Project Details	
Project Code	IIAR27Ex Vos
Title	How pathogens emerge: integrating large language models with comparative genomics to uncover the zoonotic potential of bacteria
Research Theme	Infection, Immunity & Antimicrobial Resistance
Project Type	Dry lab
Summary	Most emerging bacterial pathogens originate from animal hosts (are 'zoonotic'), yet it is hard to predict which pathogen species are likely to cross into human populations and why. This project will combine AI Large Language Models, extracting infection context from hundreds of thousands scientific publications, with comparative bacterial genomics, to identify molecular signatures of host switches. Integrating these complementary approaches will generate biologically interpretable models of zoonotic emergence and ultimately aid in developing new tools for human pathogen surveillance. This project offers training in bioinformatics, data science and microbiology, preparing students for impactful careers in infectious disease research.
Description	Background Approximately 60% of emerging infectious diseases originate in animals (termed 'zoonoses'), yet our ability to predict which bacterial species are capable of becoming human pathogens remains poor. Factors contributing to emergence are complex, resulting from ecological opportunity (e.g. changing agricultural practices, evolution of antibiotic resistance, climate change, demography) and evolutionary capability (genomic change). Although bacterial genomes are now available for thousands of species, genome sequences alone rarely explain why closely related bacteria differ dramatically in their ability to infect humans. Equally, a vast wealth of information on human infection remains locked within hundreds of thousands of scientific publications and therefore is difficult to compile and interrogate at scale, let alone integrate with genomic analyses. Recent advances in Large Language Models (LLMs) offer an unprecedented opportunity to transform the scientific literature into structured biological knowledge. Combined with comparative genomics, this creates a powerful framework to understand the factors contributing to the emergence of devastating zoonotic infections. In this project, we will first use LLMs to build a 'pathogen atlas', extracting information on all bacterial pathogens of animals and humans from the literature. Second, we will construct a predictive model to uncover the different components of ecological opportunity best predicting zoonotic status. Finally, species where prediction does not match zoonotic status will be used to generate informative hypotheses and case studies for comparative genomics analyses to identify signatures of host switching. This project will develop novel approaches to understand the fundamental biology of zoonotic bacterial emergence with the capability to inform future pathogen surveillance and provide interdisciplinary training in bioinformatics, data science and microbiology. Objectives Objective 1. Build a pathogen atlas of zoonotic potential

	Building upon recent advances in LLM-assisted literature mining developed by Vos and Costa (bioRxiv, 2025-07), the project will construct a comprehensive pathogen atlas describing bacterial species known to infect animals and/or humans. For each species, LLMs will key information such as on transmission, host range and patient characteristics from PubMed-indexed literature. The atlas will preserve uncertainty, sampling effort and provenance of evidence, rather than treating absence of reports as evidence of absence. This pathogen atlas will lead to a publicly available community resource on the diversity of bacterial pathogens of animals and humans and provide the foundation for modelling and comparative genomics described below. Objective 2. Model ecological opportunity and identify informative mismatches Utilizing the comprehensive database compiled in WP1, we will build an explicit, literature-based predictive model to identify drivers of increased zoonotic risk for different phylogenetic lineages. First, texts from literature will be transformed into networks of interconnected concepts using LLMs. Second, a zoonosis predictor will be developed using machine learning algorithms for networks. Third, the predictions will be analysed to identify characteristic concept patterns associated with zoonosis. Fourth, the predictive model will be applied to provide a quantitative measure of zoonotic potential, expressed as the predictor's confidence in the zoonotic behaviour of each known animal pathogen. These results can be used as a foundation for enhancing surveillance strategies, identifying risk factors, and guiding public health interventions. Objective 3. Comparative genomics to reveal mechanisms that allow zoonotic transmission While the literature-derived analyses above generate predictions and hypotheses, comparative genomics tests can be used to uncover genetic risk factors for zoonosis. Rather than searching for ‘zoonosis genes’, the project will aim to uncover any genomic feature (e.g. gene flow) that can explain residual differences in zoonotic outcome after accounting for the ecological-opportunity models. We will select the most informative species complexes from WP2 where there is misalignment between opportunity and outcome, for example - animal pathogens with substantial human contact but no documented zoonotic transmission or species that infect humans despite apparently limited ecological opportunity, alongside congeneric control species. Integrating genomic analyses with the outcome of the ecological-opportunity model will provide a new way to uncover potentially high-risk pathogen species. Significance The project addresses a fundamental question in infectious disease biology: why do some bacterial species become human pathogens while closely related species do not? It aims to disentangle ecological opportunity and evolutionary change as determinants of zoonotic emergence by integrating artificial intelligence with comparative genomics. The project establishes a new paradigm for using large language models in biology. Rather than employing AI simply to summarise literature, it uses LLMs to build a pathogen atlas, construct an explicit model of ecological opportunity and generate hypotheses for

	targeted comparative genomic analyses. This strategy of integrating scientific literature with genome sequence data could lay groundwork for a range of further questions in biology and medicine. Training environment and student ownership The student will receive interdisciplinary training in machine learning, large language models and explainable AI, bioinformatics and bacterial comparative genomics and microbiology and evolution and be immersed in stimulating scientific environments spanning three campuses (see sections below).
Can be completed part time?	Yes
Supervisory Team	
Lead Supervisor	
Name	Dr Michiel Vos
Affiliation	Exeter
College/Faculty	Health and Life Sciences
Department/School	Public Health and Sport Sciences
Email Address	m.vos@exeter.ac.uk
Co-Supervisor 1	
Name	Dr Fabrizio Costa
Affiliation	Exeter
College/Faculty	Faculty of Environment, Science and Economy
Department/School	Department of Computer Science
Co-Supervisor 2	
Name	Dr Sion Bayliss
Affiliation	Bristol
College/Faculty	Faculty of Health and Life Sciences
Department/School	Veterinary School
Co-Supervisor 3	
Name	
Affiliation	
College/Faculty	
Department/School	