecture that enabled the greatest flexibility close to the epitopes while at the same time enforcing pre-orientation of the epitopes through sterical constrains in the centre displayed the highest affinity. The difference in affinity between both compounds is primarily driven by on-rate (4.5-fold) but also off-rate effects (1.4-fold).This dataset resolves the structure-activity relationship of multivalent geph-binders and provides a framework for the development high-valency, ultra-high affinity interactors in general.


A
B
C
D
E
Fluor. Change [%]
Time [s]
Assoc. level [%]
e=10
e=9
e=8
e=7
e=6
e=5
e=4
o K_D [µM]
3 2.79±0.21
2 2.53±0.24
1 2.49±0.23
0 N.D.
3 4.03±0.43
2 3.11±0.34
1 4.76±0.31
3 8.46±1.31
2 3.29±0.35
1 9.66±0.64
3 4.99±0.65
2 N.D.
1 N.D.
3 6.53±1.64
2 8.89±3.02
3 N.D.
2 11.8±3.7
1 N.D.
1 14.1±1.6
o1 o2 K_D [nM]
4 0 442±12
3 1 484±17
2 2 521±18
1 3 632±28
0 4 743±31
4 0 1135±200
3 1 1368±260
2 2 1681±160
1 3 1100±100
0 4 721±129
4 0 734±125
3 1 1018±190
2 2 523±96
1 3 1351±160
0 4 357±42
4 0 1112±1790
3 1 166±199
2 2 302±509
1 3 191±450
0 4 460±184
o1 o2 o3 K_D [nM]
0 0 4 245±7
0 1 3 199±7
0 2 2 201±8
0 3 1 193±8
0 4 0 214±11
2 2 0 184±6
1 0 3 218±8
1 1 2 208±9
1 2 1 201±12
1 3 0 241±15
4 0 0 89.4±5.3
3 0 1 105±7
3 1 0 159±11
2 0 2 179±10
2 1 1 214±14
On-rate [M⁻¹s⁻¹]
Off-rate [s⁻¹]
100nM
1µM
10µM
o1=4, o2=0, o3=0
o1=3, o2=1, o3=0
o1=3, o2=0, o3=1
o1=2, o2=2, o3=0
o1=2, o2=1, o3=1
o1=2, o2=0, o3=2
o1=1, o2=3, o3=0
o1=0, o2=4, o3=0
o1=0, o2=3, o3=1
o1=1, o2=2, o3=1
o1=0, o2=2, o3=2
o1=1, o2=1, o3=2
o1=1, o2=0, o3=3
o1=0, o2=1, o3=3
o1=0, o2=0, o3=4

9

Figure 3: FPS resolves binding kinetics of dimeric, tetrameric, and octameric peptide binders in HT. (A-C) FPS curves of all dimers (A, yellow), tetramers (B, red), and octamers (C, purple) tested are displayed next to the respective association levels. For a complete list of the kinetic parameters of all compounds tested, refer to supplementary table 1. (D) Rate map of all dimers (yellow), tetramers (red), and octamers (blue) with a determinable on- and off-rate. Epitope lengths are color-coded. Note the the high dependence of dimer affinity on epitope length. 10 µM, 1 µM, and 100 nM affinities are indicated as dashed, grey lines. (E) Zoomed-in view of (D) with octameric binders in focus. Varying architectures are color-coded, and 100 nM affinity is indicated as dashed, grey line. Note that octamers with highest affinity contain ≥3 PEG building blocks in the outer o_1 position.

Prediction of multivalent binding parameters

The 40 successfully measured compounds constitute only a small fraction of the theoretical possible combinations. To discern whether the obtained dataset allows to predict multimer properties, we used machine learning. Specifically, we applied the Random Forest Regressor using the encoded amino acids and analogous building blocks as training input. Here, the peptide sequences are represented through the amino acid composition (Spänig and Heider, 2019), which demonstrated overall good performance across multiple applications and provides easy interpretability (Spänig et al., 2021). First we explored whether the observed on- and off-rates and the resulting K_D values can be reliably predicted. To this end, we applied a leave-one-out cross-validation and found a high correlation between predicted and observed K_D values (Figure 4 A, Supplementary Table 2), off-rates (Figure 4 B) and on-rates (Figure 4 C) in case of the tetrameric and octameric group. In case of the dimeric peptides, a positive correlation was only found for the K_D values. We additionally examined the correlation between observed association level and K_D, on- and off-rate for each compound. Here, positive Pearson correlations were found in case of the dimeric group for K_D (Figure 4 D) and especially off-rate (Figure 4 E) but not on-rate (Figure 4 F). In stark contrast, no or even negative correlations were found in the tetramer and octamer group when correlating the observed association level to the K_D values (Figure 4 D), off- (Figure 4 E) and on-rates (Figure 4 F).

Taken together, these results indicate that for both lower avidity dimers and higher avidity tetramers and octamers, K_D values can be reliably predicted across multivalent species using the outlined algorithm. In stark contrast, the association level may only be a representative metric for K_D and off-rate for distinct topology classes.

Peptide binders with high avidity potently neutralize native gephyrin

Our FPS studies suggest that higher-order geph-binding multimers possess enhanced potency as inhibitors compared to their dimeric counterparts. Using a complementary peptide microarray-based approach (Schulte et al., 2021) with native geph from mouse brain lysates, we probed the geph neutralizing capacity of dimeric and tetrameric geph binders. Native geph was pre-incubated with dimeric and tetrameric binders (Figure 4 G) with varying

architecture at increasing peptide competitor concentrations. Reduction in on-chip peptide binding by geph thus corresponds to neutralization of geph by competitor binding. Tetrameric binders exhibited up to two orders of magnitude more potent geph neutralization than the dimeric binders (Figure 4 H), thereby confirming the outcome of the FPS-based HT screen and further highlighting the value of the outlined approach for avidity-based binding optimization.

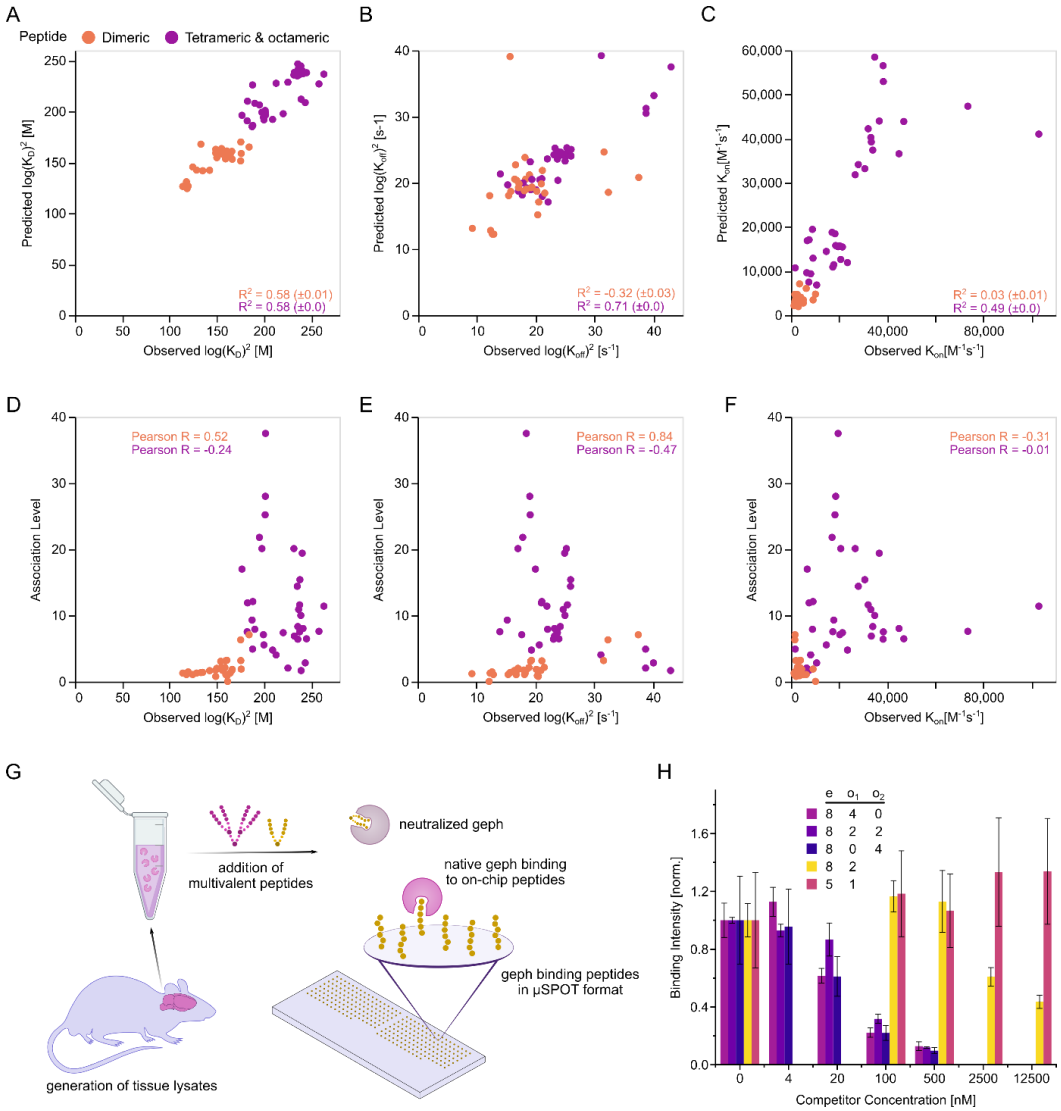


Figure 4: Multimer binding prediction and inhibition potency. Measured K_D values (A), off- (B), and on-rates (C) are plotted against predicted values in a leave-one-out cross-validation. Note the high correlation between predicted and obtained K_D values. The obtained association levels are plotted against the observed K_D values (D), off- (E), and on-rates (F). Note the low correlation between association levels and other kinetic parameters. (G) Schematic representation of µSPOT peptide microarrays, harbouring geph-binding peptides as cellulose conjugates. Native geph from mouse brain lysates was preincubated with multimeric peptides to neutralize geph-binding to on-chip peptides. (H) Normalized geph binding intensity to GlyR β-derived on-chip peptides in µSPOT format in the presence of varying competitor concentration. Native geph binding to on-chip peptides was

resolved by antibody detection and chemiluminescent readout. Note that tetrameric peptides effectively neutralized geph binding at lower concentrations than dimeric peptides. Data are presented as mean of two experiments and the corresponding standard deviations.

Discussion

FPS is a versatile technique for measuring binding affinities of binder–ligand systems, commonly DNA/protein, protein/protein and protein/small molecules. This study employs FPS in tandem with automated, low µM-scaled solid-phase peptide synthesis to establish a platform for HT real-time binding affinity determination. This setup was used to systematically characterize +100 multimeric peptides with varying architecture, binding to the target protein geph. Contrary to other examples of kinetic studies of multimeric binders (Chi et al., 2010), we observed an increase in binding affinity of higher-order multimers mainly driven by an increase in on-rate. Our work confirmed valency and monovalent binding affinity as the primarily relevant design features that govern the magnitude of avidity enhanced binding. In the same line, we found that within the complex octameric linker architecture, a high degree of flexibility close to the geph-binding epitope is preferential for binding affinity as opposed to high flexibility within the core of the octamer. This observation could be explained by an improved preorientation and/or access to different binding conformations. Further, we demonstrate the successful data-based prediction of affinities, otherwise hard to achieve using biophysical and structural data alone.

Major limitations of contemporary kinetic methods such as BLI are irresolvable off-rates in case of high avidity compounds (Supplementary Figure 2 and 3). Gratifyingly, the here presented FPS setup provided insights into the off-rates of these higher-architecture binders, which could be explained by the higher distance between the immobilized target protein in the heliX system compared to the distance on Ni-NTA biosensors in BLI, excluding complex re-binding effects on the biosensor surface. In addition, measurements of smaller and lower affinity dimeric peptides suffered from a poor SNR in BLI (Supplementary Figure 4), whereas FPS measurements provided superior SNR largely independent of ligand size. In terms of resource consumption, FPS was on par with TRIC-based measurements and vastly outperformed both BLI and ITC. Yet, in our specific system, an inverse dependence of the observed on-rate on the employed analyte concentration was found (Supplementary Figure 10), indicating that it’s required to probe selected analytes at multiple concentrations before subjecting an array of varying compounds to a single-concentration screen and validate selected hits in complementary biophysical methods such as ITC. Another possible limitation in FPS are low SNRs when screening libraries with small compound size. This could be addressed by competition FPS setups with displaceable fluorescent compounds to further boost the signal amplitude (Ponzo et al., 2019).

Importantly, this study identified novel binders with avidity enhanced inhibition capacity towards the *ex vivo* derived native protein. We anticipate that HT FPS measurements in tandem with automated approaches for ligand synthesis will aid in similar projects advancing the rational optimization of effectors with unnatural building blocks (Zhang et al., 2022) and other multimeric effectors, including multivalent protein-carbohydrate interactions (Tsouka et al., 2021). Such high avidity binders could contribute to advance our understanding of protein function and localization by expanding the toolbox of versatile chemical biology probes.

Significance (178/300 – *common significance statements are around 150-200*)

Peptide-based effectors of protein-protein Interactions (PPIs) are an attractive modality for probing protein function in chemical biology and selective pharmacological targeting of proteins in disease. Notable advantages are high selectivity and affinity when compared to small molecules and size when compared to antibodies. Leveraging peptide multimerization to further boost affinity and specificity via avidity harbours enormous potential to develop probes and therapeutics with ultra-high potency. Precise biophysical binding studies, however, remained challenging. Fluorescence proximity sensing (FPS) allowed us to precisely determine binding kinetics and thermodynamics of diverse multivalent topologies in highest throughput. A seamless transition from automated peptide synthesis to real-time affinity determination enabled of the effective engineering of multimeric architectures towards activity that was confirmed inhibiting the ex-vivo derived native protein. Subjected to machine learning the dataset enabled the affinity prediction for the here tested multivalent species, otherwise hard to achieve using conventional approaches. The results illuminate the structure-activity relationship of multimeric peptide-based effectors, establish FPS as a viable method for probing PPIs and identify highly potent inhibitors of gephyrin, the master regulator of inhibitory neuronal transmission.

Conflict of Interest/Competing Interest

Soldà A and Strasser R are employed at Dynamic Biosensors which commercializes the heliX® system for FPS measurements. Adams N, Bekić I and Streicher W are employed at Nanotemper which commercializes the Dianthus® system for TRIC measurements. The other authors declare no competing interests.

Acknowledgements

We thank Sonja Kachler for her excellent technical assistance. This work was funded by the DFG (DFG MA6957/1-1) to H.M.M. and C.S. and by the Bundesministerium für Wirtschaft und Energie (BMWi) in the project MoDiPro-ISOB (16KN0742325) to A.S., N.A., I.B., S.S., R.S., D.H. W.S..

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Hans Michael Maric (Hans.Maric@uni-wuerzburg.de)

Author Contributions

Conceptualization: H.M.M., C.S.; *Methodology*: C.S., A.S., N.A., I.B., S.S.; *Software*: S.S., D.H., *Formal Analysis*: C.S., A.S., N.A., I.B., S.S.; *Investigation*: C.S., A.S., N.A., I.B.; ; *Writing – Original Draft*: C.S., H.M.M.; *Writing – Review & Editing*: A.S., S.S., N.A., I.B., R.S., D.H.; *Visualization*: C.S., S.S.; *Supervision*: H.M.M., D.H., R.S., W.S.; *Project Administration*: H.M.M., R.S., W.S., D.H.; *Funding Acquisition*: H.M.M., R.S., W.S., D.H.

Methods

Unless otherwise stated, amino acids and reagents were purchased from either Iris Biotech or Carl Roth. All solvents were purchased from commercial sources and used without further purification.

Automated Solid-Phase Peptide Synthesis

µSPOT peptide arrays (Dikmans et al., 2006) were synthesized using a MultiPep RSi robot (CEM GmbH, Kamp-Lindford, Germany) on in-house produced, acid-labile, amino-functionalized, cellulose membrane discs containing 9-fluorenylmethyloxycarbonyl-β-alanine (Fmoc-β-Ala) linkers (average loading: 130 nmol/disc – 4 mm diameter). Synthesis was initiated by Fmoc deprotection using 20% piperidine (pip) in dimethylformamide (DMF) followed by washing with DMF and ethanol (EtOH). Peptide chain elongation was achieved using a coupling solution consisting of preactivated amino acids (aas, 0.5 M) with ethyl 2-cyano-2-(hydroxyimino)acetate (oxyma, 1 M) and *N*,*N*′-diisopropylcarbodiimide (DIC, 1 M) in DMF (1:1:1, aa:oxyma:DIC). Couplings were carried out for 3×30 min, followed by capping (4% acetic anhydride in DMF) and washes with DMF and EtOH. Synthesis was finalized by deprotection with 20% pip in DMF (2×4 µL/disc for 10 min each), followed by washing with DMF and EtOH. Dried discs were transferred to 96 deep-well blocks and treated, while shaking, with sidechain deprotection solution, consisting of 90% trifluoracetic acid (TFA), 2% dichloromethane (DCM), 5% H_2O and 3% triisopropylsilane (TIPS) (150 µL/well) for 1.5 h at room temperature (rt). Afterwards, the deprotection solution was removed, and the discs were solubilized overnight (ON) at rt, while shaking, using a solvation mixture containing 88.5% TFA, 4% trifluoromethanesulfonic acid (TFMSA), 5% H_2O and 2.5% TIPS (250 µL/well). The resulting peptide-cellulose conjugates (PCCs) were precipitated with ice-cold ether (0.7 mL/well) and spun down at 2000×g for 10 min at 4 °C, followed by two additional washes of the formed pellet with ice-cold ether. The resulting pellets were dissolved in DMSO (250 µL/well) to give final stocks. PCC solutions were mixed 2:1 with saline-sodium citrate (SSC) buffer (150 mM NaCl, 15 mM trisodium citrate, pH 7.0) and transferred to a 384-well plate. For transfer of the PCC solutions to white coated CelluSpot blank slides (76×26 mm, Intavis AG), a SlideSpotter (CEM GmbH) was used. After completion of the printing procedure, slides were left to dry ON.

Preparative Peptide Synthesis

Standard solid-phase peptide synthesis with Fmoc chemistry was applied, shortly, 2-chlorotrityl resin (1.6 mmol/g) was swollen in dry DCM with 2 eq. of N,N-Diisopropylethylamine (DIEA). Then, the desired aa (1eq) and the orthogonally protected

Boc-Gly-OH (1eq) were loaded. Boc-Gly-OH reduces resin loading in order to prevent aggregation of the elongating peptide chain. After ON reaction, the resin was capped with MeOH and washed with DCM and DMF. Deprotection and conjugation cycles followed, where 20% pip solution in DMF was used to deprotect, and after washes, the peptide chain was elongated by adding aa (4eq.) with oxyma (4eq.) and DIC (4eq.). Coupling efficiency was monitored by measuring the absorption of the dibenzofulvene–pip adduct after deprotection. The peptides were cleaved from the resin using a cocktail of 90.5% TFA, 4% H_2O, 3% TIPS 5% thioanisole, 2.5% 1,2-Dithiothreitol for 2 h at rt. The peptides were precipitated and washed twice with ice-cold ether, then purified with high-performance liquid chromatography (HPLC), and analyzed by liquid chromatography-mass spectrometry (LCMS) (Supplementary Table 3).

Unmodified peptides synthesized in 2 µmol scale were bought from Intavis Peptide Services (SKU: 90.215) with a free N-terminal amino end and C-terminal amide group and were used for FPS and BLI measurements without further purification. Crude peptide purity was assessed by LC-MS similar to preparatively synthesized peptides (Supplementary Table 4).

Preparation of Mouse Tissue Lysates

Whole mouse brains were obtained from C57BL/6J mice at >4 weeks of age and immediately flash-frozen in liquid N_2. Before lysis, whole mouse brains were weighed and cut into four pieces along the horizontal and vertical axis. To prepare one lysate, two diagonally opposite pieces were transferred into a 1.5 mL reaction tube (Sarsted). Lysis was carried out on ice in 500 µL HEPES lysis buffer (20 mM HEPES, 100 mM KCH_3COO,40 mM KCl, 5 mM $MgCl_2$, 5 mM DTT, 1 mM PMS, 5 mM EDTA, 1% Triton X-100, 1% complete EDTA-free protease inhibitor cocktail (Roche) (all v/v)), by hand crushing the brain material with a hand pestle in a 1.5 mL reaction tube. Lysis was completed by 1 min sonification on ice with a Sartorius Labsonic M Sonificator at 20% amplitude with care to avoid heating the suspensions. Finally, Lysates were centrifuged for 15 min at 17,200×g and 4 °C. The SN was subsequently collected, transferred to a new 1.5 mL reaction tube, flash-frozen in liquid N_2 and stored at -80 °C until use.

Microarray Binding Assay

µSPOT slides were blocked by incubation with 2.5 mL 5% (w/v) blotting grade milkpowder (MP, Carl Roth) in PBS for 60 min at ~70 revolutions per minute (rpm) and RT. Afterwards, slides were incubated with 0.8% (v/v) mouse brain lysate 5% MP in 1 × PBS for 15 min before slides were washed with 3×2.5 mL 1×PBS for 1 min. To label native geph for

detection, the slides were incubated with 2.5 mL of a 1:5,000 diluted primary antibody (anti-gephyrin (3B11, SynapticSystems) in 5% MP in 1×PBS for 15 min, after which the slides were washed with 3×2.5 mL 1×PBS for 1 min. Afterwards, the slides were incubated with a secondary HRP-coupled Anti-mouse antibody (31430, Invitrogen) in 5% MP in 1×PBS for 15 min, after which the slides were washed with 3×2.5 mL 1×PBS for 1 min. Peptide binding was detected through chemiluminescent detection (Lowest Sensitivity, 30s exposure time) after application of 200 µL of SuperSignal West Femto Maximum Sensitive Substrate (Thermo Scientific) per slide using a c400 imaging system (Azure).

For on-chip peptide competition, native geph was preincubated with the indicated peptides in 5% MP in PBS for 30 min on ice before being put on an array slide.

Binding intensities were evaluated using FIJI including the Microarray Profile addon (OptiNav). After background subtraction of the mean greyscale value of the microarray surface surrounding the spots, raw greyscale intensities for each position were obtained for the left and right sides of the internal duplicate on each microarray slide. The standard deviation (STDEV) between both sides was obtained using formula (1).

$$STDEV = \sqrt{\frac{\sum(x-\bar{x})}{n}} \quad (1)$$

with

n The total number of data points
$\bar{x}$ The mean intensity value

Afterwards, the raw intensities of all spots of interest were summed and normalized to the summed intensity of the condition without competitor peptide.

Protein Expression and Purification

GephE (gephyrin P2 splice variant residues 318–736) was expressed in Escherichia coli and purified in a two-step purification as described earlier (Kim et al., 2006, Schrader et al., 2004). Concisely, the protein was purified using via Intein-tag (Chitin beads, New England BioLabs), and after self-cleavage the protein could be obtained by size-exclusion chromatography (SEC) column (HiLoad 16/600 Superdex 200pg, GE Healthcare) on an ÄKTA explorer system (GE Healthcare). His-tagged gephE was produced similarly with the exception of purification on IMAC coloumns before SEC purification.

Temperature Related Intensity Change (TRIC) Assays

For the TRIC assay, 16-point affinity measurements with each peptide against a target complex in duplicates were performed on the Dianthus NT.23PicoDuo. The experiment was performed in a single Dianthus 384-microwell plate using an assay buffer of 1× PBS, 2 mM reduced L-Glutathione and 0.1 % Pluronic® F-127, pH 7.4. Target protein and tracer peptide was diluted to 40 nM gephE and 20 nM NN1D-Cy5 in assay buffer and incubated on ice for one hour to create the target complex. All peptides were first pre-diluted to 2 mM in assay buffer and subsequently, a 16 point, 1:1 dilution series of each peptide was performed with an electronic multichannel pipette to a final volume of 10 µl directly in the Dianthus plate. Afterward, each dilution was mixed with 10 µl target complex, resulting in 16-point dilutions series of the peptides with a final concentration from 1 mM to 30.52 nM in the assay with 20 nM gephE and 10 nM NN1D-Cy5. The plate was centrifuged for 30 sec at 1000 ×*g* and incubated at 25 °C for 30 min. The final measurement of the plate was performed at 25 °C where the fluorescence signal of the samples was measured for 1 sec with the IR-laser off and for 5 sec on, resulting in TRIC traces where the detected fluorescence values are displayed as the relative fluorescence over time and under influence of the IR-laser induced heating and normalized to a value of one. For further analysis of the assay, the fluorescent signal is again normalized by dividing the fluorescence values after IR laser activation with the fluorescence values prior to the activation giving the normalized fluorescence F_{norm} in ‰. For a competitions assay of this kind, the affinity is evaluated by K_i values which are obtained by applying a Hill-fit to a plot of F_{norm} vs. ligand concentration to determine an EC_{50} value (Formula 2 and 3).

The affinity of the tracer peptide to gephE was determined in the same assay buffer as the TRIC experiments were performed. gephE was diluted to 1000 nM and subsequently a 16-point dilution series of the protein was performed directly in a Dianthus plate in triplicate to a final volume of 10 µL. The gephE dilutions were mixed directly with 10 µL 2 nM NN1D-Cy5 to a final volume of 20 µL at 1 nM NN1D-Cy5 with protein concentration between 500 and 0.015 nM. The samples were subject to the same Dianthus parameters as above but analysed with a K_D fit for later use in the determination of K_i values.

$$K_i = \frac{K_D}{2-\gamma} \bullet \left(\frac{EC_{50}}{\frac{[T]_t}{\gamma} - \frac{K_D}{2-\gamma} - \frac{[C]_t}{2}} - \gamma \right) \quad (2)$$

with

$$\gamma = \frac{[T]_t + [C]_t + K_D - \sqrt{([T]_t + [C]_t + K_D)^2 - 4[T]_t[C]_t}}{2[C]_t} \quad (3)$$

and

$[T]_t$ Final concentration of the target protein (gephE)

$[C]_t$ Final concentration of fluorescent tracer peptide (NN1D-Cy5) that is in competition with unlabelled peptide ligand in the assay

K_D The determined K_D between the fluorescent tracer and the target protein from a direct binding affinity measurement

EC_{50} The EC_{50} obtained from the above-described competition assay between the unlabelled peptide ligand with the target complex

Isothermal titration calorimetry (ITC)

ITC measurements were performed using an ITC200 (MicroCal) at 25 °C and 1000 rpm stirring. PBS pH 7.4 was used as the standard solvent. Specifically, 40 µL of a solution 200 µM of dimeric, or 100 µM of tetrameric compounds was titrated into the 200 µL sample cell containing 20 µM GephE. In each experiment, a volume of 2.5 µL of ligand was added, resulting in 15 injections and a final molar ratio between 1:0.5 (tetrameric compounds) and 1:1 (dimeric compounds). The dissociation constant (K_D) and stoichiometry (N) were obtained by data analysis using NITPIC, SEDPHAT, and GUSSI (Brautigam et al., 2016). Measurements were conducted three times for each probe and are given as mean values with the resulting standard deviations.

Biolayer interferometry (BLI)

BLI measurements were carried out using the ForteBio Octet RED96 system. The chamber temperature was kept constant at 25 °C with a plate agitation speed of 1000 rpm. Briefly, Ni-NTA-coated biosensors were dipped into 200 µL of a 200 nM His-GephE solution (in a kinetic buffer (KB): 1×PBS with 0.1% (w/v) BSA, 0.05% (v/v) Tween20, 2 mM GSH) for protein immobilization. The loaded sensors were moved to solutions containing various concentrations (200 – 0.781 nM) of dimeric, tetrameric and octameric peptides solubilized in KB to obtain the association curve. After the 180-300s association step, the sensors were moved to KB to obtain the dissociation curve. A buffer only condition with a loaded biosensor was used as a reference for background subtraction. The association and dissociation curve

were fitted with the ForteBio Biosystems Data Analysis HT Software (local fitting algorithm, 1:1 model).

Preparation of Protein-DNA Conjugates

GephE was covalently coupled via its primary amines to the 5'end of ssDNA (cNL-A48, ligand strand) (coupling kit HK-NHS-1, Dynamic Biosensors, Martinsried, DE). The protein-DNA conjugate was purified from the free protein and free DNA using the proFIRE® system (Dynamic Biosensors, Martinsried, DE) (Wiener et al., 2020, Reinking and Stingele, 2021). The purification gives a good first impression of the status of the protein after conjugation and can be used as quality control of the protein sample to be immobilized onto the surface (for the chromatogram, see Supplementary Figure 11). The embedded Data Viewer software provides protein-DNA conjugate purity and concentration based on the chromatogram. The yield of the gephE-DNA (1:1 ratio) is sufficient for approximately 300 chip functionalizations, considering a chip density of 100% and a ligand concentration of 100 nM). After liquid nitrogen freezing, the conjugates were stored at a concentration of 500 nM in PE40 buffer (10 mM Na_2HPO_4/NaH_2PO_4, 40 mM NaCl, 0.05 % Tween20, 50 µM EDTA, 50 µM EGTA) at −80 °C and were freshly thawed before each experiment.

Chip Functionalization

All switchSENSE experiments were performed on a dual-color heliX$^+$ instrument using a standard heliX Adapter Biochip (ADP-48-2-0, Dynamic Biosensors, Martinsried, DE), in which single-stranded DNA (anchor strands) are covalently attached to the chip surface. Each chip is equipped with 2 gold electrodes (or spots), with different DNA anchor strands. Herein, we used spot 1 as measurement spot with the conjugated target protein (gephE-DNA) and spot 2 as real time referencing (only DNA), in order to monitor possible unspecific binding of the peptides on the anchor DNA and/or gold electrodes. Firstly, the conjugate gephE-DNA (ligand strand) was preincubated with the complementary ssDNA carrying the Gb fluorophore (adapter strand), for 20min at RT upon shacking (600rpm). Secondly, the whole ligand construct was immobilized on the biochip via hybridization of complementary anchor strand (for a schematic representation, see Supplementary Figure 12). The chip was regenerated and freshly functionalized before each measurement series. For chip regeneration, the double stranded DNA nanolevers were denatured by disrupting the hydrogen bonds between base pairs using a high-pH regeneration solution (HK-REG-1, Dynamic Biosensors). The conjugate is washed away while the covalently attached single-stranded nanolevers remained on the surface and could be reused for a new functionalization step. Using FPS mode, a DNA-based biochip can be regenerated up to 50 times.

Fluorescence Proximity Sensing (FPS) Mode – switchSENSE interaction analysis

Interaction analysis was performed in fluorescence proximity sensing (FPS) mode with a constant voltage of -0.4 V applied, which forces the surface-tethered DNA into a fixed angle. When the protein analyte binds to the DNA target, it affects the average distance of the fluorescent label from the fluorescence-quenching gold surface. Besides the change in DNA orientation, a change in close proximity to the fluorescent dye or direct interaction of the protein with the fluorescent dye lead to measurable changes in the fluorescence intensity. In the FPS measurements, the series of peptides were being flushed at specified concentrations over the two electrodes of the biochip. When the peptide reaches the target protein (gephE) present in spot 1, we observed an increase in the fluorescence signal on the timescale of seconds. Hence, the concentration jump itself may be considered instantaneous, and the time dependence of the fluorescence signal directly reflects the protein-peptide kinetics. After flushing out the peptide and replacing the bulk solution with pure buffer, only dissociation can take place. During measurements the sample tray containing the protein/peptide samples was set to 25°C, as well as the experiment temperature on the biochip. Peptide samples were diluted and measured in PE140 buffer (10 mM Na_2HPO_4/NaH_2PO_4, 140 mM NaCl, 0.05 % Tween20, 50 µM EDTA, 50 µM EGTA). Flow rate for association and dissociation reactions was set to 200 µL/min. The green LED power was set to 4. Experiment design, workflow and data analysis were performed with the heliOS software from Dynamic Biosensors. The association and dissociation rates (k_{on} and k_{off}), dissociation constants (K_D) and the respective error values were derived from a global single exponential fit model, upon double referencing correction (blank and real-time).

Machine Learning

We employed Snakemake v6.9.1 using Python v3.8.5 to develop the machine learning workflow (Köster and Rahmann, 2012). First, we removed all sequences with no available K_D values, on-, or off-rates. Moreover, we used the median values for duplicated sequences, i.e., repeated measurements. Afterward, we log-square-transformed K_D and k_{off} to retain issues with floating-point arithmetic. Specifically, we applied the FunctionTransformer from scikit-learn v1.0 using $log(x)^2$, with log being the natural logarithm (Pedregosa et al., 2011). We encoded the peptides using the amino acid composition (AAC) (Spänig and Heider, 2019) and the linker sequence through the one-hot encoding. Thus, we introduced three binary representations to transform the linker into a machine-readable format and assigned the actual linker (J) to $[1, 0]$ and the spacer (O) to $[0, 1]$. Since the model requires a fixed-length input, we also introduced gaps, denoted as $[0, 0]$.

The AAC encoding counts the number of all amino acids concerning the total sequence length:

$$f(t) = \frac{N(t)}{N}$$ (Chen et al., 2018).

$N(t)$ denotes the number of amino acids t, N refers to the peptide length, and $f(t)$, finally, is the composition of t (Chen et al., 2018). The resulting matrix X contains 79 peptides represented by 20 proteinogenic amino acids and a binary vector of length 14, thus, 34 features. Note that we removed all AAC features with zero variance before model training.

Afterward, we used the Random Forest Regressor with default arguments. We verified the model employing leave-one-out cross-validation (LOOCV), i.e., we trained k models using $k - 1$ peptides to predict the k-th peptide. For model evaluation, we computed the correlation coefficient R^2, which is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Specifically, y_i is the i-th observed K_D value, on-, or off-rate, $\hat{y}_i$ is the i-th predicted K_D value, on-, or off-rate, and $\underline{y}$ is the average K_D value, on-, or off-rate. To score the correlation between the association level and K_D values, on- and off-rates, we utilized Pearson's product-moment correlation coefficient:

$$Pearson\ R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

We used the implementations provided by the scikit-learn library. Finally, we conducted a 1000-fold bootstrapping for the total R^2 and confidence interval (CI) calculation.

References

BACH, A., CLAUSEN, B. H., MØLLER, M., VESTERGAARD, B., CHI, C. N., ROUND, A., SØRENSEN, P. L., NISSEN, K. B., KASTRUP, J. S. & GAJHEDE, M. 2012. A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1-2 and protects against ischemic brain damage. *Proceedings of the National Academy of Sciences,* 109**,** 3317-3322.

BRAUTIGAM, C. A., ZHAO, H., VARGAS, C., KELLER, S. & SCHUCK, P. 2016. Integration and global analysis of isothermal titration calorimetry data for studying macromolecular interactions. *Nature protocols,* 11**,** 882-894.

BRUNETTI, J., FALCIANI, C., BRACCI, L. & PINI, A. 2018. Branched peptides as bioactive molecules for drug design. *Peptide Science,* 110**,** e24089.

CHEN, Z., ZHAO, P., LI, F., LEIER, A., MARQUEZ-LAGO, T. T., WANG, Y., WEBB, G. I., SMITH, A. I., DALY, R. J. & CHOU, K.-C. 2018. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics,* 34**,** 2499-2502.

CHI, C. N., BACH, A., GOTTSCHALK, M., KRISTENSEN, A. S., STROMGAARD, K. & JEMTH, P. 2010. Deciphering the kinetic binding mechanism of dimeric ligands using a potent plasma-stable dimeric inhibitor of postsynaptic density protein-95 as an example. *The Journal of biological chemistry,* 285**,** 28252-60.

DEMMER, O., DIJKGRAAF, I., SCHUMACHER, U., MARINELLI, L., COSCONATI, S., GOURNI, E., WESTER, H.-J. R. & KESSLER, H. 2011. Design, synthesis, and functionalization of dimeric peptides targeting chemokine receptor CXCR4. *Journal of medicinal chemistry,* 54**,** 7648-7662.

DIKMANS, A., BEUTLING, U., SCHMEISSER, E., THIELE, S. & FRANK, R. 2006. SC2: a novel process for manufacturing multipurpose high-density chemical microarrays. *Qsar & Combinatorial Science,* 25**,** 1069-1080.

ERLENDSSON, S. & TEILUM, K. 2021. Binding revisited—avidity in cellular function and signaling. *Frontiers in Molecular Biosciences***,** 470.

ERRINGTON, W. J., BRUNCSICS, B. & SARKAR, C. A. 2019. Mechanisms of noncanonical binding dynamics in multivalent protein–protein interactions. *Proceedings of the National Academy of Sciences,* 116**,** 25659-25667.

HÄUßERMANN, K., YOUNG, G., KUKURA, P. & DIETZ, H. 2019. Dissecting FOXP2 oligomerization and DNA binding. *Angewandte Chemie,* 131**,** 7744-7749.

IMAIDE, S., RICHING, K. M., MAKUKHIN, N., VETMA, V., WHITWORTH, C., HUGHES, S. J., TRAINOR, N., MAHAN, S. D., MURPHY, N. & COWAN, A. D. 2021. Trivalent PROTACs enhance protein degradation via combined avidity and cooperativity. *Nature chemical biology,* 17**,** 1157-1167.

KIM, E. Y., SCHRADER, N., SMOLINSKY, B., BEDET, C., VANNIER, C., SCHWARZ, G. & SCHINDELIN, H. 2006. Deciphering the structural framework of glycine receptor anchoring by gephyrin. *The EMBO journal,* 25**,** 1385-95.

KITOV, P. I. & BUNDLE, D. R. 2003. On the nature of the multivalency effect: a thermodynamic model. *Journal of the American Chemical Society,* 125**,** 16271-16284.

KNEZEVIC, J., LANGER, A., HAMPEL, P. A., KAISER, W., STRASSER, R. & RANT, U. 2012. Quantitation of affinity, avidity, and binding kinetics of protein analytes with a dynamically switchable biosurface. *Journal of the American Chemical Society,* 134**,** 15225-15228.

KÖSTER, J. & RAHMANN, S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics,* 28**,** 2520-2522.

LEE, C. C., MACKAY, J. A., FRÉCHET, J. M. & SZOKA, F. C. 2005. Designing dendrimers for biological applications. *Nature biotechnology,* 23**,** 1517-1526.

LONDON, N., RAVEH, B. & SCHUELER-FURMAN, O. 2013. Druggable protein–protein interactions–from hot spots to hot segments. *Current opinion in chemical biology,* 17**,** 952-959.

LU, H., ZHOU, Q., HE, J., JIANG, Z., PENG, C., TONG, R. & SHI, J. 2020. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy,* 5**,** 1-23.

MARIANAYAGAM, N. J., SUNDE, M. & MATTHEWS, J. M. 2004. The power of two: protein dimerization in biology. *Trends in biochemical sciences,* 29**,** 618-625.

MARIC, H. M., HAUSRAT, T. J., NEUBERT, F., DALBY, N. O., DOOSE, S., SAUER, M., KNEUSSEL, M. & STRØMGAARD, K. 2017. Gephyrin-binding peptides visualize postsynaptic sites and modulate neurotransmission. *Nature Chemical Biology,* 13**,** 153-160.

MARIC, H. M., KASARAGOD, V. B., HAUSRAT, T. J., KNEUSSEL, M., TRETTER, V., STROMGAARD, K. & SCHINDELIN, H. 2014. Molecular basis of the alternative recruitment of GABA(A) versus glycine receptors through gephyrin. *Nat Commun,* 5**,** 5767.

NOMIZU, M., YAMAMURA, K., KLEINMAN, H. K. & YAMADA, Y. 1993. Multimeric forms of Tyr-Ile-Gly-Ser-Arg (YIGSR) peptide enhance the inhibition of tumor growth and metastasis. *Cancer research,* 53**,** 3459-3461.

PATCHING, S. G. 2014. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochimica et Biophysica Acta (BBA)-Biomembranes,* 1838**,** 43-55.

PAWSON, T. & NASH, P. 2003. Assembly of cell regulatory systems through protein interaction domains. *science,* 300**,** 445-452.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research,* 12**,** 2825-2830.

PONZO, I., MÖLLER, F. M., DAUB, H. & MATSCHEKO, N. 2019. A DNA-based biosensor assay for the kinetic characterization of ion-dependent aptamer folding and protein binding. *Molecules,* 24**,** 2877.

REINKING, H. K. & STINGELE, J. 2021. Protein-oligonucleotide conjugates as model substrates for DNA-protein crosslink repair proteases. *STAR Protocols,* 2**,** 100591.

ROSENBAUM, M. I., CLEMMENSEN, L. S., BREDT, D. S., BETTLER, B. & STROMGAARD, K. 2020. Targeting receptor complexes: a new dimension in drug discovery. *Nat Rev Drug Discov,* 19**,** 884-901.

SAINLOS, M., ISKENDERIAN, W. S. & IMPERIALI, B. 2009. A general screening strategy for peptide-based fluorogenic ligands: probes for dynamic studies of PDZ domain-mediated interactions. *Journal of the American Chemical Society,* 131**,** 6680-6682.

SCHRADER, N., KIM, E. Y., WINKING, J., PAULUKAT, J., SCHINDELIN, H. & SCHWARZ, G. 2004. Biochemical characterization of the high affinity binding between the glycine receptor and gephyrin. *The Journal of biological chemistry,* 279**,** 18733-41.

SCHULTE, C., KHAYENKO, V., NORDBLOM, N. F., TIPPEL, F., PECK, V., GUPTA, A. J. & MARIC, H. M. 2021. High-throughput determination of protein affinities using unmodified peptide libraries in nanomolar scale. *iScience,* 24**,** 101898.

SCHULTE, C. & MARIC, H. M. 2021. Expanding GABAAR pharmacology via receptor-associated proteins. *Current Opinion in Pharmacology,* 57**,** 98-106.

SPÄNIG, S. & HEIDER, D. 2019. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData mining,* 12**,** 1-29.

SPÄNIG, S., MOHSEN, S., HATTAB, G., HAUSCHILD, A.-C. & HEIDER, D. 2021. A large-scale comparative study on peptide encodings for biomedical classification. *NAR genomics and bioinformatics,* 3**,** lqab039.

STEIN, J. A., IANESELLI, A. & BRAUN, D. 2021. Kinetic Microscale Thermophoresis for Simultaneous Measurement of Binding Affinity and Kinetics. *Angewandte Chemie*.

SULTANA, A. & LEE, J. E. 2015. Measuring protein-protein and protein-nucleic acid interactions by biolayer interferometry. *Current protocols in protein science,* 79**,** 19.25. 1-19.25. 26.

TSOUKA, A., HOETZEL, K., MENDE, M., HEIDEPRIEM, J., PARIS, G., EICKELMANN, S., SEEBERGER, P. H., LEPENIES, B. & LOEFFLER, F. F. 2021. Probing Multivalent Carbohydrate-Protein Interactions With On-Chip Synthesized Glycopeptides Using Different Functionalized Surfaces. *Frontiers in Chemistry***,** 931.

TYAGARAJAN, S. K. & FRITSCHY, J. M. 2014. Gephyrin: a master regulator of neuronal function? *Nature reviews. Neuroscience,* 15**,** 141-56.

WALPORT, L. J., LOW, J. K., MATTHEWS, J. M. & MACKAY, J. P. 2021. The characterization of protein interactions–what, how and how much? *Chemical Society Reviews*.

WIENER, J., KOKOTEK, D., ROSOWSKI, S., LICKERT, H. & MEIER, M. 2020. Preparation of single-and double-oligonucleotide antibody conjugates and their application for protein analytics. *Scientific reports,* 10**,** 1-11.

WOOLDRIDGE, L., LISSINA, A., COLE, D. K., VAN DEN BERG, H. A., PRICE, D. A. & SEWELL, A. K. 2009. Tricks with tetramers: how to get the most from multimeric peptide–MHC. *Immunology,* 126**,** 147-164.

ZHANG, G., BROWN, J. S., QUARTARARO, A. J., LI, C., TAN, X., HANNA, S., ANTILLA, S., COWFER, A. E., LOAS, A. & PENTELUTE, B. L. 2022. Rapid de novo discovery of peptidomimetic affinity reagents for human angiotensin converting enzyme 2. *Communications Chemistry,* 5**,** 1-10.

7
Discussion

Various aspects of antimicrobial resistance (AMR) have been approached in the present dissertation. In particular, waste- and freshwater-based monitoring of AMR and prediction of host-defense peptides (HDPs), specifically antimicrobial peptides (AMPs), have been covered. An initial introduction on AMR and AMP provided the basics for the remaining parts. The study's goal was to illustrate general aspects of AMR, tools to examine the dissemination, and the role of computational methods to replace conventional antibiotics. To this end, we conducted a comprehensive literature search to demonstrate the particular topics' diversity, results, and challenges. We contributed a survey on European freshwater lakes to point out the advantages of standardized examination of multi-omic datasets. After shedding light on the relevance of AMR and its dissemination, the necessity of alternative strategies to tackle multi-resistant pathogens is evident.

Standardization is also the primary concern of the machine learning (ML) section. The great variety of literature-known peptide encodings, models, and applications requires integrative workflows. Moreover, users should not be bothered with the complexity of such workflows, such as hyper-parameter optimization. Consequently, we contributed multiple studies, ranging from peptide encodings and their performance across biomedical domains to an unsupervised selection and ensemble classifier configuration. As a proof-of-concept, we applied our methods to predict binding affinities of modified peptides. The easy transformation to other biomedical domains stresses the benefit of an integrative approach and demonstrates the significance of ML for peptidomics.

The integration of multi-omic datasets, followed by the computational analysis of various modalities, is only one aspect. AMR, environmental epidemiology (EE), AMPs, and ML are very complex topics per se. Researchers must consider the nuances when addressing particular issues. In the following, we discuss critical aspects, highlight potential weaknesses, and suggest further research directions.

7.1 Antimicrobial Resistance

Several antibiotics are already of critical importance (see Table 2.1), and administration must comply with prioritization criteria to hamper resistance forming[250]. The significance of certain antimicrobial agents is in concert with the tremendous impact resulting from novel insensitivity acquirement of bacteria. Jian *et al.* (2021) stressed the economic context of AMR by summarizing associated costs and deaths until 2017, many thereof recurring annually[115]. The World Health Organization (WHO) designated various multi-resistant species as critical pathogens[115]. Prioritization of antibiotics research is therefore required to mitigate the consequences of AMR[115]. High-priority species, such as *Acinetobacter baumannii*, developed resistance against carbapenems or third-generation cephalosporins[115]. An effective measure to control future AMR dissemination in the environment concerns the inactivation of biological waste[115]. Jian *et al.* (2021) referred to Selvam *et al.* (2012), who has observed that simple composting significantly reduced levels of antimicrobial resistance genes (ARGs)[207].

The integration of social factors could further support comprehension of AMR spread[141]. Li *et al.* (2021) related ecological conditions, socio-economic factors, and antibiotic consumption with AMR prevalence[141]. The study examined the infantile intestinal tract and revealed that significant rates of *Escherichia coli* in combination with an underdeveloped microbiome drives AMR of infants[141]. The authors observed diverse environmental factors affecting the distribution of ARGs within a child's gut (α-diversity) and ARG changes between the children (β-diversity), thereby determining potential AMR drivers[118,141].

VanOeffelen *et al.* (2021) stressed the significance of the large-scale integration of genome data and related antibiotic sensitivity[234]. The authors collected 67,000 genomes comprising more than 100 species and demonstrated the potential of ML guided resistance mechanism prediction[234]. VanOeffelen *et al.* (2021) used, among others, the Comprehensive Antibiotic Resistance Database (CARD) as a source for ARGs[3,188].

To illuminate multiple facets of AMR, researchers should integrate environmental and socio-economic factors from multi-omic resources, such as metagenome and whole-genome sequencing (WGS) data[141,234]. Relating those studies with public resources on ARGs, such as the CARD[3], the detection of resistance mechanisms can be advanced. The benefit of various facets on AMR is in accordance with our large-scale study on AMR in European freshwater lakes[221]. The integration of 16S rRNA amplicon sequencing data, metagenomes, and statis-

tics on socio-economic factors, in particular, surrounding agriculture, enabled quantification of the current AMR burden and screening of future progression[221].

7.2 Environmental Epidemiology

A potential drawback of metagenome-based AMR analysis is the overestimating of ARGs[87]. Spänig *et al.* (2021) argued that based on their data, a distinction between chromosomal- and plasmid-encoded resistance genes is challenging[221]. This follows Gupta *et al.* (2020), who stressed the importance of reconsidering such findings thoroughly[87]. First, it is unclear whether the ARG origin is located on a mobile genetic element (MGE); hence, facilitating readily dissemination[87]. Second, the ARG could belong to strain, natural in the examined environment, ultimately aggravating conclusions about AMR dissemination trajectories[87]. To mitigate these issues, Gupta *et al.* (2020) also provided a standardized workflow to survey the resistome prospectively, comprising DNA preprocessing, read mapping, and ARG detection[87]. The authors suggested integrating long and short metagenomic reads using OPERA-MS to enable the determination of MGEs[17].

Moreover, various tools assess the risk of ARGs by incorporating the clinical relevance and location[87]. In this way, intrinsic resistance can be distinguished from putatively acquired AMR[87]. To attribute ARGs to the host strain, several studies correlated the genomic context and AMR determinants[87]. However, Gupta *et al.* (2020) referred to potential false-positive predictions and urged scrutinizing such findings carefully[87].

The comparison with other environments by the inclusion of different sampling sites is crucial[221]. Researchers should integrate additional metagenome datasets from public repositories to enhance the validity of the results[37]. Chen *et al.* (2019) collected and examined specimens in Lake Tai (China) and additionally integrated metagenome data from different countries, for instance, Australia[37]. The comprehensive approach enabled the authors to determine substantially higher AMR pollution in Lake Tai[37].

Qu *et al.* (2019) highlighted the threat of multi-resistant pathogens in the future, specifically in China, and referred to it as the "post-antibiotic era"[190]. The authors pointed out the dramatic consequences of AMR, ultimately resulting in decreased life expectancy[190]. The Chinese government already implemented several measures to mitigate implications, for instance, the reduction of antibiotic administration in medicine or livestock farming[190]. In addition, Qu *et al.* (2019) suggested incorporating additional multi-modal datasets on a large scale, comprising WGS and AMR profiles of health care institutions[190]. According to the authors, such "big data" approaches could reveal unknown interaction and AMR transmission paths[190]. Finally, Qu *et al.* (2019) recommended a broader application of artificial intelligence for targeted treatment[190].

Arango-Argoty *et al.* (2018) developed a Deep Learning (DL) model to classify ARGs in genomic datasets[5]. In this study, DNA fragments or complete genes have been encoded using the similarity to known ARGs as features[5]. The classifier returns the probability of an input gene being part of the investigated resistance categories[5]. The model achieved comparable or even higher performance than a "best hit" approach utilizing database queries; however, the fast prediction significantly decreases ARG annotation time for future studies[5].

The fast evaluation of sequencing data could increase the experimental throughput; thus, potentially ensuring the topicality of AMR. However, the processed data must also be accessible and easily understood[8]. To this end, governmental institutions and research groups have already developed interactive AMR dashboards. Dashboards provide related overview graphics and tabular data, allowing multiple views of different aspects. Non-technical users, such as healthcare stakeholders, can quickly grasp the current threat of AMR.

The dashboard by the European Centre for Disease Prevention and Control (ECDC) visualizes resistance against various antibiotic drug classes Europe-wide and per country*. In addition, the Center For Disease Dynamics, Economics & Policy (CDDEP) integrated global data on AMR from 2000 to 2015[123] into the ResistanceMap†. The ResistanceMap features resistance levels visualizations of pathogens and the progress of insensitivity concerning diverse antimicrobials. Furthermore, Stedtfeld *et al.* (2016) collected samples from 43 locations, associated global data on AMR from public studies, and integrated them all in an interactive dashboard[224]. The application denoted as the "antimicrobial resistance dashboard" can be utilized to fetch information about multiple aspects of AMR, including dissemination and endemic clusters[224].

The urgency of successive data generation and integration is evident. Large-scale examination of AMR, for instance, through standardized sampling and evaluation of various European freshwater lakes, ensures data integrity[221]. In addition, such studies enable retrospective tracing of AMR, as conducted by Schar *et al.* (2021)[203]. The authors reviewed multiple studies on AMR and collected the respective data[203]. Schar *et al.* (2021) interpolated the locations; thus, enabling continuous AMR prediction for places lacking actual samples[203]. As mentioned above, fast screening of genomic data reduces the computational demand; therefore, speeding up the AMR integration[5]. Legacy AMR dashboards additionally stress the significance of updated results. According to their manuals, the ECDC relies on data from 2015 and the CDDEP dashboard from 2017. Thus, emphasizing the demand for a novel approach for continuous data integration, analysis, and visual processing[68].

Up-to-date, readily accessible, visually depicted data on AMR could also support clinicians in their antibiotics administration. The University of Bern provides the INterface For Empirical

*`https://multimedia.efsa.europa.eu/dataviz-2015/index.htm`, accessed January 21, 2022
†`https://resistancemap.cddep.org/index.php`, accessed January 21, 2022

antimicrobial ChemoTherapy (INFECT) on AMR in Switzerland[‡]. INFECT provides information on various pathogens, including statistics about the ascertained susceptibility and the number of isolates. Moreover, according to the manual, INFECT is updated monthly. The data of INFECT is directly acquired from patients[§]. Thus, similar, steadily updated resources are required incorporating EE data since infirmary sewage is one of the main drivers for AMR development[20,100,178]. Ad-hoc notification of physicians might positively impact their decisions for alternative treatments.

7.3 Host-Defense Peptides

HDPs, specifically AMPs, are a vital drug class and possess broad usability beyond pharmaceutical application[26]. In humans, AMPs are expressed in various organs, for instance, the brain, intestinal tract, or skin[239]. Metabolites, including amino acids, vitamins, or acids, improve the therapeutic application of AMPs, fostering alternative therapies for infectious diseases[239]. Nevertheless, for clinical relevance, pharmacokinetics and efficacy must be greatly improved[70]. Thus, Fjell *et al.* (2012) illustrated limitations of AMP activity studies based on the primary sequence[71]. In particular, studies on structure-based peptide-membrane interaction, such as molecular docking, benefit from complex interactions of amino acids in higher dimensions[71]. Moreover, the optimization of physicochemical properties, peptide length, and hydrophobic moment for increasing selectivity and reducing toxicity is crucial[106]. However, these parameters influence each other, which aggravates a measurement of individual influences on antimicrobial activity[106].

Although resistance to AMPs is unlikely, some species were able to adapt using multiple AMP defense measures[10,71]. One countermeasure is chemical alteration of the outer lipid bi-layer, ultimately mitigating the AMP's cell wall selectivity[10]. In particular, modification of phospholipids, such as acylation, mitigates peptide accumulation[10]. The affinity of AMPs can be also reduced by polar proteins inserted in the outer leaflet, such that the interaction between the polar face of the AMP with the membrane is inhibited[10]. Some species release non-polar proteins, which covalently bind to the positive charged AMPs; hence, resulting in the inactivation of the peptide[10]. Efflux pumps can be utilized to remove AMPs from the cell[10,117]. Finally, intracellular peptide digestion through proteolytic proteins has also been observed[10].

Nevertheless, Wimley *et al.* (2011) pointed out that different experiments, considering certain molecular aspects of AMPs, could unveil specific details on the mode of action[248]. Consequently, it is uncertain whether an individual effect is observed due to a particular condition or if different modes of action exist[248]. The significance of translocation is also partially

[‡]`https://infect.info/`, accessed January 22, 2022

[§]`https://www.anresis.ch/antibiotic-resistance/laboratories/`, accessed January 22, 2022

understood since it is unclear whether it is a coincidental process owing to a by-product of membrane lysis or whether intracellular components are indeed the actual target[248]. Moreover, Lazzaro *et al.* (2020) indicated the potential involvement of various endogenous factors, which can be hardly considered experimentally, resulting in an antimicrobial response[136].

Besides cationic peptides, anionic AMPs demonstrated potential[94]. Anionic antimicrobials are expressed in various species, including vertebrates, invertebrates, and plants[94]. The sources for AMPs are numerous, and researchers can adopt a great variety of parameters for activity and pharmacokinetics. Nevertheless, synthesis remains, considering around 400 $ per gram, expensive[50]. Albeit biological synthesis, such as insertion of an AMP-encoding gene in microbial DNA, followed by large-scale breeding and purification of the proteome, chemical synthesis, including solid-phase synthesis, and premature biological production ensued by a chemical synthesis step can be employed, further research is necessary to reduce the economic impact[50]. However, the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) simplified AMP admission for a faster transition into clinical practice[72]. Accordingly, Rathinakumar *et al.* (2009) suggested extending the focus beyond broad-spectrum peptides to AMPs with targeted activity[191].

Lazar *et al.* (2018) underpinned that parallel administration of AMPs, and conventional antibiotics could enhance the susceptibility[135]. The dual strategy reduced resistance development since the cell membrane adapted biochemically to the treatment[135]. In particular, the findings suggested that on environmental pressure, the gene expression pattern changes, resulting in an altered AMP susceptible phenotype[135]. Wiesner and Vilcinskas (2010) have also acknowledged the interaction of conventional antibiotics and AMPs[246]. The authors noted that the silenced gene encoding for the human θ-defensin could be reactivated by aminoglycoside application[246]. Since θ-defensin possesses antiviral effects, this is highly beneficial for human immunodeficiency virus (HIV) treatment[246]. However, if pathogens develop resistance to human-derived AMPs, the innate immune system could be undermined[136].

7.4 Machine Learning

Various studies have been published concerning the prediction of multiple peptide characteristics. In these studies, researchers are faced with many encodings, further complicating the hyper-parameter optimization. Thus, Feurer *et al.* (2015) highlighted the benefit of automated ML pipelines, specifically for non-technical users[69]. Concerning biomedicine, wet-lab scientists are encouraged to conduct computational experiments autonomously. Fortunately, several packages already provide ready access to encoding and ML algorithms[21,142,210]. A graphical interface is a further step towards user-friendly biomedical ML workflows[38]. Albeit the selection is simplified, the sheer choice of the encodings remains overwhelming and requires extensive preprocessing work[220,223]. Our recently published encoding benchmark

provides semi-automated encoding selection[223]. This work has been continued by a study on unsupervised encoding selection, further paving the way for automated solutions in the biomedical domain. However, more research is necessary to achieve similar automation levels as, for instance, auto-sklearn[69].

Encodings contribute, in addition to the ML model, significantly to the performance[223]. Accordingly, we discovered that similar encoding configurations result in a similar performance; nevertheless, raising the question about the impact of distant values[223]. Moreover, some parameters depend on the sequence length; short peptides require a fixed gap or the window length. Consequently, a variable parameter space, thus, the interaction of non-adjacent amino acids, resulting in high accuracy, could possess biological meaning. Additional research in this direction might be fruitful.

The diversity of individual models and encodings, accompanied by the mutual compensation of misclassified instances, remains an open research topic. Although Kuncheva (2014) pointed out that maximizing the diversity alone is insufficient to compile efficient ensembles, it is, however, a crucial property[129]. Moreover, Heider *et al.* (2014) argued that different encodings reflect various characteristics of the amino acid sequences; consequently, suggested the diversity as a measure for selecting relevant encodings[97,220]. However, Spänig *et al.* (2021) detected a minor impact of the diversity on the class separability; hence, performance[223]. More research is required to examine to which extent encodings can participate in the final prediction.

Spänig *et al.* (2019) suggested the Disagreement Measure D[220], which describes the mean false positive and false negative rates of the base classifiers[129]. In contrast, Spänig *et al.* (2021) employed the Interrater Agreement κ[223], including the entire confusion matrix[129]. Kuncheva has described additional metrics (2014)[129]. Thus, the versatility of diversity measures demands further studies on this topic. Consequently, diversity metrics are conceivable, which acknowledge the specific requirements of the encoding field, for instance, utilizing penalization, weights, or the consensus with the true class. More research in this direction is of utmost significance, as the diversity might unveil biological contributions of individual encodings.

Furthermore, we demonstrated that sequence-based encodings (SeBEs) and structure-based encodings (StBEs) achieved good performance, although the former group is in general superior[223]. The inferiority of StBEs is noteworthy since the protein structure determines its purpose[237]. According to Spänig *et al.* (2021), the results can be explained by the low agreement between estimated and actual structures[223]. Additional investigations should address the accuracy of the structure approximation and the root-mean-square deviation from model structures.

The structure approximation could also be flawed due to a Basic Local Alignment Search Tool (BLAST) parameter misinterpretation. We used max_target_seqs to determine the num-

ber of hits to be returned[223]. This parameter defines the number of hits exceeding the minimum e-value, which are not necessarily the best matches[211].

The long run-time of *ab initio* structure prediction[223] could be evaded by tools, specifically tackling peptides[132]. The Collection of Antimicrobial Peptides (CAMP) hosts structures of several hundred AMPs[238]. A novel dataset could be constructed utilizing known structures complemented by negative sequences with confirmed three-dimensional conformations for future studies. However, according to Section 5.1, scientists must address the intersection of the classes. Ultimately, such a benchmark dataset could provide an unbiased view of the relation between SeBEs and StBEs.

Further studies should also consider the structure resolution. Snyder *et al.* (2005) referred to the complementary of X-ray crystallography and NMR spectroscopy; thus, the resulting structure is potentially incomplete, depending on the approach[218]. As of January 2022, the CAMP contains 757 AMP structures, thereof approx. 55 % NMR spectroscopy-derived. Thus, the structure determination technology additionally complicates the creation of a benchmark dataset. This aspect could also impair the findings of Spänig *et al.* (2021)[223]. In summary, albeit structure approximation is tentatively sufficient[223], researchers must consider various aspects to improve StBEs.

In addition, Burdukiewicz *et al.* (2021) referred to the issue of inconsistent strategies for constructing negative test data[28]. The authors stressed the consequences on benchmark studies if negative sequences, such as non-AMPs, are randomly selected[28]. The dataset disparity also concerns the encoding benchmark of Spänig *et al.* (2021)[223]. Since the study included datasets from the same domain, researchers should address the impact of negative data sampling in the future.

Moreover, Bourgade *et al.* (2014) identified antimicrobial activity of *β*-amyloid peptides[23]. A pathological accumulation of these peptides in the human brain leads to Alzheimer's disease[23]. The observed pleiotropy of *β*-amyloid peptides questions the reliability of negative datasets fundamentally. If it is unclear whether putative non-AMPs are indeed non-effective, the model training is heavily impaired.

A countermeasure could be Positive-Unlabeled learning (PUL)[14]. According to Bekker and Davis (2020), PUL employs, besides the labeled positive instances, unlabeled data, hence, a second group, where the class membership is unknown[14]. In the simplest case, the labeled positive data is randomly and uniformly gathered from the actual positive data, the classes are well-separable, and unlabeled instances approximate their real class[14]. If the preconditions are fulfilled, users can conduct the training by first assigning "reliable negative examples", using, for instance, the k-nearest neighbors algorithm[14]. For the actual training, Bekker and Davis (2020) suggested various specialized algorithms[14]. Nevertheless, putative negative and positive peptides ultimately require *in vitro* verification to rule out any uncertainty. With that being said, several working groups successfully predicted, synthesized, and validated

HDPs, neglecting the issue of unlabeled data[78,177,204].

An increasing emphasis has been also put on the interpretability of ML algorithms[60]. For instance, DL models convince with outstanding performance, accompanied with a complex architecture, which hampers decision-making comprehension[60]. Specifically for amino acid sequences, sophisticated methods to evaluate the decision on the amino acid level are lacking. In particular, for complex encodings comprising multiple, non-adjacent amino acids, more details would be helpful. Interpretable artificial intelligence enables researchers to verify biological observations or establish novel hypotheses. In this light, Heider *et al.* (2014) developed a model which confirmed the 11/25 rule[97]. The rule states that positively charged residues at the eleventh and twenty-fifth position influence HIV co-receptor tropism[97]. The authors conceded that the respective features solely represent the electrostatic potential close to the actual amino acids[97]. Ultimately, the 11/25 rule would remain undiscovered without prior knowledge since the decision is not attributable to the amino acids.

8
Conclusion

This dissertation covered four core topics on multiple aspects of antimicrobial resistance (AMR) based on a comprehensive literature review. The first chapter introduced AMR, specifically microbial resistance mechanisms, the background of dissemination events, and the characteristics of antibiotic drug classes. Afterward, environmental epidemiology (EE), particularly the state-of-the-art concerning waste- and freshwater-based studies, was presented. As a proof of concept, we examined multiple European freshwater lakes to pave the way for standardized data integration[221]. The results revealed baseline levels of AMR[221], underpinning the significance of alternative therapies. Thus, the third chapter covered host-defense peptides (HDPs), specifically, antimicrobial peptides (AMPs). In addition, the working principles of AMPs are illuminated. The chapter concluded with a survey on multiple applications and the challenge and the capability of abiotic modification. Years of AMP research generated numerous amino acid sequences and structures, enabling the application of artificial intelligence. Thus, the last chapter introduced all aspects of machine learning (ML), including data repositories, amino acid encodings, biomedical domains, and concluded with the respective publications contributed to this dissertation.

AMR, EE, and AMPs are the essential data resources discussed. AMR has been outlined utilizing the current literature; however, the review demonstrates the importance of genome sequencing and susceptibility testing. Various studies investigated EE, specifically stressing metagenome sampling and molecular diagnosis. Antimicrobial peptidomics is the foundation of ML-based prediction of novel antibiotics. Ultimately, the diversity hampers streamlined

integration of multi-omic datasets, which is expected due to the different biological descent.

In this light, the thesis revealed the necessity of standardized datasets. Several studies on AMR in waste- and freshwaters, already initialized crucial attempts, for instance, Czekalski *et al.* (2015)[49], Schar *et al.* (2021)[203], and Spänig *et al.* (2021)[223]. Undoubtedly, the multi-modal character of the studies requires the collection of diverse data; thus, completely unified integration might be practically impossible. However, researchers should globally agree on a minimum standardization, which allows at least to set the initial baseline level of AMR and future monitoring.

Concerning AMPs, researchers pointed out that the variety of positive and negative training data aggravates the comparison of ML models. Thus, it is of utmost importance to define a gold-standard test dataset. The dataset could be employed to verify the integrity of new models. An independent consortium could contribute additional AMPs if novel sequences have been discovered. In perspective, the central repository could follow the principles of the UCI Machine learning repository by providing standardized peptide datasets to assess new models[62]. These datasets would enable researchers an unbiased evaluation beyond model performance, comprising encoding accuracy and the utility of unsupervised encoding selection.

A recent study confirmed the tremendous negative impact of AMR on health care[176]. In particular, Murray *et al.* (2022) observed that in 2019 almost two million deaths were due to ineffective antibiotics[176]. Major involved pathogens included four ESKAPE species as well as *Escherichia coli* and *Streptococcus pneumoniae*[176]. Millions of deaths are implicitly related to AMR[176]. These figures again stress effective countermeasures in the future, including prediction, development, and introduction of novel AMPs into clinical practice.

List of Tables

List of Abbreviations

Miscellaneous

AMP	antimicrobial peptide
ACP	anticancer peptide
AMP	antimicrobial peptide
AMR	antimicrobial resistance
ANOVA	Analysis of Variance
ARG	antimicrobial resistance gene
ARO	Antibiotic Resistance Ontology
AUC	area under the curve
BLAST	Basic Local Alignment Search Tool
BLOSUM	blocks substitution matrix
CDDEP	Center For Disease Dynamics, Economics & Policy
COVID-19	coronavirus disease 2019
DNA	deoxyribonucleic acid
ECDC	European Centre for Disease Prevention and Control
EE	environmental epidemiology
EMA	European Medicines Agency
ESKAPE	Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter spp.
EU	European Union
FAIR	findable, accessible, interchangeable, and reproducible
FBE	freshwater-based epidemiology
FDA	Food and Drug Administration

FPS	fluorescence proximity sensing
HDP	host-defense peptide
HGT	horizontal gene transfer
HIV	human immunodeficiency virus
HPLC	High Performance Liquid Chromatography
HPLC-MSMS	HPLC Tandem Mass Spectrometry
IL17	interleukin-17
INFECT	INterface For Empirical antimicrobial ChemoTherapy
LC-MS/MS	Liquid Chromatography - tandem Mass Spectrometry
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight
MANOVA	Multivariate Analysis of Variance
MCC	Matthews correlation coefficient
MGE	mobile genetic element
MIC	minimum inhibitory concentration
ML	machine learning
MODI	Modelability Index
MRSA	multi-resistant *Staphylococcus aureus*
MVO	multi-verse optimizer
omp	outer-membrane protein
OTU	Operative Taxonomic Unit
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
PCR	polymerase chain reaction
PPI	protein-protein interaction
PSI-BLAST	Position-Specific Iterative BLAST
PSSM	position-specific scoring matrix
qPCR	quantitative PCR
QSP	quorum-sensing peptide

RGI	Resistance Gene Identifier
RNA	ribonucleic acid
RND	resistance-nodulation-division
rRNA	ribosomal RNA
SARS-CoV-2	severe acute respiratory syndrome coronavirus type 2
TMP	trimethoprim
WBE	wastewater-based epidemiology
WGS	whole-genome sequencing
WHO	World Health Organization
WWTP	wastewater treatment plant
***α*-MSH**	*α*-melanocyte-stimulating hormone

Machine Learning Algorithms

ANN	Artificial Neural Network
BRT	Boosted Regression Tree
CNN	Convolutional Neural Network
DL	Deep Learning
DTC	Decision Tree classifier
ECC	Ensemble Classifier Chain
LASSO	Least Absolute Shrinkage and Selection Operator
LRC	Logistic Regression classifier
NBC	Naïve Bayes classifier
PUL	Positive-Unlabeled learning
RFC	Random Forest classifier
RFR	Random Forest regressor
SVM	Support Vector Machine

Encodings

AAC	amino acid composition
AAP	amino acid pairs
CGR	chaos game representation
CTDC	composition/transition/distribution-composition
DPC	dipeptide composition
GPC	g-gap dipeptide composition
MoBE	model-based encoding
PAAC	pseudo amino acid composition
PKRAAC	pseudo k-tuple reduced amino acids composition
QSAR	quantitative structure-activity relationship
SeBE	sequence-based encoding
StBE	structure-based encoding

Databases

APD	Antimicrobial Peptide Database
CAMP	Collection of Antimicrobial Peptides
CARD	Comprehensive Antibiotic Resistance Database
PDB	Protein Data Bank
LAMP	Linking Antimicrobial Peptides Database

References

[1] Abramović, B. F., Uzelac, M. M., Armaković, S. J., Gašić, U., Četojević Simin, D. D., & Armaković, S. (2021). Experimental and computational study of hydrolysis and photolysis of antibiotic ceftriaxone: Degradation kinetics, pathways, and toxicity. *Science of the Total Environment*, 768.

[2] Al-Madi, N., Faris, H., & Mirjalili, S. (2019). Binary multi-verse optimization algorithm for global optimization and discrete problems. *International Journal of Machine Learning and Cybernetics*, 10(12), 3445–3465.

[3] Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. L. V., Cheng, A. A., Liu, S., et al. (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 48, D517–D525.

[4] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25.

[5] Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6.

[6] Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Heijne, G. V., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2.

[7] Aronica, P. G., Reid, L. M., Desai, N., Li, J., Fox, S. J., Yadahalli, S., Essex, J. W., & Verma, C. S. (2021). Computational Methods and Tools in Antimicrobial Peptide Research. *Journal of Chemical Information and Modeling*.

[8] Arya, B. K., Robert, D., Bhattacharya, S. D., & Mukhopadhyay, J. (2013). A framework for web based geographical information systems for country wide antimicrobial resistance monitoring. *Health Policy and Technology*, 2, 85–93.

[9] Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., et al. (2021). Self-driving cars: A survey. *Expert Systems with Applications*, 165, 113816.

[10] Bahar, A. A. & Ren, D. (2013). Antimicrobial peptides. *Pharmaceuticals*, 6(12), 1543–1575.

[11] Bala, S., Khanna, R., Dadhwal, M., Prabagaran, S. R., Shivaji, S., Cullum, J., & Lal, R. (2004). Reclassification of Amycolatopsis mediterranei DSM 46095 as Amycolatopsis rifamycinica sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 54, 1145–1149.

[12] Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., et al. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49, D480–D489.

[13] Bednarska, N. G., Wren, B. W., & Willcocks, S. J. (2017). The importance of the glycosylation of antimicrobial peptides: Natural and synthetic approaches. *Drug Discovery Today*, 22, 919–926.

[14] Bekker, J. & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109, 719–760.

[15] Berg, J. M., Tymoczko, J. L., Gatto jr Gregory J., & Stryer, L. (2018). *Stryer Biochemie 8. Auflage*, chapter 2,3, (pp. 36–40,96). Springer Spektrum, 8 edition.

[16] Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: A Python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, 8.

[17] Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., Dvornicic, M., Soldo, J. P., Koh, J. Y., Tong, C., et al. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology*, 37, 937–944.

[18] Bialvaei, A. Z. & Kafil, H. S. (2015). Colistin, mechanisms and prevalence of resistance. *Current Medical Research and Opinion*, 31, 707–721.

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

[20] Bojar, B., Sheridan, J., Beattie, R., Cahak, C., Liedhegner, E., Munoz-Price, L. S., Hristova, K. R., & Skwor, T. (2021). Antibiotic resistance patterns of Escherichia coli isolates from the clinic through the wastewater pathway. *International Journal of Hygiene and Environmental Health*, 238, 113863.

[21] Bonidia, R. P., Domingues, D. S., Sanches, D. S., & de Carvalho, A. C. P. L. F. (2021). MathFeature: Feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. *Briefings in Bioinformatics*.

[22] Bose, P. & Harrison, R. W. (2011). Encoding protein structure with functions on graphs. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, (pp. 338–344).

[23] Bourgade, K., Garneau, H., Giroux, G., Page, A. Y. L., Bocti, C., Dupuis, G., Frost, E. H., & Fülöp, T. (2015). β-Amyloid peptides display protective activity against the human Alzheimer's disease-associated herpes simplex virus-1. *Biogerontology*, 16, 85–98.

[24] Bradford, P. A. (2001). Extended-spectrum *β*-lactamases in the 21st century: Characterization, epidemiology, and detection of this important resistance threat. *Clinical Microbiology Reviews*, 14(4), 933–951.

[25] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

[26] Brogden, N. K. & Brogden, K. A. (2011). Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals? *International Journal of Antimicrobial Agents*, 38(3), 217–225.

[27] Bueno, I., Travis, D., Gonzalez-Rocha, G., Alvarez, J., Lima, C., Benitez, C. G., Phelps, N. B., Wass, B., Johnson, T. J., Zhang, Q., et al. (2019). Antibiotic Resistance Genes in Freshwater Trout Farms in a Watershed in Chile. *Journal of Environmental Quality*, 48, 1462–1471.

[28] Burdukiewicz, M., Sidorczuk, K., Gagat, P., Pietluch, F., Kała, J., Rafacz, D., Bakala, M., Slowik, J., Kolenda, R., Rödiger, S., et al. (2021). The impact of negative data sampling on antimicrobial peptide prediction. German Conference on Bioinformatics. Oral presentation.

[29] Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Costanzo, L. D., Duarte, J. M., et al. (2021). RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49, D437–D451.

[30] Cao, D. S., Liang, Y. Z., Yan, J., Tan, G. S., Xu, Q. S., & Liu, S. (2013). PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *Journal of Chemical Information and Modeling*, 53, 3086–3096.

[31] Cederlund, A., Gudmundsson, G. H., & Agerberth, B. (2011). Antimicrobial peptides important in innate immunity. *FEBS Journal*, 278(20), 3942–3951.

[32] Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107, 1477–1494.

[33] Chakraborty, J., Sapkale, V., Rajput, V., Shah, M., Kamble, S., & Dharne, M. (2020). Shotgun metagenome guided exploration of anthropogenically driven resistomic hotspots within Lonar soda lake of India. *Ecotoxicology and Environmental Safety*, 194.

[34] Chandra, N. & Kumar, S. (2017). Antibiotics Producing Soil Microorganisms. In M. Hashmi, V. Strezov, & A. Varma (Eds.), *Antibiotics and Antibiotics Resistance Genes in Soils*, volume 51 (pp. 1–18). Springer.

[35] Charles, P. E., Dalle, F., Aube, H., Doise, J. M., Quenot, J. P., Aho, L. S., Chavanet, P., & Blettery, B. (2005). Candida spp. colonization significance in critically ill medical patients: A prospective study. *Intensive care medicine*, 31, 393–400.

[36] Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). BERT4Bitter: A bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*, 37, 2556–2562.

[37] Chen, H., Jing, L., Yao, Z., Meng, F., & Teng, Y. (2019). Prevalence, source and risk of antibiotic resistance genes in the sediments of Lake Tai (China) deciphered by metagenomic assembly: A comparison with other global lakes. *Environment International*, 127, 267–275.

[38] Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R. J., Webb, G. I., Zhao, Q., et al. (2021). iLearnPlus: A comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Research*.

[39] Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K. C., & Song, J. (2018). iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34, 2499–2502.

[40] Cho, S., Barrett, J. B., Frye, J. G., & Jackson, C. R. (2020). Antimicrobial Resistance Gene Detection and Plasmid Typing Among Multidrug Resistant Enterococci Isolated from Freshwater Environment. *Microorganisms*, 8, 1–15.

[41] Choi, P. M., Tscharke, B., Samanipour, S., Hall, W. D., Gartner, C. E., Mueller, J. F., Thomas, K. V., & O'Brien, J. W. (2019). Social, demographic, and economic correlates of food and chemical consumption measured by wastewater-based epidemiology. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 21864–21873.

[42] Choi, P. M., Tscharke, B. J., Donner, E., O'Brien, J. W., Grant, S. C., Kaserzon, S. L., Mackie, R., O'Malley, E., Crosbie, N. D., Thomas, K. V., et al. (2018). Wastewater-based epidemiology biomarkers: Past, present and future. *TrAC - Trends in Analytical Chemistry*, 105, 453–469.

[43] Choi, U. & Lee, C. R. (2019). Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia coli. *Frontiers in Microbiology*, 10.

[44] Chung, C. R., Jhong, J. H., Wang, Z., Chen, S., Wan, Y., Horng, J. T., & Lee, T. Y. (2020). Characterization and identification of natural antimicrobial peptides on different organisms. *International Journal of Molecular Sciences*, 21.

[45] Coin, I., Beyermann, M., & Bienert, M. (2007). Solid-phase peptide synthesis: From standard procedures to the synthesis of difficult sequences. *Nature Protocols*, 2, 3247–3256.

[46] Coskun, M., Ucar, A., Yildirim, O., & Demir, Y. (2017). Face recognition based on convolutional neural network. In *Proceedings of the International Conference on Modern Electrical and Energy Systems, MEES 2017*, volume 2018-January (pp. 376–379).: Institute of Electrical and Electronics Engineers Inc.

[47] Cox, G. & Wright, G. D. (2013). Intrinsic antibiotic resistance: Mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology*, 303(6-7), 287–292.

[48] Cunningham, J. M., Koytiger, G., Sorger, P. K., & AlQuraishi, M. (2020). Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature Methods*, 17, 175–183.

[49] Czekalski, N., Sigdel, R., Birtel, J., Matthews, B., & Bürgmann, H. (2015). Does human activity impact the natural antibiotic resistance background? Abundance of antibiotic resistance genes in 21 Swiss lakes. *Environment International*, 81, 45–55.

[50] da Costa, J. P., Cova, M., Ferreira, R., & Vitorino, R. (2015). Antimicrobial peptides: An alternative for innovative medicines? *Applied Microbiology and Biotechnology*, 99(5), 2023–2040.

[51] Dahouda, M. K. & Joe, I. (2021). A Deep-Learned Embedding Technique for Categorical Features Encoding. *IEEE Access*, 9, 114381–114391.

[52] Damiati, S. A., Alaofi, A. L., Dhar, P., & Alhakamy, N. A. (2019). Novel machine learning application for prediction of membrane insertion potential of cell-penetrating peptides. *International Journal of Pharmaceutics*, 567.

[53] Dantas, G. & Sommer, M. O. (2012). Context matters - the complex interplay between resistome genotypes and resistance phenotypes. *Current Opinion in Microbiology*, 15, 577–582.

[54] Daughton, C. G. (2020). Wastewater surveillance for population-wide COVID-19: The present and future. *Science of the Total Environment*, 736.

[55] de Jesus, A. J. & Allen, T. W. (2013). The role of tryptophan side chains in membrane protein anchoring and hydrophobic mismatch. *Biochimica et Biophysica Acta - Biomembranes*, 1828, 864–876.

[56] Deslouches, B., Phadke, S. M., Lazarevic, V., Cascio, M., Islam, K., Montelaro, R. C., & Mietzner, T. A. (2005). De Novo Generation of Cationic Antimicrobial Peptides : Influence of Length and Tryptophan Substitution on Antimicrobial Activity. *ANTIMICROBIAL AGENTS AND CHEMOTHERAPY*, 49(1), 316–322.

[57] Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14, 927–930.

[58] Dong, G. F., Zheng, L., Huang, S. H., Gao, J., & Zuo, Y. C. (2021). Amino Acid Reduction Can Help to Improve the Identification of Antimicrobial Peptides and Their Functional Activities. *Frontiers in Genetics*, 12.

[59] Dong, J., Yao, Z. J., Zhang, L., Luo, F., Lin, Q., Lu, A. P., Chen, A. F., & Cao, D. S. (2018). PyBioMed: A python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *Journal of Cheminformatics*, 10.

[60] Du, M., Liu, N., & Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, 63, 68–77.

[61] Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., & Liu, S. Q. (2016). Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences*, 17.

[62] Dua, D. & Graff, C. (2017). UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, accessed January 27, 2022. University of California, Irvine, School of Information and Computer Sciences.

[63] EFSA (European Food Safety Authority) & ECDC (European Centre for DiseasePrevention and Control) (2020). The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2017/2018. *EFSA Journal*, 18(3).

[64] Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1982). The helical hydrophobic moment: A measure of the amphiphilicity of a helix. *Nature*, 299, 793–804.

[65] Ellabaan, M. M., Munck, C., Porse, A., Imamovic, L., & Sommer, M. O. (2021). Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. *Nature Communications*, 12.

[66] European Centre for Disease Prevention and Control (ECDC) (2020). Antimicrobial resistance in the EU/EEA (EARS-Net). `https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019`, accessed December 2021.

[67] Farrell, M. L., Joyce, A., Duane, S., Fitzhenry, K., Hooban, B., Burke, L. P., & Morris, D. (2021). Evaluating the potential for exposure to organisms of public health concern in naturally occurring bathing waters in Europe: A scoping review. *Water Research*, 206.

[68] Feng, Y., Zou, S., Chen, H., Yu, Y., & Ruan, Z. (2021). BacWGSTdb 2.0: A one-stop repository for bacterial whole-genome sequence typing and source tracking. *Nucleic Acids Research*, 49, D644–D650.

[69] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems*, 2015-Janua, 2962–2970.

[70] Findlay, B., Zhanel, G. G., & Schweizer, F. (2010). Cationic amphiphiles, a new generation of antimicrobials inspired by the natural antimicrobial peptide scaffold. *Antimicrobial Agents and Chemotherapy*, 54(10), 4049–4058.

[71] Fjell, C. D., Hiss, J. A., Hancock, R. E., & Schneider, G. (2012). Designing antimicrobial peptides: Form follows function. *Nature Reviews Drug Discovery*, 11(1), 37–51.

[72] Fox, J. L. (2013). Antimicrobial peptides stage a comeback. *Nature Biotechnology*, 31(5), 379–382.

[73] Frieri, M., Kumar, K., & Boutin, A. (2017). Antibiotic resistance. *Journal of Infection and Public Health*, 10(4), 369–378.

[74] Fritsche, O. (2016). *Mikrobiologie*, chapter 1,5, (pp. 4,5–6,180,185). Springer Spektrum: Berlin, Heidelberg.

[75] Fu, X., Cai, L., Zeng, X., & Zou, Q. (2020). StackCPPred: A stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics*, 36, 3028–3034.

[76] Fuchs, J. A., Grisoni, F., Kossenjans, M., Hiss, J. A., & Schneider, G. (2018). Lipophilicity prediction of peptides and peptide derivatives by consensus machine learning. *MedChemComm*, 9, 1538–1546.

[77] Gabere, M. N. & Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics*, 33(13), 1921–1929.

[78] Gabernet, G., Gautschi, D., Müller, A. T., Neuhaus, C. S., Armbrecht, L., Dittrich, P. S., Hiss, J. A., & Schneider, G. (2019). In silico design and optimization of selective membranolytic anticancer peptides. *Scientific reports*, 9, 11282.

[79] Galani, A., Alygizakis, N., Aalizadeh, R., Kastritis, E., Dimopoulos, M. A., & Thomaidis, N. S. (2021). Patterns of pharmaceuticals use during the first wave of COVID-19 pandemic in Athens, Greece as revealed by wastewater-based epidemiology. *Science of the Total Environment*, 798.

[80] Gebreyohannes, G., Nyerere, A., Bii, C., & Sbhatu, D. B. (2019). Challenges of intervention, treatment, and antibiotic resistance of biofilm-forming microorganisms. *Heliyon*, 5(8).

[81] Giuliani, A., Pirri, G., & Nicoletto, S. F. (2007). Antimicrobial peptides : An overview of a promising class of therapeutics. *Central European Journal of Biology*, 2(1), 1–33.

[82] Golbraikh, A., Muratov, E., Fourches, D., & Tropsha, A. (2014). Data set modelability by QSAR. *Journal of Chemical Information and Modeling*, 54, 1–4.

[83] Gonzalez, R., Islas, L., Obregon, A.-M., Escalante, L., & Sanchez, S. (1995). Gentamicin formation in micromonospora purpurea: Stimulatory effect of ammonium. *The Journal of Antibiotics*, 48.

[84] Grisoni, F., Neuhaus, C. S., Gabernet, G., Müller, A. T., Hiss, J. A., & Schneider, G. (2018). Designing Anticancer Peptides by Constructive Machine Learning. *ChemMedChem*, 13, 1300–1302.

[85] Guina, T., Yi, E. C., Wang, H., Hackett, M., & Miller, S. I. (2000). A PhoP-Regulated Outer Membrane Protease of Salmonella enterica Serovar Typhimurium Promotes Resistance to Alpha-Helical Antimicrobial Peptides. *JOURNAL OF BACTERIOLOGY*, 182, 4077–4086.

[86] Guo, Y., Yan, K., LV, H., & Liu, B. (2021). PreTP-EL: Prediction of therapeutic peptides based on ensemble learning. *Briefings in Bioinformatics*, 22.

[87] Gupta, C. L., Tiwari, R. K., & Cytryn, E. (2020). Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes. *Environment International*, 138.

[88] Gupta, S., Mittal, P., Madhu, M. K., & Sharma, V. K. (2017a). IL17eScan: A tool for the identification of peptides inducing IL-17 response. *Frontiers in Immunology*, 8.

[89] Gupta, S., Sharma, A. K., Shastri, V., Madhu, M. K., & Sharma, V. K. (2017b). Prediction of anti-inflammatory proteins/peptides: An insilico approach. *Journal of Translational Medicine*, 15(1).

[90] Hahn, M. W., Kasalický, V., Jezbera, J., Brandt, U., Jezberová, J., & Šimek, K. (2010). Limnohabitans curvus gen. nov., sp. nov., a planktonic bacterium isolated from a freshwater lake. *International Journal of Systematic and Evolutionary Microbiology*, 60(6), 1358–1365.

[91] Hamp, T. & Rost, B. (2015). More challenges for machine-learning protein interactions. *Bioinformatics*, 31, 1521–1525.

[92] Hancock, R. E., Haney, E. F., & Gill, E. E. (2016). The immunology of host defence peptides: Beyond antimicrobial activity. *Nature Reviews Immunology*, 16(5), 321–334.

[93] Harder, J., Bartels, J., Christophers, E., & Schröder, J.-M. (1997). A peptide antibiotic from human skin. *Nature*, 387, 861.

[94] Harris, F., Dennison, S., & Phoenix, D. (2009). Anionic Antimicrobial Peptides from Eukaryotic Organisms. *Current Protein & Peptide Science*, 10(6), 585–606.

[95] Hatosy, S. M. & Martiny, A. C. (2015). The ocean as a global reservoir of antibiotic resistance genes. *Applied and Environmental Microbiology*, 81, 7593–7599.

[96] Heidary, M., Khosravi, A. D., Khoshnood, S., Nasiri, M. J., Soleimani, S., & Goudarzi, M. (2018). Daptomycin. *Journal of Antimicrobial Chemotherapy*, 73, 1–11.

[97] Heider, D., Dybowski, J. N., Wilms, C., & Hoffmann, D. (2014). A simple structure-based model for the prediction of HIV-1 co-receptor tropism. *BioData Mining*, 7, 14.

[98] Heider, D. & Hoffmann, D. (2011). Interpol: An R package for preprocessing of protein sequences. *BioData Mining*, 4(1), 2–7.

[99] Hendriksen, R. S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S. K., Kjeldgaard, J., et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications*, 10(1).

[100] Hernando-Amado, S., Coque, T. M., Baquero, F., & Martínez, J. L. (2019). Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nature Microbiology*, 4(9), 1432–1442.

[101] Hilchie, A. L., Wuerth, K., & Hancock, R. E. W. (2013). Immune modulation by multifaceted cationic host defense (antimicrobial) peptides. *Nature Chemical Biology*, 9(12), 761–768.

[102] Hocquet, D., Muller, A., & Bertrand, X. (2016). What happens in hospitals does not stay in hospitals: Antibiotic-resistant bacteria in hospital wastewater systems. *Journal of Hospital Infection*, 93(4), 395–402.

[103] Hogg, T., Mesters, J. R., & Hilgenfeld, R. (2002). Inhibitory Mechanisms of Antibiotics Targeting Elongation Factor Tu. *Current Protein and Peptide Science*, 3, 121–131.

[104] Hooban, B., Fitzhenry, K., Cahill, N., Joyce, A., Connor, L. O., Bray, J. E., Brisse, S., Passet, V., Syed, R. A., Cormican, M., & Morris, D. (2021). A Point Prevalence Survey of Antibiotic Resistance in the Irish Environment, 2018–2019. *Environment International*, 152.

[105] Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Frontiers in Microbiology*, 11, 2559.

[106] Huang, Y., Huang, J., & Chen, Y. (2010a). Alpha-helical cationic antimicrobial peptides: Relationships of structure and function. *Protein and Cell*, 1(2), 143–152.

[107] Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010b). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26, 680–682.

[108] Hughes, J. P., Rees, S. S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, 162, 1239–1249.

[109] Hutchings, M., Truman, A., & Wilkinson, B. (2019). Antibiotics: Past, present and future. *Current Opinion in Microbiology*, 51, 72–80.

[110] Interagency Coordination Group on Antimicrobial Resistance (2019). *No time to wait: Securing the future from drug-resistant infections*. Report, World Health Organization (WHO).

[111] Ito, E. A., Katahira, I., da Rocha Vicente, F. F., Pereira, L. F. P., & Lopes, F. M. (2018). BASiNET — Biological sequences network: A case study on coding and non-coding RNAs identification. *Nucleic Acids Research*, 46.

[112] Janairo, J. I. B. (2021). Machine Learning for the Cleaner Production of Antioxidant Peptides. *International Journal of Peptide Research and Therapeutics*, 27, 2051–2056.

[113] Jhong, J. H., Chi, Y. H., Li, W. C., Lin, T. H., Huang, K. Y., & Lee, T. Y. (2019). dbAMP: An integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Research*, 47, D285–D297.

[114] Jhong, J.-H., Yao, L., Pang, Y., Li, Z., Chung, C.-R., Wang, R., Li, S., Li, W., Luo, M., Ma, R., et al. (2022). dbAMP 2.0: Updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Research*, 50, D460–D470.

[115] Jian, Z., Zeng, L., Xu, T., Sun, S., Yan, S., Yang, L., Huang, Y., Jia, J., & Dou, T. (2021). Antibiotic resistance genes in bacteria: Occurrence, spread, and control. *Journal of Basic Microbiology*.

[116] Jing, X., Dong, Q., Hong, D., & Lu, R. (2020). Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17, 1918–1931.

[117] Joo, H. S., Fu, C. I., & Otto, M. (2016). Bacterial strategies of resistance to antimicrobial peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371.

[118] Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7, 1–17.

[119] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(SUPPL. 1), 202–205.

[120] Kennedy, P. G. (2013). Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness). *The Lancet Neurology*, 12, 186–194.

[121] Khatun, M. S., Hasan, M. M., Shoombuatong, W., & Kurata, H. (2020). ProIn-Fuse: Improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. *Journal of Computer-Aided Molecular Design*, 34(12), 1229–1236.

[122] Kim, D. W. & Cha, C. J. (2021). Antibiotic resistome from the One-Health perspective: Understanding and controlling antimicrobial resistance transmission. *Experimental and Molecular Medicine*, 53, 301–309.

[123] Klein, E. Y., Boeckel, T. P. V., Martinez, E. M., Pant, S., Gandra, S., Levin, S. A., Goossens, H., & Laxminarayan, R. (2018). Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E3463–E3470.

[124] Kong, M., Bu, Y. Q., Zhang, Q., Zhang, S. H., Xing, L. Q., Gao, Z. Q., Bi, F. Z., & Hu, G. J. (2021). Distribution, abundance, and risk assessment of selected antibiotics in a shallow freshwater body used for drinking water, China. *Journal of Environmental Management*, 280.

[125] Koo, H. B. & Seo, J. (2019). Antimicrobial peptides under clinical investigation. *Peptide Science*, 111(5).

[126] Köster, J. & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2522.

[127] Kriegeskorte, A. & Peters, G. (2012). Horizontal gene transfer boosts MRSA spreading. *Nature Medicine*, 18, 662–663.

[128] Kuncheva, L. I. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51, 181–207.

[129] Kuncheva, L. I. (2014). *Combining Pattern Classifiers*, chapter 2,3,4,5,8,8,8, (pp. 49,104,113,177–178,247,247–255,276). Wiley: New Jersey, 2 edition.

[130] Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157, 105–132.

[131] Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, 7, 1511–1522.

[132] Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., & Tufféry, P. (2016). PEP-FOLD3: Faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, 44, W449–W454.

[133] LaPlante, K. L. & Rybak, M. J. (2004). Daptomycin - A novel antibiotic against Gram-positive pathogens. *Expert Opinion on Pharmacotherapy*, 5, 2321–2331.

[134] Larsson, D. G. J. & Flach, C.-F. (2021). Antibiotic resistance in the environment. *Nature Reviews Microbiology*.

[135] Lázár, V., Martins, A., Spohn, R., Daruka, L., Grézal, G., Fekete, G., Számel, M., Jangir, P. K., Kintses, B., Csörgo, B., et al. (2018). Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nature Microbiology*, 3(6), 718–731.

[136] Lazzaro, B. P., Zasloff, M., & Rolff, J. (2020). Antimicrobial peptides: Application informed by evolution. *Science*, 368(6490).

[137] Le, Q. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of Machine Learning Research*, volume 32 (pp. 1188–1196).: PMLR.

[138] Lee, E. Y., Fulan, B. M., Wong, G. C., & Ferguson, A. L. (2016). Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 13588–13593.

[139] Lee, E. Y., Lee, M. W., Fulan, B. M., Ferguson, A. L., & Wong, G. C. (2017). What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface Focus*, 7.

[140] Li, S.-C., Goto, N. K., Williams, K. A., Deber, C. M., & Fasman, G. D. (1996). Alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment. *Biochemistry*, 93, 6676–6681.

[141] Li, X., Stokholm, J., Brejnrod, A., Vestergaard, G. A., Russel, J., Trivedi, U., Thorsen, J., Gupta, S., Hjelmsø, M. H., Shah, S. A., et al. (2021). The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host and Microbe*, 29(6), 975–987.e4.

[142] Liu, B., Gao, X., & Zhang, H. (2019). BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic acids research*, 47(20), e127.

[143] Liu, S., Liu, C., & Deng, L. (2018a). Machine learning approaches for protein-protein interaction hot spot prediction: Progress and comparative assessment. *Molecules*, 23.

[144] Liu, Y., Bai, P., Woischnig, A. K., Hamri, G. C.-E., Ye, H., Folcher, M., Xie, M., Khanna, N., & Fussenegger, M. (2018b). Immunomimetic Designer Cells Protect Mice from MRSA Infection. *Cell*, 174, 259–270.e11.

[145] Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, 8.

[146] Löchel, H. F., Riemenschneider, M., Frishman, D., & Heider, D. (2018). SCOTCH : Subtype A Coreceptor Tropism Classification in HIV-1. *Bioinformatics*, 34(15), 2575–2580.

[147] Löchel, H. F. & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 19, 6263–6271.

[148] Löchel, H. F., Welzel, M., Hattab, G., Hauschild, A.-C., & Heider, D. (2021). Fractal construction of constrained code words for DNA storage systems. *Nucleic Acids Research*.

[149] Ma, H., Bandos, A. I., Rockette, H. E., & Gur, D. (2013). On use of partial area under the roc curve for evaluation of diagnostic performance. *Statistics in Medicine*, 32, 3449–3458.

[150] Mader, J. S. & Hoskin, D. W. (2006). Expert Opinion on Investigational Drugs Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment Cationic antimicrobial peptides as novel cytotoxic agents for cancer. *Expert Opinion on Investigational Drugs*, 15(8), 933–946.

[151] Magana, M., Pushpanathan, M., Santos, A. L., Leanse, L., Fernandez, M., Ioannidis, A., Giulianotti, M. A., Apidianakis, Y., Bradfute, S., Ferguson, A. L., et al. (2020). The value of antimicrobial peptides in the age of resistance. *The Lancet Infectious Diseases*, 20(9), e216–e230.

[152] Manavalan, B., Basith, S., & Lee, G. (2021). Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2. *Briefings in Bioinformatics*.

[153] Manavalan, B., Basith, S., Shin, T. H., Wei, L., & Lee, G. (2019). mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics*, 35, 2757–2765.

[154] Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., & Lee, G. (2018). Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *Journal of Proteome Research*, 17(8), 2715–2726.

[155] Mancuso, G., Midiri, A., Gerace, E., & Biondo, C. (2021). Bacterial Antibiotic Resistance: The Most Critical Pathogens. *Pathogens*, 10(10), 1310.

[156] Manna, M. S., Tamer, Y. T., Gaszek, I., Poulides, N., Ahmed, A., Wang, X., Toprak, F. C., Woodard, D. N. R., Koh, A. Y., Williams, N. S., et al. (2021). A trimethoprim derivative impedes antibiotic resistance evolution. *Nature Communications*, 12.

[157] Mannoor, M. S., Zhang, S., Link, A. J., & McAlpine, M. C. (2010). Electrical detection of pathogenic bacteria via immobilized antimicrobial peptides. *Proceedings of the National Academy of Sciences*, 107(45), 19207–19212.

[158] Mao, K., Zhang, H., Pan, Y., & Yang, Z. (2021). Biosensors for wastewater-based epidemiology for monitoring public health. *Water Research*, 191.

[159] Margineantu, D. D. & Dietterich, T. G. (1997). Pruning Adaptive Boosting. In *ICML*.

[160] marion Hutinel, Huijbers, P. M. C., Fick, J., Åhrén, C., Larsson, D. G. J., & Flach, C.-F. (2019). Population-level surveillance of antibiotic resistance in Escherichia coli through sewage analysis. *Eurosurveillance*, 24, 1–11.

[161] Marti, E., Huerta, B., Rodríguez-Mozaz, S., Barceló, D., Marcé, R., & Balcázar, J. L. (2018). Abundance of antibiotic resistance genes and bacterial community composition in wild freshwater fish species. *Chemosphere*, 196, 115–119.

[162] Mayer, C. D., Lorent, J., & Horgan, G. W. (2011). Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. *Statistical Applications in Genetics and Molecular Biology*, 10(1).

[163] McGinnis, W. D., Siu, C., S, A., & Huang, H. (2018). Category Encoders: A scikit-learn-contrib package of transformers for encoding categorical data. *The Journal of Open Source Software*, 3, 501.

[164] Mcguinness, W. A., Malachowa, N., & Deleo, F. R. (2017). Vancomycin Resistance in Staphylococcus aureus. *YALE JOURNAL OF BIOLOGY AND MEDICINE*, 90, 269–281.

[165] Meher, P. K., Sahu, T. K., Saini, V., & Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, 7.

[166] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv*.

[167] Mikut, R., Ruden, S., Reischl, M., Breitling, F., Volkmer, R., & Hilpert, K. (2016). Improving short antimicrobial peptides despite elusive rules for activity. *Biochimica et Biophysica Acta - Biomembranes*, 1858(5), 1024–1033.

[168] Miller, S. M., Simon, R. J., Ng, S., Zuckermann, R. N., Kerr, J. M., & Moos, W. H. (1995). Comparison of the Proteolytic Susceptibilities of Homologous L-Amino Acid, D-Amino Acid, and N-Substituted Clycine Peptide and Peptoid Oligomers. *Drug Development Research*, 35, 20–32.

[169] Mirzaei, R., Mesdaghinia, A., Hoseini, S. S., & Yunesian, M. (2019). Antibiotics in urban wastewater and rivers of Tehran, Iran: Consumption, mass load, occurrence, and ecological risk. *Chemosphere*, 221, 55–66.

[170] Mohanram, H. & Bhattacharjya, S. (2016). Salt-resistant short antimicrobial peptides. *Biopolymers*, 106(3), 345–356.

[171] Mohapatra, S., Hartrampf, N., Poskus, M., Loas, A., Gómez-Bombarelli, R., & Pentelute, B. L. (2020). Deep Learning for Prediction and Optimization of Fast-Flow Peptide Synthesis. *ACS Central Science*, 6, 2277–2286.

[172] Mollica, A., Pinnen, F., Azzurra, S., & Costante, R. (2014). The Evolution of Peptide Synthesis: From Early Days to Small Molecular Machines. *Current Bioactive Compounds*, 9, 184–202.

[173] Monti, M., Vertogen, B., Masini, C., Donati, C., Lilli, C., Zingaretti, C., Musuraca, G., Giorgi, U. D., Cerchione, C., Farolfi, A., et al. (2020). Hydroxychloroquine as Prophylaxis for COVID-19: A Review. *Frontiers in Pharmacology*, 11.

[174] Mtetwa, H. N., Amoah, I. D., Kumari, S., Bux, F., & Reddy, P. (2021). Wastewater-Based Surveillance of Antibiotic Resistance Genes Associated with Tuberculosis Treatment Regimen in KwaZulu Natal, South Africa. *Antibiotics*, 10, 1362.

[175] Mulani, M. S., Kamble, E. E., Kumkar, S. N., Tawre, M. S., & Pardesi, K. R. (2019). Emerging strategies to combat ESKAPE pathogens in the era of antimicrobial resistance: A review. *Frontiers in Microbiology*, 10.

[176] 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*.

[177] Müller, A. T., Hiss, J. A., & Schneider, G. (2018). Recurrent Neural Network Model for Constructive Peptide Design. *Journal of Chemical Information and Modeling*, 58, 472–479.

[178] Ngigi, A. N., Magu, M. M., & Muendo, B. M. (2020). Occurrence of antibiotics residues in hospital wastewater, wastewater treatment plant, and in surface water in Nairobi County, Kenya. *Environmental Monitoring and Assessment*, 192.

[179] Nicolas, P. (2009). Multifunctional host defense peptides: Intracellular-targeting antimicrobial peptides. *FEBS Journal*, 276(22), 6483–6496.

[180] Nnadozie, C. F. & Odume, O. N. (2019). Freshwater environments as reservoirs of antibiotic resistant bacteria and their role in the dissemination of antibiotic resistance genes. *Environmental Pollution*, 254.

[181] Noskin, G. A. (2005). Tigecycline: A New Glycylcycline for Treatment of Serious Infections. *Clinical Infectious Diseases*, 41, 303–317.

[182] Olson, R. S., Edu, O., & Moore, J. H. (2016). TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning. In *JMLR: Workshop and Conference Proceedings*, volume 64 (pp. 66–74).

[183] Pages, J. M., Lavigne, J. P., Leflon-Guibout, V., Marcon, E., Bert, F., Noussair, L., & Nicolas-Chanoine, M. H. (2009). Efflux pump, the masked side of ß-lactam resistance in Klebsiella pneumoniae clinical isolates. *PLoS ONE*, 4.

[184] Papo, N. & Shai, Y. (2005). A Molecular Mechanism for Lipopolysaccharide Protection of Gram-negative Bacteria from Antimicrobial Peptides. *The Journal of Biological Chemistry*, 280(11), 10378–10387.

[185] Parnham, M. J., Haber, V. E., Giamarellos-Bourboulis, E. J., Perletti, G., Verleden, G. M., & Vos, R. (2014). Azithromycin: Mechanisms of action and their relevance for clinical applications. *Pharmacology and Therapeutics*, 143, 225–245.

[186] Peng, J. & Xu, J. (2011). RaptorX: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function and Bioinformatics*, 79(SUPPL. 10), 161–171.

[187] Peschel, A. & Sahl, H. G. (2006). The co-evolution of host cationic antimicrobial peptides and microbial resistance. *Nature Reviews Microbiology*, 4, 529–536.

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

[189] Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175(4), 7–9.

[190] Qu, J., Huang, Y., & Lv, X. (2019). Crisis of antimicrobial resistance in China: Now and the future. *Frontiers in Microbiology*, 10.

[191] Rathinakumar, R., Walkenhorst, W. F., & Wimley, W. C. (2009). Broad-spectrum antimicrobial peptides by rational combinatorial design and high-throughput screening: The importance of interfacial activity. *Journal of the American Chemical Society*, 131(22), 7609–7617.

[192] Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21, 4–21.

[193] Raza, A., Ngieng, S. C., Sime, F. B., Cabot, P. J., Roberts, J. A., Popat, A., Kumeria, T., & Falconer, J. R. (2021). Oral meropenem for superbugs: Challenges and opportunities. *Drug Discovery Today*, 26, 551–560.

[194] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2021). Classifier Chains: A Review and Perspectives. *Journal of Artificial Intelligence Research*, 70, 683–718.

[195] Regina, A. L. A., Medeiros, J. D., Teixeira, F. M., Côrrea, R. P., Santos, F. A. M., Brantes, C. P. R., Pereira, I. A., Stapelfeldt, D. M. A., Diniz, C. G., & da Silva, V. L. (2021). A watershed impacted by anthropogenic activities: Microbial community alterations and reservoir of antimicrobial resistance genes. *Science of the Total Environment*, 793.

[196] Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., Hauschild, A.-C., Schwengers, O., & Heider, D. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38, 325–334.

[197] Rogers, D. & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50, 742–754.

[198] Roguet, A., Therial, C., Catherine, A., Bressy, A., Varrault, G., Bouhdamane, L., Tran, V., Lemaire, B. J., Vincon-Leite, B., Saad, M., et al. (2018). Importance of Local and Regional Scales in Shaping Mycobacterial Abundance in Freshwater Lakes. *Microbial Ecology*, 75(4), 834–846.

[199] Romero-Molina, S., Ruiz-Blanco, Y. B., Green, J. R., & Sanchez-Garcia, E. (2019). ProtDCal-Suite: A web server for the numerical codification and functional analysis of proteins. *Protein Science*, 28, 1734–1743.

[200] Ruiz-Blanco, Y. B., Paz, W., Green, J., & Marrero-Ponce, Y. (2015). ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics*, 16.

[201] Santafe, G., Inza, I., & Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4), 467–508.

[202] Sato, H. & Feix, J. B. (2006). Peptide – membrane interactions and mechanisms of membrane destruction by amphipathic *α* -helical antimicrobial peptides. *Biochimica et Biophysica Acta*, 1758, 1245–1256.

[203] Schar, D., Zhao, C., Wang, Y., Larsson, D. G., Gilbert, M., & Boeckel, T. P. V. (2021). Twenty-year trends in antimicrobial resistance from aquaculture and fisheries in Asia. *Nature Communications*, 12.

[204] Schissel, C. K., Mohapatra, S., Wolfe, J. M., Fadzen, C. M., Bellovoda, K., Wu, C. L., Wood, J. A., Malmberg, A. B., Loas, A., Gómez-Bombarelli, R., & Pentelute, B. L. (2021). Deep learning to design nuclear-targeting abiotic miniproteins. *Nature Chemistry*, 13, 992–1000.

[205] Schmidt, N. W., Mishra, A., Lai, G. H., Davis, M., Sanders, L. K., Tran, D., Garcia, A., Tai, K. P., McCray, P. B., Ouellette, A. J., et al. (2011). Criterion for amino acid composition of defensins and antimicrobial peptides based on geometry of membrane destabilization. *Journal of the American Chemical Society*, 133(17), 6720–6727.

[206] Schulte, C., Soldà, A., Spänig, S., Adams, N., Streicher, W., Heider, D., Strasser, R., & Maric, H. M. (2022). Fluorescence Proximity Sensing for Real Time Affinity Determination of Multivalent Peptide-Protein Interactions. *Manuscript submitted for publication.*

[207] Selvam, A., Xu, D., Zhao, Z., & Wong, J. W. (2012). Fate of tetracycline, sulfonamide and fluoroquinolone resistance genes and the changes in bacterial diversity during composting of swine manure. *Bioresource Technology*, 126, 383–390.

[208] Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14.

[209] Seo, M.-D., Won, H.-S., Kim, J.-H., Mishig-Ochir, T., & Lee, B.-J. (2012). Antimicrobial Peptides for Therapeutic Applications: A Review. *Molecules*, 17(12), 12276–12286.

[210] Sequeira, A. M., Lousa, D., & Rocha, M. (2021). ProPythia: A Python package for protein classification based on machine and deep learning. *Neurocomputing*.

[211] Shah, N., Nute, M. G., Warnow, T., & Pop, M. (2019). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*, 35, 1613–1614.

[212] Shen, Y., Liu, C., Chi, K., Gao, Q., Bai, X., Xu, Y., & Guo, N. (2022). Development of a machine learning-based predictor for identifying and discovering antioxidant peptides based on a new strategy. *Food Control*, 131.

[213] Shoombuatong, W., Schaduangrat, N., Pratiwi, R., & Nantasenamat, C. (2019). THPep: A machine learning-based approach for predicting tumor homing peptides. *Computational Biology and Chemistry*, 80, 441–451.

[214] Sila, A. & Bougatef, A. (2016). Antioxidant peptides from marine by-products: Isolation, identification and application in food systems. A review. *Journal of Functional Foods*, 21, 10–26.

[215] Simeon, S., Li, H., Win, T. S., Malik, A. A., Kandhro, A. H., Piacham, T., Shoombuatong, W., Nuchnoi, P., Wikberg, J. E., Gleeson, M. P., & Nantasenamat, C. (2017). PepBio: Predicting the bioactivity of host defense peptides. *RSC Advances*, 7, 35119–35134.

[216] Singh, D., Singh, P., & Sisodia, D. S. (2019). Evolutionary based ensemble framework for realizing transfer learning in HIV-1 Protease cleavage sites prediction. *Applied Intelligence*, 49, 1260–1282.

[217] Smiline, A. S., Vijayashree, J. P., & Paramasivam, A. (2018). Molecular characterization of plasmid-encoded blaTEM, blaSHV and blaCTX-M among extended spectrum β-lactamases [ESBLs] producing Acinetobacter baumannii. *British Journal of Biomedical Science*, 75, 200–202.

[218] Snyder, D. A., Chen, Y., Denissova, N. G., Acton, T., Aramini, J. M., Ciano, M., Karlin, R., Liu, J., Manor, P., Rajan, P. A., et al. (2005). Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. *Journal of the American Chemical Society*, 127, 16505–16511.

[219] Song, Y. M., Yang, S. T., Lim, S. S., Kim, Y., Hahm, K. S., Kim, J. I., & Shin, S. Y. (2004). Effects of L- or D-Pro incorporation into hydrophobic or hydrophilic helix face

of amphipathic α-helical model peptide on structure and cell selectivity. *Biochemical and Biophysical Research Communications*, 314, 615–621.

[220] Spänig, S. & Heider, D. (2019). Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Mining*, 12(1), 1–29.

[221] Spänig, S., Eick, L., Nuy, J. K., Beisser, D., Ip, M., Heider, D., & Boenigk, J. (2021a). A multi-omics study on quantifying antimicrobial resistance in European freshwater lakes. *Environment International*, 157.

[222] Spänig, S., Michel, A., & Heider, D. (2022). Unsupervised encoding selection through ensemble pruning for biomedical classification. *bioRxiv*. Manuscript submitted for publication.

[223] Spänig, S., Mohsen, S., Hattab, G., Hauschild, A.-C., & Heider, D. (2021b). A large-scale comparative study on peptide encodings for biomedical classification. *NAR Genomics and Bioinformatics*, 3.

[224] Stedtfeld, R. D., Williams, M. R., Fakher, U., Johnson, T. A., Stedtfeld, T. M., Wang, F., Khalife, W. T., Hughes, M., Etchebarne, B. E., Tiedje, J. M., & Hashsham, S. A. (2016). Antimicrobial resistance dashboard application for mapping environmental occurrence and resistant pathogens. *FEMS Microbiology Ecology*, 92.

[225] Sultan, I., Rahman, S., Jan, A. T., Siddiqui, M. T., Mondal, A. H., & Haq, Q. M. R. (2018). Antibiotics, resistome and resistance mechanisms: A bacterial perspective. *Frontiers in Microbiology*, 9.

[226] Surette, M. D., Spanogiannopoulos, P., & Wright, G. D. (2021). The Enzymes of the Rifamycin Antibiotic Resistome. *Accounts of Chemical Research*, 54, 2065–2075.

[227] Tariq, A., Siddiqui, M. R., Kumar, J., Reddy, D., Negi, P. S., Chaudhary, M., Srivastava, S. M., & Singh, R. K. (2010). Development and validation of high performance liquid chromatographic method for the simultaneous determination of ceftriaxone and vancomycin in pharmaceutical formulations and biological samples. *ScienceAsia*, 36, 297–304.

[228] Teillant, A., Gandra, S., Barter, D., Morgan, D. J., & Laxminarayan, R. (2015). Potential burden of antibiotic resistance on surgery and cancer chemotherapy antibiotic prophylaxis in the USA: A literature review and modelling study. *The Lancet Infectious Diseases*, 15, 1429–1437.

[229] Teixeira, V., Feio, M. J., & Bastos, M. (2012). Role of lipids in the interaction of antimicrobial peptides with membranes. *Progress in Lipid Research*, 51(2), 149–177.

[230] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. *CoRR*.

[231] Toke, O. (2005). Antimicrobial Peptides : New Candidates in the Fight Against Bacterial Infections. *Peptide Science*, 80(6), 717–735.

[232] Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., & Raghava, G. P. (2015). CancerPPD: A database of anticancer peptides and proteins. *Nucleic Acids Research*, 43, D837–D843.

[233] Vaara, M. (1992). Agents That Increase the Permeability of the Outer Membrane. *MICROBIOLOGICAL REVIEWS*, (pp. 395–411).

[234] VanOeffelen, M., Nguyen, M., Aytan-Aktug, D., Brettin, T., Dietrich, E. M., Kenyon, R. W., Machi, D., Mao, C., Olson, R., Pusch, G. D., et al. (2021). A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes. *Briefings in Bioinformatics*, 22.

[235] Velez, R. & Sloand, E. (2016). Combating antibiotic resistance, mitigating future threats and ongoing initiatives. *Molecular Ecology*, 25(13-14), 1886–1889.

[236] Vens, C., Rosso, M. N., & Danchin, E. G. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27, 1231–1238.

[237] von der Saal, K. (2020). *Biochemie*, chapter 2,2,6,19, (pp. 12,19,60,242). Springer Berlin Heidelberg: Berlin, Heidelberg.

[238] Waghu, F. H., Barai, R. S., Gurung, P., & Idicula-Thomas, S. (2016). CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, 44, D1094–D1097.

[239] Wang, G. (2014). Human antimicrobial peptides and proteins. *Pharmaceuticals*, 7(5), 545–594.

[240] Wang, G., Li, X., & Wang, Z. (2016). APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44, D1087–D1093.

[241] Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., Song, J., Chou, K. C., & Lithgow, T. (2017). POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics*, 33, 2756–2758.

[242] Wang, S., Yang, Z., Liu, Y., Zhao, M.-T., Zhao, J., Zhang, H., Liu, Z.-Y., Wang, X.-L., Ma, L., & Yang, Y.-H. (2021). Application of topical gentamicin—a new era in the treatment of genodermatosis. *World Journal of Pediatrics*, 17, 568–575.

[243] Wang, Z. & Wang, G. (2004). APD: The antimicrobial peptide database. *Nucleic Acids Research*, 32.

[244] Watkinson, A. J., Murby, E. J., Kolpin, D. W., & Costanzo, S. D. (2009). The occurrence of antibiotics in an urban watershed: From wastewater to drinking water. *Science of the Total Environment*, 407, 2711–2723.

[245] Wei, L., Hu, J., Li, F., Song, J., Su, R., & Zou, Q. (2018). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Briefings in Bioinformatics*, 21, 106–119.

[246] Wiesner, J. & Vilcinskas, A. (2010). Antimicrobial peptides: The ancient arm of the human immune system. *Virulence*, 1(5), 440–464.

[247] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

[248] Wimley, W. C. & Hristova, K. (2011). Antimicrobial peptides: Successes, challenges and unanswered questions. *Journal of Membrane Biology*, 239(1-2), 27–34.

[249] Winter, M., Buckling, A., Harms, K., Johnsen, P. J., & Vos, M. (2021). Antimicrobial resistance acquisition via natural transformation: Context is everything. *Current Opinion in Microbiology*, 64, 133–138.

[250] World Health Organization (WHO) (2019). Critically Important Antimicrobials for Human Medicine. `https://apps.who.int/iris/handle/10665/312266`, accessed February 7, 2021.

[251] Worth, C. L., Gong, S., & Blundell, T. L. (2009). Structural and functional constraints in the evolution of protein families. *Nature Reviews Molecular Cell Biology*, 10, 709–720.

[252] Xiao, X., Wang, P., Lin, W. Z., Jia, J. H., & Chou, K. C. (2013). IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry*, 436(2), 168–177.

[253] Xu, J., Li, F., Leier, A., Xiang, D., Shen, H. H., Lago, T. T. M., Li, J., Yu, D. J., & Song, J. (2021a). Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, 22.

[254] Xu, L., Huang, H., Wei, W., Zhong, Y., Tang, B., Yuan, H., Zhu, L., Huang, W., Ge, M., Yang, S., et al. (2014). Complete genome sequence and comparative genomic analyses of the vancomycin-producing Amycolatopsis orientalis. *BMC Genomics*, 15.

[255] Xu, Z., Luo, M., Lin, W., Xue, G., Wang, P., Jin, X., Xu, C., Zhou, W., Cai, Y., Yang, W., Nie, H., & Jiang, Q. (2021b). DLpTCR: An ensemble deep learning framework for predicting immunogenic peptide recognized by T cell receptor. *Briefings in Bioinformatics*, 22.

[256] Xue, A. Y., Szymczak, L. C., Mrksich, M., & Bagheri, N. (2017). Machine Learning on Signal-to-Noise Ratios Improves Peptide Array Design in SAMDI Mass Spectrometry. *Analytical Chemistry*, 89, 9039–9047.

[257] Yamashita, H., Fujitani, M., Shimizu, K., Kanie, K., Kato, R., & Honda, H. (2020). Machine Learning-Based Amino Acid Substitution of Short Peptides: Acquisition of Peptides with Enhanced Inhibitory Activities against *α*-Amylase and *α*-Glucosidase. *ACS Biomaterials Science and Engineering*, 6, 6117–6125.

[258] Yeung, A. T., Gellatly, S. L., & Hancock, R. E. (2011). Multifunctional cationic host defence peptides and their clinical applications. *Cellular and Molecular Life Sciences*, 68(13), 2161–2176.

[259] Yin, L. M., Edwards, M. A., Li, J., Yip, C. M., & Deber, C. M. (2012). Roles of hydrophobicity and charge distribution of cationic antimicrobial peptides in peptide-membrane interactions. *Journal of Biological Chemistry*, 287(10), 7738–7745.

[260] Yuan, S. F., Liu, Z. H., Huang, R. P., Yin, H., & Dang, Z. (2016). Levels of six antibiotics used in china estimated by means of wastewater-based epidemiology. *Water Science and Technology*, 73, 769–775.

[261] Zaiou, M. (2007). Multifunctional antimicrobial peptides: Therapeutic targets in several human diseases. *Journal of Molecular Medicine*, 85, 317–329.

[262] Zhang, A. N., Gaston, J. M., Dai, C. L., Zhao, S., Poyet, M., Groussin, M., Yin, X., Li, L. G., van Loosdrecht, M. C., Topp, E., et al. (2021). An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nature Communications*, 12.

[263] Zhang, J., Li, S., Ding, Y., Tang, J., & Guo, F. (2020). An two-layer predictive model of ensemble classifier chain for detecting antimicrobial peptides. In *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020* (pp. 56–61).: Institute of Electrical and Electronics Engineers Inc.

[264] Zhao, J., Zhao, C., Liang, G., Zhang, M., & Zheng, J. (2013a). Engineering antimicrobial peptides with improved antimicrobial and hemolytic activities. *Journal of Chemical Information and Modeling*, 53, 3280–3296.

[265] Zhao, X., Wu, H., Lu, H., Li, G., & Huang, Q. (2013b). LAMP: A Database Linking Antimicrobial Peptides. *PLoS ONE*, 8.

[266] Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, 2, 349–360.

[267] Zhu, Y. G., Zhao, Y., Li, B., Huang, C. L., Zhang, S. Y., Yu, S., Chen, Y. S., Zhang, T., Gillings, M. R., & Su, J. Q. (2017). Continental-scale pollution of estuaries with antibiotic resistance genes. *Nature Microbiology*, 2.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive mit (x11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.