Integration of Bioinformatics and Machine Learning to characterize *Fusobacterium nucleatum*'s pathogenicity

Zihan Tian[1,2] and Pietro Liò[2]
[1]Dept of Biology, John Hopkins University,
[2]Department of Computer Science and Technology,
University of Cambridge, Cambridge, UK
ztian12@jh.edu pl219@cam.ac.uk

Abstract

Fusobacterium nucleatum has been found to be associated with cancer lesions in both oral and colon cancers. Although important studies have dissected the clinical aspects of its remarkable pathogenicity, there is a lack of molecular studies. Our computational work based on bioinformatics and machine learning methodologies has predicted potential pathogenicity islands. The study has involved the analysis of genome-based compositional bias, promoter maps, codon adaptation index, protein structure, and characterized these genomic regions on the basis of predicting base compositional bias, promoter mapping, protein abundances, and interactions, metabolic model to characterize these regions. Although most of the currently in use pathogenicity islands finder software detect the presence of three pathogenicity islands, our analysis suggests that only one is present. Furthermore, we have investigated and discussed the metabolic advantages of pathogenicity, particularly iron ion scavenging activity. Our work has two immediate and important benefits: the improved understanding of the biological processes that shape the pathogenicity and evolution of *Fusobacterium nucleatum* at the molecular level and the improved ability of integrating and automating the state-of-the-art bioinformatics tools and machine learning approaches in the inference of the mechanistic interpretability of a pathogenic phenotype.

Introduction

Various studies indicate that microbial dysbiosis is associated with oral and systemic diseases (Han, 2015). *Fusobacterium nucleatum* (*F.nucleatum*), a Gram-negative anaerobe naturally found in the oral cavity of humans, contributes to periodontal disease, as well as oral and colorectal cancers (Patel et al., 2022; Queen et al., 2025). The association of *F. nucleatum* with oral squamous cell carcinoma (OSCC) highlights its role via oncogenic signaling pathways and has metabolic adaptability in the tumor microenvironments (Lim et al., 2025).

Beside being a pathogen, *F. nucleatum* could induce resistance to antitumor therapies, tumor recurrence, and be associated with a poor prognosis in patients. Chemotherapy resistance is one of the major contributors to the poor prognosis in colorectal cancer. Mechanistic studies have shown that F. nucleatum can inhibit pyroptosis, a form of programmed cell death induced by chemotherapeutic agents, by down-regulating the Hippo signaling pathway and up-regulating antiapoptotic BCL2 (Han et al., 2020; Wang et al., 2024). In addition, growing evidence suggests that removal of tumor-associated *F. nucleatum* can induce the host immune response to recognize tumor cells, suggesting potential future antibacteria / tumor vaccine applications (Chen and Huang, 2024).

Notably, *F.nucleatum* can coexist with more than 500 other species, and its elongated shape allows it to bridge other bacteria, maintaining the development of biofilm and facilitating the spreading of periodontal disease in humans. *F. nucleatum* enhances the invasive potential of other microorganisms, such as *Porphyromonas gingivalis*. Together, these bacteria form a pathogenic alliance that can significantly alter host responses and promote cancer progression (How et al., 2016; Mohammed et al., 2020; Patel et al., 2022).

Despite *F. nucleatum*'s association with various cancers in many different tissues and its contribution to therapeutic resistance, the genomic, proteomic, and metabolic characteristics of its specific pathogenic function remain poorly characterized (Ma et al., 2023). To address this gap, this study

focuses on a specific strain of *F. nucleatum* ATCC 25586 that is naturally present in the human oral cavity. An integrated bioinformatics and machine learning framework is used to predict potential pathogenicity islands (PAIs). Protein and metabolic pathways that may be associated with cancer are identified. Structural predictions (AlphaFold, TMHMM, and String) are used to evaluate protein-protein interactions and potential transmembrane regions. In parallel, genome-scale metabolic modeling using COBRApy and flux balance analysis was used to explore iron dependency and pH tolerance. This study aims to better understand what contributes to *F. nucleatum* pathogenicity and offers insights into its connection to cancer progression and therapeutic resistance.

Methodology: Building a Bioinformatics-Machine Learning integrated workflow

Genomic analysis pipeline focusing on *Fusobacterium nucleatum* subsp. nucleatum strain ATCC 25586, a critical periodontal pathogen with established links to various oral diseases and colorectal cancers.

The genome of *Fusobacterium nucleatum* subsp. nucleatum strain ATCC 25586 (NCBI accession: AE009951.2) was downloaded from the NCBI Genome database. Genome annotation was performed using Prokka (Seemann, 2014) to generate a detailed catalog of its genetic components, including coding sequences, rRNAs, and tRNAs.

To identify potential pathogenic islands (PAI) in bacteria of interest, GC skew analysis, homology-based functional analysis (UniProt + BLAST), and IslandViewer (Bertelli et al., 2017) genomic island predictions were used. Promoter predictions were generated using the ProPr: Prokaryote Promoter Prediction v2.0 (de Jong et al., 2012) and integrated with gene annotations from NCBI GenBank to evaluate open reading frames.

Synonymous codon usage bias was assessed using the Codon Adaptation Index (CAI). The ribosomal protein genes served as the reference set, and the CAI is calculated based on the API Reference (Bagnoli and Liò, 1995; Lee, 2019). High CAI values suggest strong codon optimization, likely for efficient translation. AlphaFold and String database are used to model protein structures and interactions (Jumper et al., 2021; Szklarczyk et al., 2024). TMHMM2.0 analysis is used to identify transmembrane proteins (Krogh et al., 2001).

Flux Balance Analysis (FBA) of a genome-scale metabolic model(generated with CarveMe) was used to explore iron dependency (70 iron-related reactions, 40 metabolites). Computational approaches were adapted from Ratzke and Gore (Ratzke and Gore, 2018), which employ differential equations and Gaussian preference functions, to predict bacterium adaptation to acidic conditions, relevant to tumor microenvironment colonization potential.

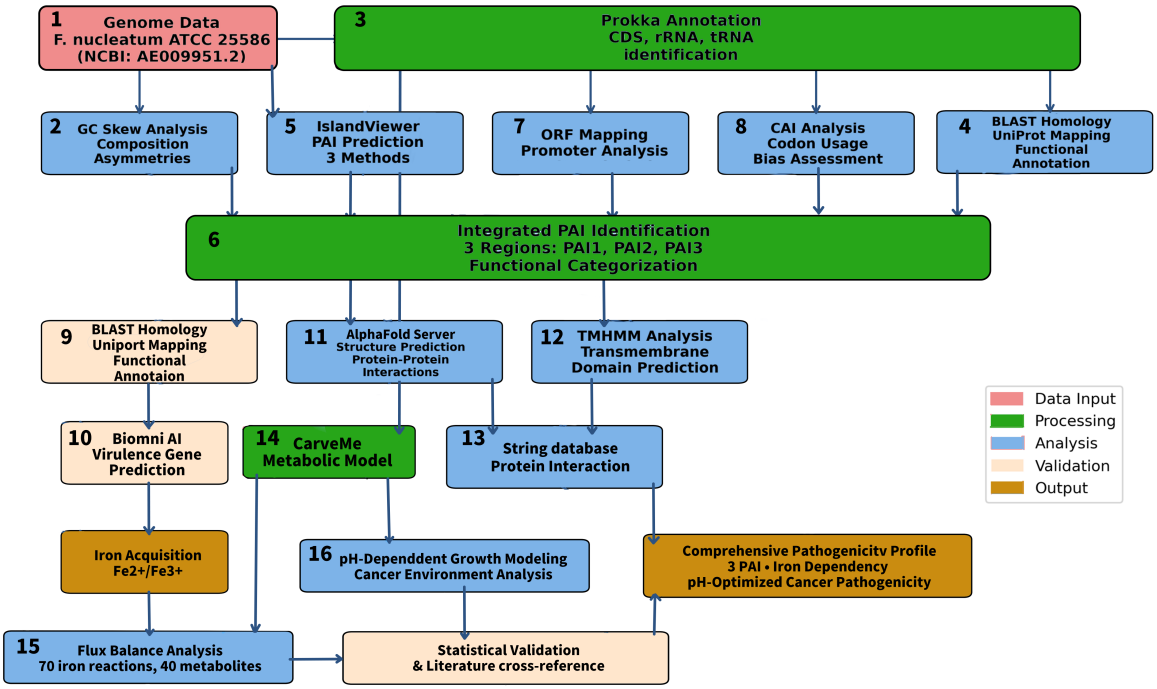

Figure 1: Flowchart illustrating the analytic pipeline. 1) Genome retrieval from NCBI. 2) GC skew compositional analysis. 3) Genome annotation with Prokka. 4) Using UniProt and BLAST to identify potential identities for uncharacterized proteins, categorize proteins in certain regions that are identified as deviating from GC skew. 5) Genomic island predictions using Island Viewer. 6) Predicting 3 potential PAI. 7) Promoter and open reading frame (ORF) mapping. 8) Ribosomal protein genes were used as the reference set, and the CAI was calculated by comparing each gene's codon usage to the Relative Synonymous Codon Usage (RSCU) values of the ribosomal proteins. 9) Functional annotation is expanded by BLAST. 10) Virulence gene prediction via the biomni AI tool. 11) Structural analyses include AlphaFold protein structure prediction, 12) TMHMM-based transmembrane domain detection, and 13) Using the String database to predict multi-protein interaction networks. 14) Metabolic modeling generated by CarveMe. 15) flux balance analysis (FBA) targeting 70 iron-related reactions and 40 metabolites. 16) pH tolerance is assessed through pH-dependent growth modeling. Data is statistically validated and cross-referenced with the literature.

Results and Discussion

Genome Analysis

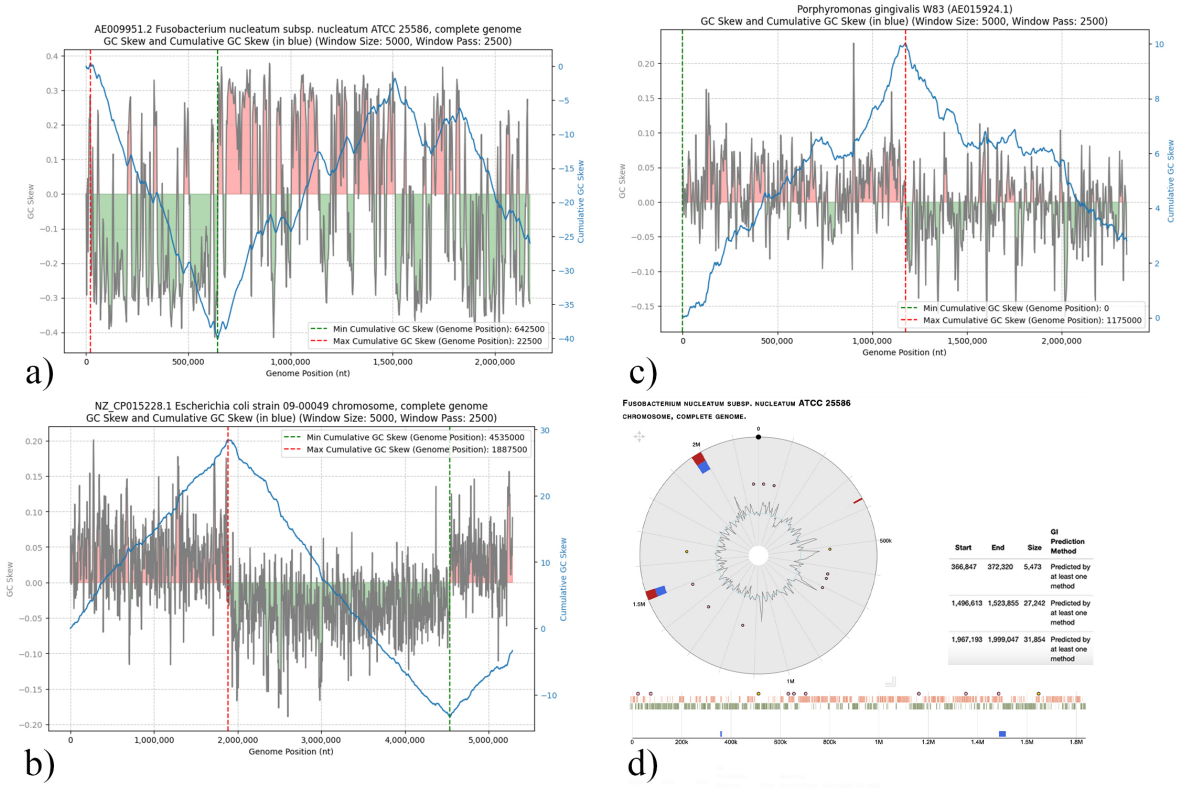

Figure 2: Panels a–c show GC Skew (in grey, where positive GC skew values have a red pattern and negative ones have a green one) and Cumulative GC Skew (blue) analyses of the complete genomes of Fusobacterium nucleatum subsp. nucleatum ATCC 25586 (GenBank: AE009951.2), Escherichia coli strain 09-00049 (GenBank: NZ_CP015228.1), and Porphyromonas gingivalis W83 (GenBank: AE015924.1), respectively. These plots were generated using the GC_skew-plot tool by Sergi Vallvé and highlight nucleotide composition asymmetries across each genome (Vallvé, 2023). Panel d presents genomic island predictions in F. nucleatum ATCC 25586 using IslandViewer, identifying three potential pathogenicity islands (PAIs) in Fusobacterium nucleatum subsp. nucleatum ATCC 25586. These regions were predicted by at least one of the integrated methods: IslandPick, IslandPath-DIMOB, or SIGI-HMM (Bertelli et al., 2017).

As shown in Figure 2, notable fluctuations in GC skew were observed between the 1.6 Mb and 1.9 Mb region, suggesting atypical nucleotide composition that may indicate genomic anomalies such as prophages or integrated mobile genetic elements (MGEs). In addition, genomic island predictions were performed using IslandViewer Bertelli et al. (2017), a computational tool that integrates multiple prediction methods, including IslandPick, IslandPath-DIMOB, and SIGI-HMM. Three regions (366,847-372,320, 1,496,613-1,523,855, 1,967,193-1,999,047) that have the potential to be PAI have been identified.

To characterize the genetic content of the predicted PAIs, gene sequences were downloaded from NCBI (GFF3 format), and corresponding protein sequences were extracted. Since a large number of hypothetical proteins were initially identified, protein sequences were queried against the UniProt database for id mapping and using BLAST to identify the top three potential identities for each protein (Altschul et al., 1990). The result, combined with annotations from NCBI and IslandViewer, was cross-referenced to make a best-fit prediction for each protein's identity.

The functional profile for each region was done through a combined analysis of manual review, homology searches, and Biomni (Huang et al., 2025).

Region 1967193–1999047 was found to be dominated by genes related to metabolic functions. The proteins within this region primarily suggest a systemic or intracellular role, potentially related to nutrient acquisition or host interaction in deeper tissues.

Region 366847–372320 contains toxins, such as hemolysin, cluster. This region is of particular interest because it is the shortest of all 3 regions and only contains 10 identified open reading frames. Most of its genes could be associated with drug resistance, such as a nitroreductase family protein, acyl-CoA thioesterase, and a metallothionein. However, the presence of an SSU ribosomal protein S20P raises questions about its classification as a classic pathogenicity island (PAI). Generally, essential genes like those for ribosomes are not typically mobile genetic elements and are not associated with virulence.

Region 1496613–1523855 exhibited the strongest characteristics of a pathogenicity island. It contains mobile genetic elements and the presence of toxin-antitoxin systems. Furthermore, this region corresponded directly to a significant fluctuation identified during the initial GC skew analysis.

The Map of the Promoters reveals coexpression and gene organisation

The *F.nucleatum* ATCC 25586 genome contains 2,067 predicted protein-coding genes based on the National Institutes of Health (NIH) database. The PePPER analysis predicts that within this genome, there are 680 promoters. Hence, the genome-wide promoter-to-gene ratio is 0.33. When estimating 3 predicts PAI region using Biopython (Cock et al., 2009), PAI 2 and PAI 3 have promoter-to-gene ratios closely matching the genome average, with 0.31 and 0.33, respectively. However, PAI 1 high a notably higher ratio of 0.7. Focusing on PAI 1 and its surrounding genomic region found a hemolysin cluster is found located between 296–307 kb, which includes PAI1. Within PAI1 itself (coordinates: 366,847–372,320 bp), the gene at locus tag FN1885 encodes hemolysin III, a pore-forming cytolytic toxin with a high pathogenic potential due to its direct membrane-damaging activity.

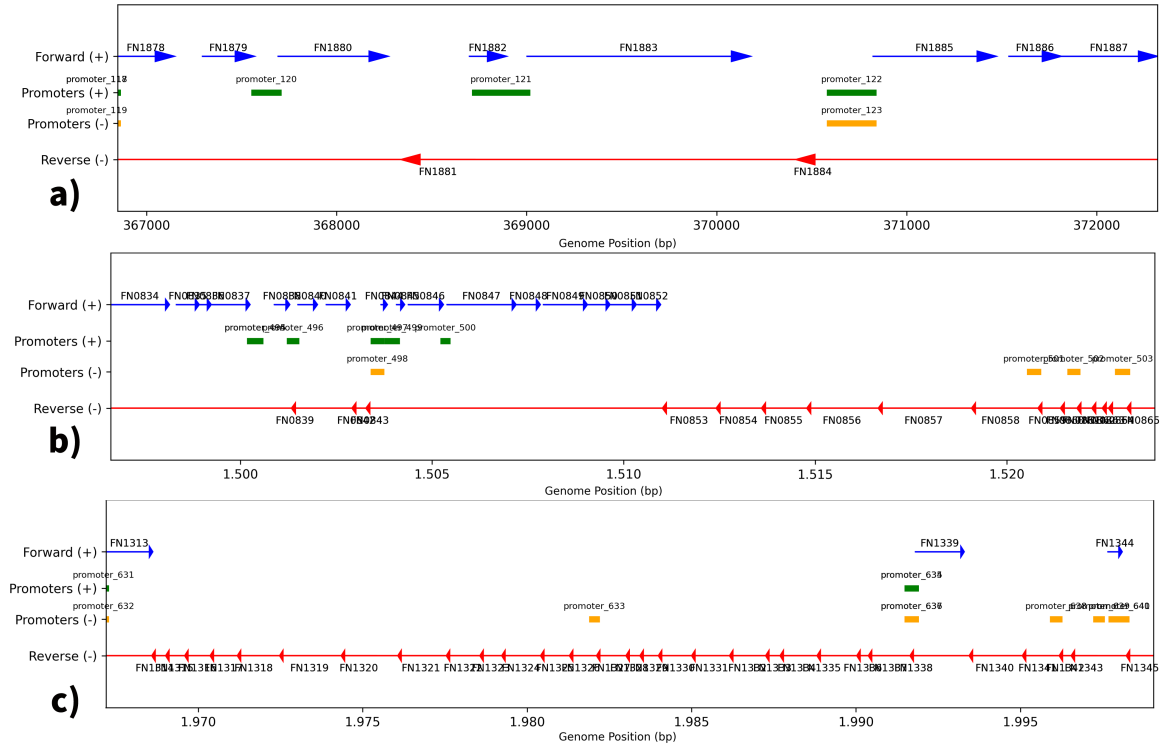

Figure 3: Promoter mapping of 3 predicts potential Pathogenicity Islands (PAIs), showing the distribution of genes on both strands and ORF. Genes are represented by arrows, with blue arrows indicating forward strand (+) genes and red arrows indicating reverse strand (−) genes. Promoters are shown as bars, with green bars representing predicted promoters on the forward strand and yellow bars representing predicted promoters on the reverse strand. a) PAI 1 (region 366847–372320) includes 7 predicted promoters (promoter_117 to promoter_123) and 10 genes (locus tags FN1878 to FN1887). b) PAI 2(region 1496613–1523855) with 10 predicted promoters (promoter_494 to promoter_503) and 32 genes (FN0834 to FN0865). c) PAI 3(region 1967193–1999047) contain 11 predicted promoters (promoter_631 to promoter_641) and 33 genes (FN1313 to FN1345).

After promoter mapping of predicted PAIs, several functional clustering sappear to demonstrate potential coexpression and co-regulation patterns. In PAI 1, the hemolysin III (locus tag FN1885) and the ots adjacent unknown protein (locus tag FN1886) on the forward strand, with only one promoter upstream, suggest a potential linkage in their transcriptional regulation. Hemolysin III (FN1885), as a pore-forming toxin, is direct proof of virulence potential by allowing host cell lysis and immune evasion. The downstream unknown protein (FN1886) could imply an accessory factor that will either enhance the function of hemolysin or represent some form of interaction with the host in an unknown and uncharacterized manner.

In PAI 2 there seems to be the biotin synthesis cluster on forward strand FN0849 (synthase) → FN0850 → FN0851 (bioC) → FN0852 (hydrolase) and the glycogen metabolism operon on the reverse strand(FN0853-FN0858, all glucose related), potentially implementing co-adaptation to metabolic processes possibly coordinating vitamin biosynthesis with energy storage/utilization under nutrient limited situations. The 2 predicted operons do not directly contribute virulence factors, but may facilitate colonization in host cells. Biotin synthesis genes (FN0849–FN0852) may aid the survival in nutrient-restricted niches where the host limits vitamin availability. The glycogen metabolism operon (FN0853–FN0858) allows energy storage, supporting persistence during fluctuating nutrient conditions. Over all both operons contribute to pathogenicity through nutritional adaptation, helping the bacterium withstand stresses imposed by its host.

In PAI 3, 2 potential coexpression clusters could be RNA polymerase sigma factors cluster on the reverse strand with FN1317 to FN1318 (rpoD) and the close link metabolic enzymes with FN1322 (metalloprotease) → FN1323 (thymidylate kinase)→ FN1324 (reductoisomerase). One could be regulatory

control of metabolic actives, while the other could function in tissue degradation (protease activity) and nucleotide/amino acid metabolism (thymidylate kinase, reductoisomerase), ensuring replication inside the host. Those 2 operons support PAI3's role as a metabolic regulator but not as a pathogenicity island.

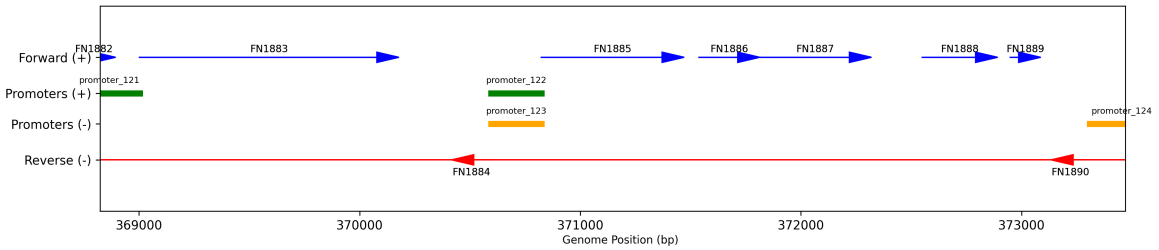

Figure 4: Genomic organization of the region spanning 368,823–373,471 bp surrounding locus tag FN1885. Blue arrows show coding sequences (ORFs) on the forward strand (+), while red arrows show on the reverse strand (−). Predicted promoters are depicted as colored bars. Green forward, yellow reverse. Promoter positions are plotted relative to their genomic coordinates and strand orientation.

Analyzing the genomic environment of the gene with locus tag FN1885 includes 9 genes and 4 putative promoters, of which FN1885 is in a forward-strand gene arrangement with 7 other genes in the same direction and has no large intergenic gaps. The reverse-strand genes (FN1884 and FN1889) are interspersed but not aligned in the same transcriptional direction, making them less likely to be part of the same operon. Within the forward strand gene, there are 2 upstream promoters, particularly promoter_122, located proximal to FN1885, indicating overlapping or sequential transcriptions and coregulation of this gene cluster. Based on the promoter positioning, FN1885 is most likely coexpressed with FN1886 to 1889, potentially in response to signals for coordinated biological function.

Codon Adaptation Index as a Proxy for Protein Abundance

High CAI values suggest strong codon optimization, likely for efficient translation. A large proportion of the top genes in the bacterial genome were ribosomal proteins (for example, L24P, S19P, L20P), which are typically highly expressed and exhibit high CAI values. Some non-ribosomal genes (e.g., flavodoxin, transcriptional regulators, hypothetical proteins) also show high CAI, which may indicate functional importance or horizontal gene transfer.

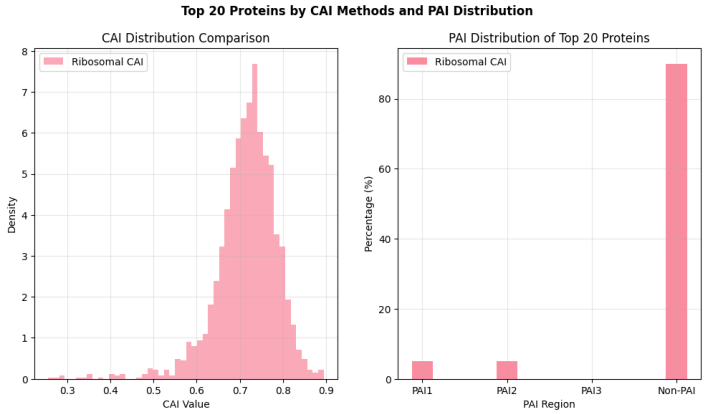

Figure 5: Codon Adaptation Index (CAI) analysis using ribosomal protein sequences as reference. Among the top 20 proteins ranked by CAI, several are located within predicted pathogenicity islands (PAI1: 366,847–372,320; PAI2: 1,496,613–1,523,855; PAI3: 1,967,193–1,999,047).

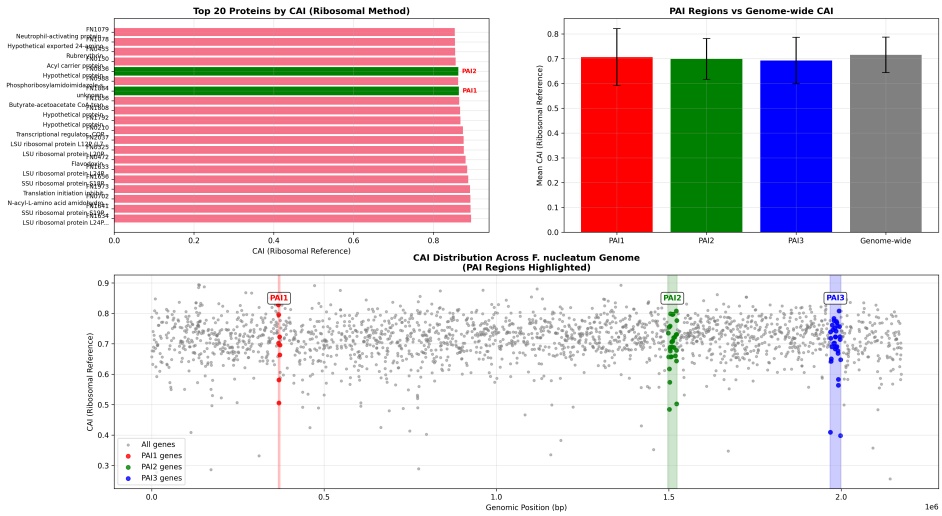

Figure 6: Distribution of the Codon Adaptation Index (CAI) along the genome of Fusobacterium nucleatum, with specific attention to potential Pathogenicity Islands (PAI1: 366,847–372,320; PAI2: 1,496,613–1,523,855; PAI3: 1,967,193–1,999,047). The analysis was performed using a ribosomal gene set as the reference for optimal codon usage.

PAI1 (region 366847–372320) contains genes with moderate to high CAI, including a ribosomal protein and a hemolysin, which may suggest functional integration into the host genome.

Table 1: Summary of genes in the predicted PAI1 region in *Fusobacterium nucleatum*, including predicted products, pathogenic potential, coexpression evidence, genomic context, and Codon Adaptation Index (CAI) values. High CAI values suggest codon optimization.

Locus Tag	Predicted Product	Pathogenic Potential	Coexpression Evidence	PAI Support	CAI
FN1885	Hemolysin III	HIGH – Direct cytolysis	Target gene	Primary virulence factor	0.722699
FN1884	Hypothetical protein	Unknown	Same strand, 406 bp upstream	Typical PAI composition	0.862872
FN1886	Hypothetical protein	Unknown	Same strand, 715 bp downstream	Typical PAI composition	0.663412
FN1881	Esterase	Medium – Membrane modification	Opposite strand, overlapping region	Metabolic modification	0.795734
FN1880	Oxygen-insensitive NAD(P)H nitroreductase	Low – Metabolic advantage	Same strand, 3.1 kb upstream	Anaerobic adaptation	0.702174
FN1879	Ribosomal protein S20P	Low – Housekeeping	Same strand, 3.5 kb upstream	Housekeeping (contradicts PAI)	0.827737
FN1882	Hypothetical protein	Unknown	Same strand, 2.1 kb upstream	Typical PAI composition	0.505487
FN1883	Transposase	Low – Mobile element	Same strand, 1.8 kb upstream	HGT evidence	0.581472
FN1887	Transposase	Low – Mobile element	Same strand, 1.0 kb downstream	HGT evidence	0.696104

CAI values ranged from 0.505 to 0.863, with most genes exceeding 0.65, indicative of moderate-to-high codon optimization relative to ribosomal gene reference sets. High CAI values in hypothetical proteins AAL93983.1 (locus tag: FN1884, CAI: 0.863) and AAL93978.1 (SSU ribosomal protein S20P, CAI: 0.828) suggest potential functional relevance, although one has uncertain annotation. At the same time, lower CAI scores in protein AAL93981.1 (locus tag: FN1882, CAI: 0.505) and AAL93982.1 (locus tag: FN1883, CAI: 0.581) may indicate either lower expression under standard conditions or acquisition from distinct genomic backgrounds via horizontal gene transfer. Functionally, the cluster contains a highly pathogenic-potential hemolysin gene and a likely coexpressed gene that could potentially facilitate in hemolysin pathway. Also, this region contains accessory hypothetical proteins and mobile genetic elements (transposases), consistent with horizontally acquired virulence modules. However, the inclusion of a ribosomal protein (housekeeping function) and metabolic enzymes (e.g., NAD(P)H nitroreductase) diverges from the traditional pathogenicity islands' composition. PAI1 is a virulence gene cluster with coordinated expression under pathogenic conditions, but whether it is a complete pathogenicity island should be proven experimentally.

PAI2 (region 1496613–1523855) mostly contains hypothetical proteins and metabolic enzymes. The high-CAI hypothetical proteins, which could be recently acquired genes that are being adapted for expression, and the presence of metabolic enzymes, suggest potential functional roles in niche adaptation.

PAI3 (region 1967193–1999047) contains genes involved in translation and nucleotide metabolism, which are typically core functions. Core functional regions are generally not integrated from other genes.

Proteomics

PAI predicted interactome

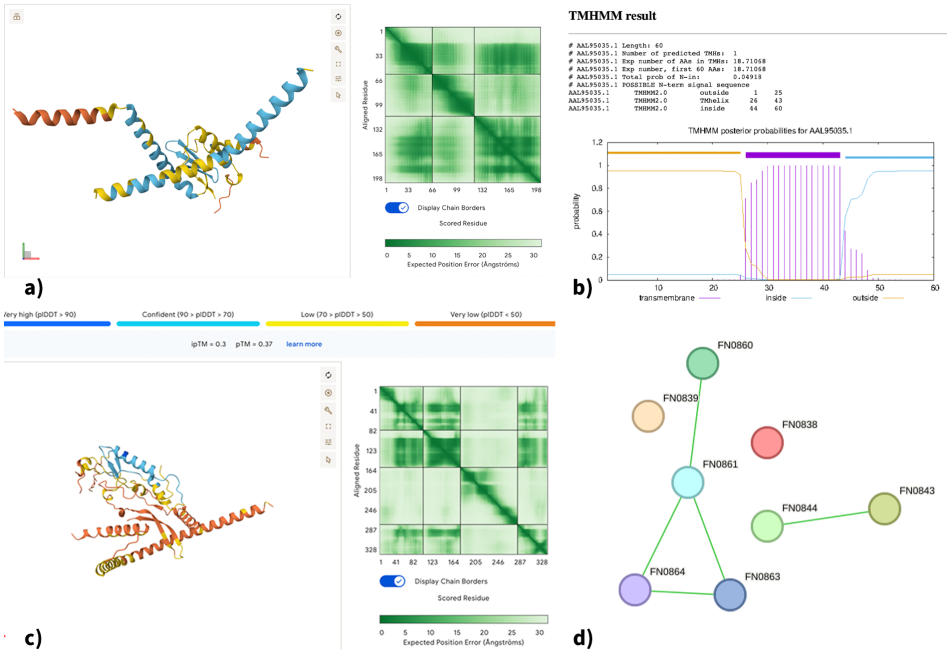

Figure 7: Predicting structural and interaction analysis for the uncharacterized proteins in region 1496613–1523855 (PAI2). a) AlphaFold-Multimer prediction of the interactions among protein AAL95035.1, AAL95040.1, and AAL95056.1 (Jumper et al., 2021). The contact map displays several off-diagonal green blocks, indicating potentially intermolecular interactions. There is a large green block between AAL95035.1 and AAL95056.1, indicating a strong interaction, whereas AAL95040.1 appears to have the weakest connection. b) TMHMM2.0 prediction of transmembrane topology for AAL95035.1 (Krogh et al., 2001). c) AlphaFold-Multimer prediction intearction of proteins AAL95056.1 (FN0860), AAL95057.1 (FN0861), AAL95059.1 (FN0863), and AAL95060.1 (FN0864). The contact map suggests that there may be several possible interaction interfaces, although the confidence of prediction is moderate and should be considered with caution. d) STRING-based protein-protein interaction network for the uncharacterized proteins within PAI2 (Szklarczyk et al., 2024). Two main clusters are predicted: one with AAL95056.1 (Locus tag: FN0860), AAL95057.1 (Locus tag: FN0861), AAL95059.1 (Locus tag: FN0863), and AAL95060.1 (Locus tag: FN0864); the other with AAL95040.1 (FN0844) and AAL95039.1 (FN0843).

The genomic region PAI2 (1496613–1523855) contains 32 protein-coding genes. They encode elements involved in mobility, such as tyrosine-type recombinase/integrase and transposases, consistent with acquisition via horizontal gene transfer. Out of these 32 genes, 8 genes encode proteins that remain uncharacterized after genome analysis. AlphaFold has been used to predict the structure and interaction of proteins encoded by these 8 genes, which suggests interaction between three proteins with id AAL95035.1, AAL95040.1, and AAL95056.1(Jumper et al., 2021). TMHMM2.0 analysis reveals that within the 8 uncharacterized proteins, AAL95035.1 and AAL95040.1 are transmembrane proteins (Krogh et al., 2001). Further analysis of these 2 proteins, AAL95035.1 and AAL95040.1, using Alphafold has shown that AAL95035.1 has the capacity for dimers or trimers, and the interaction region of the multimer form is predicted to be encoded by codons 25-41, which corresponds to the transmembrane region predicted by TMHMM2.0. In addition, AAL95035.1 interacts with another protein containing the TM2 domain (protein id AAL95058.1). It is predicted that AAL95035.1 may be involved in membrane-associated complexes. To verify if the transmembrane protein identified in PAI2 will contribute to making pore-forming toxins, experimental evidence is needed.

STRING, another protein interaction software, suggests multi-protein interaction networks be-

tween the 8 uncharacterized proteins: 1) AAL95056.1 (Locus tag: FN0860), AAL95057.1 (Locus tag: FN0861), AAL95059.1 (Locus tag: FN0863), and AAL95060.1 (Locus tag: FN0864); 2) AAL95040.1 (FN0844) and AAL95039.1 (FN0843) (Szklarczyk et al., 2024). However, these associations are not supported by AlphaFold, highlighting discrepancies between predictive tools.

The functional annotations indicate atoms associated with virulence, such as toxin-antitoxin systems (YwqK family antitoxin), a RiboL-PSP-HEPN domain-containing protein involved in stress survival, and several outer membranes and lipoproteins, which may facilitate interactions with the host. In association with these virulence predictions, the metabolic adaptation genes, such as those involved in biotin biosynthesis (8-amino-7-oxononanoate synthase) and glycogen metabolism (glycogen synthase, glucan branching enzyme, glucose-1-phosphate adenylyltransferase, and glucanotransferase), demonstrate that the identified genes could contribute to fitness in a host with limited nutrient availability. In our analyses, the results from the PAI2 indicated a combination of mobility-associated genes, putative virulence factors, genes indicating metabolic survival, as well as TM2 domain-containing protein and hypothetical proteins that were predicted to be structurally associated in membrane-associated complexes, which suggests that PAI2 could function as a pathogenicity island.

Pathogenic cluster analysis and interesting pathogenic activity

Since region 1496613–1523855 is identified as a potential PAI, regions around it that also seem to deviate significantly from the normal GC skew pattern were analyzed. Adjacent to region 1496613–1523855, a 150-kb region (1.7–1.85 Mb), which contains 136 annotated proteins, was analyzed for pathogenic potential. The same process of ID mapping, homology-based search (BLAST), and functional categorization was performed on this region. Of the 136 proteins present in this region, 37/136 (27.21%) are unannotated in the NCBI database. The incomplete characterization of this region underscores the potential presence of novel or strain-specific functions.

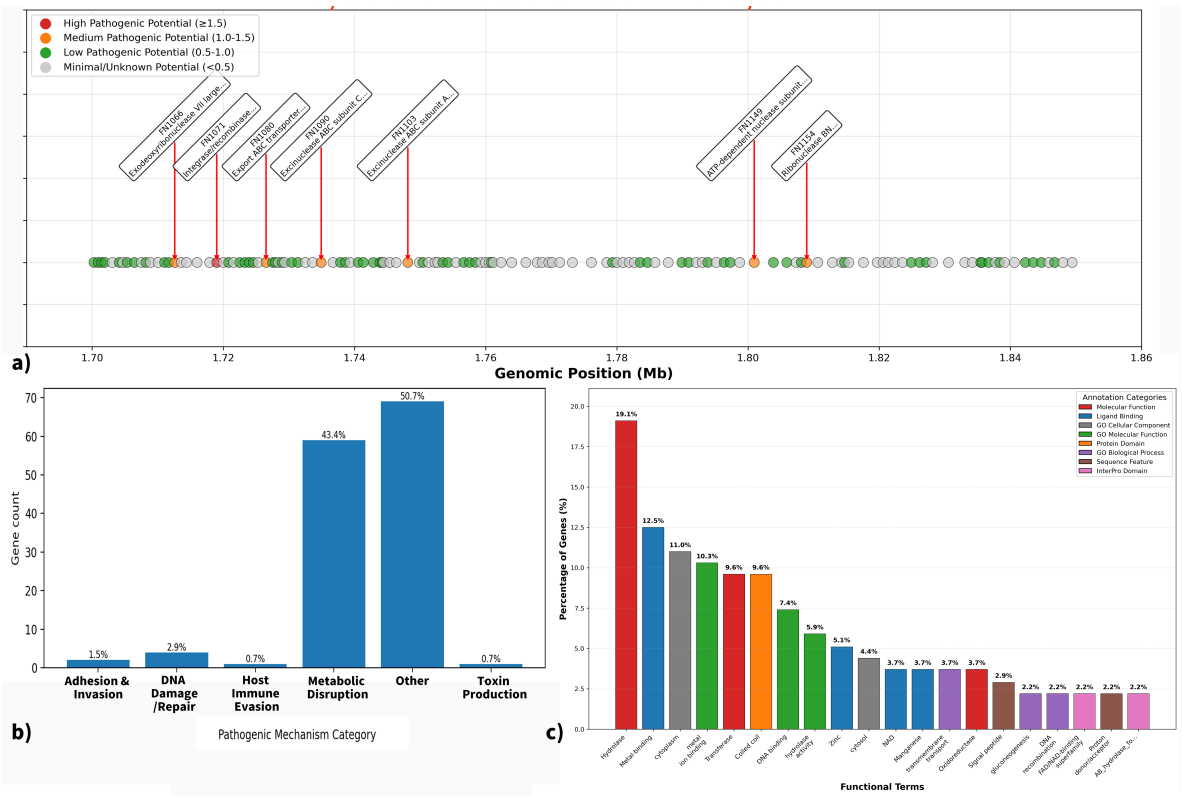

Figure 8: Functional and Pathogenic Annotation of the 1.7–1.85 Mb Genomic Region in *F.nucleatum* ATCC 25586. 1) Gene organization schematic for 1.70-1.85 Mb region. The genes of high and medium pathogenicity are labeled. Classification of potential pathogenicity was completed by Biomni's (Huang et al., 2025) annotation pipeline. b) Predicted pathogenicity mechanisms among genes in these genomic regions based on PubMed-assisted text mining. c) The functional annotation output of DAVID for the same gene set (Huang et al., 2008; Sherman et al., 2022). The total percentages are higher than 100%, as this is a feature of DAVID where gene-term associations overlap and are counted multiple times across similar or redundant terms.

Functional classification based on PubMed-assisted text mining showed that Genes for Metabolic Disruption had the highest abundance, while Genes for DNA Damage/Repair were second. This suggests that these regions may be playing a key role in the metabolic flexibility and / or genomic integrity of bacteria. DAVID functional annotation provided further support, where the most enriched categories were for Hydrolase activity (19.1%), Metal-ion binding (12.5%), and Ion Transport (11%). These functions are likely involved in host interaction, nutrient acquisition, and stress-facilitation, providing more evidence of possible pathogenic capacity. The biomni analysis needs to be interpreted cautiously, because the method of evaluating pathogenic potential is unknown, and the prediction of pathogenic protein products could potentially involve either virulence or normal cellular pathways.

Similarly, constraints are present for the PubMed-assisted text mining analysis. Most genes have roles in multiple pathways, and the same protein in different pathways may play entirely different roles. The abundance of hypothetical proteins limits functional interpretability, and automatically mined text introduces variability in the results based on the time of database indexing, revisions to entries, and the sensitivity of the keyword search. In addition, the percentage values in DAVID are greater than 100%; this is a recognized artifact resulting from overlapping gene-term associations, for instance, where genes belong to more than one functional category.

The analysis suggests that this area of the genome may represent a contribution to the overall pathogenic potential, and the metal-ion pathway potentially related to iron acquisition should be given special attention. However, these are computational results that may be strain dependent. Experimental evidence, likely involving transcriptomics or functional assays, could help validate the involvement of this region in virulence.

The iron in "bacterial economy": hemolysin and iron ions transporters

Hemolysin genes are identified throughout the genome. This gene encodes a TlyA family protein, present in phylogenetically distant bacteria, being primarily found in distantly related bacteria such as *Mycobacterium*, *Helicobacter*, and *Brachyspira* species, rather than within the Fusobacteriota phylum (Arenas et al., 2011; Rahman et al., 2015). This TlyA protein shows pathogen-specific distribution patterns, where it is observed in pathogenic species like *M. tuberculosis* and *H. pylori* but is absent in non-pathogenic relatives, suggesting a strong selective pressure for its retention in pathogens (Keith et al., 2022; Sharma and Singh, 2022). Functionally, hemolysin operates as a pore-forming cytotoxin that lyses host erythrocytes, releasing iron and nutrients, allowing F. nucleatum to acquire nutrients in inflamed periodontal pockets (Ma et al., 2023; Samainukul et al., 2019).

Hemolysin and iron acquisition mechanism in F. nucleatum

The identification of hemolysin implies that F. nucleatum (Fn) specializes in a heme utilization pathway to acquire iron from the host, indicating its ability to colonize and thrive in blood-rich environments such as periodontal pockets and tumor microenvironments.

From genome annotation, key proteins and genes involved in the heme-based iron acquisition pathway are identified. This system relies on ABC transporters (HmuUV), which comprise Hemin Transport System Permease Protein HmuU (encoded by genes hmuU_1 and hmuU_2) and the Heme Import ATP-Binding Protein HmuV (encoded by gene hmuV). HmuV, as the ATP-binding component, provides the energy to power HmuU. HmuU then facilitates the physical transport of the heme molecule from the blood and tissues across the cell membrane to the bacterial cells. While these proteins are dedicated to heme uptake, other ABC transporters for various metals, such as an Inner Membrane ABC Transporter Permease Protein (ydcV), a Zinc Import ATP-Binding Protein (znuC), Manganese ABC transporter substrate-binding lipoprotein (psaA), Iron import ATP-binding/permease protein (irtA), and many others, were also identified, highlighting a broader strategy for metal acquisition (Delepelaire, 2019; Louvel et al., 2006; McGregor et al., 2023).

Following import, the heme molecule is enzymatically degraded to release iron, a process in which flavodoxin (encoded by fldA, ispg, and isiB; highly expressed, cai of 0.880656) acts as an electron-transfer protein, reducing the ferrous complex and facilitating catalysis (Boynton et al., 2009; McGregor et al., 2023).

The iron is processed so it can be used by the cell for various cellular processes. In addition to basic metabolism sustaining bacteria, these iron acquisition mechanisms can disrupt erythrocytes and promote metastasis in colorectal cancer. The virulence protein, DNA protection during starvation protein (Dps), for instance, serves as a bacterial ferritin to protect DNA from oxidative stress while also acting as a potent pathogenic factor. Besides disrupting and lysing the host's erythrocytes to acquire iron, Dps also upregulates the chemokine CCL2/CCL7 to enhance the bacterium's survival in macrophages. The expression of CCL2/CCL7 also contributes to cancer progression by inducing epithelial-mesenchymal transition (EMT) and promoting colorectal cancer (CRC) metastasis in vivo (Wu et al., 2023).

Genomic evidence showed complete iron acquisition machinery, and *F. nucleatum* relies on iron-dependent enzymes for core metabolic processes. An additional metabolic model analysis is done to see if iron is essential for *F. nucleatum* survival and proliferation.

How scavenging iron helps metabolic growth

The annotated genome *F. nucleatum* ATCC 25586 was utilized with CarveMe (Machado et al., 2018) to reconstruct a genome-scale metabolic model, comprising 1,914 reactions and 1,339 metabolites, with 582 associated genes. Flux Balance Analysis (FBA) was applied to determine the optimal growth rate, which was predicted to be 67.42 mmol/gDW/hr under ideal nutrient-rich conditions. The flux through the biomass reaction represents the rate at which the bacterium produces those metabolites necessary for its growth (e.g., amino acids, lipids, cofactors, and proteins). The biomass flux returned by COBRA is the specific growth rate that can be used to calculate the specific biomass of the bacteria at a specific time using an exponential function. Therefore, biomass flux can be used as an indicator of

the growth rate of the bacterium. Iron is required in a minimal medium for maintaining the growth rate proven by the CarveMe-generated metabolic model. The model identified uptake fluxes of 0.452702 mmol/gDW/h for ferrous iron (EX_fe2_e) and 0.524388 mmol/gDW/h for ferric iron (EX_fe3_e) as necessary to maintain the organism's growth rate in a minimal medium.

Although it contains iron requirements in the composition of the biomass and the model contains reactions involving iron metabolites, COBRApy analyses performed using the metabolic model generated by Carveme evaluate biomass production under all iron conditions, including complete absence of iron, limitation of iron uptake, and unlimited iron availability, give identical results (67.416545 mmol / gDW / hour). This indicates missing iron dependencies or unrealistic constraints in the model.

The failure of the stimulation could be explained by FBA's inherent reliance on functional optimization in the face of biological redundancy. The metabolic networks contain multiple iron-chelate compounds (enterobactin, salmochelin, heme) and alternative metabolic pathways that can perform equivalent functions, masking true iron dependencies. In this simulation, only the EX_fe2_e and EX_fe3_e are set to (0,0), while other iron exchange reactions allow unrealistic, unlimited iron uptake. The FBA algorithm, pursuing optimization of biomass flux output, can simply re-route metabolic flux through these redundant pathways. Consequently, the simulation predicts no impact on growth, leading to a classification of the gene as non-essential (Boone, 2025).

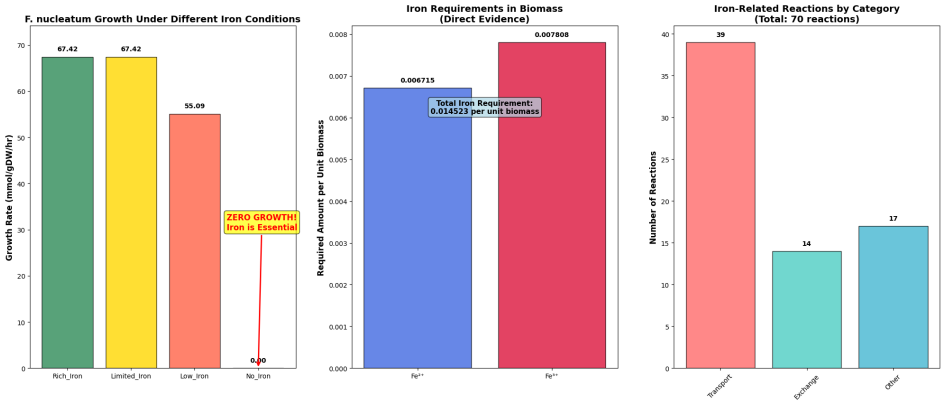

Figure 9: Multi-panel visualization that provides a comprehensive overview of iron's essentiality for *F. nucleatum*, FBA flux analyzed using cobrapy. The panels show that iron deprivation completely halts growth, while rich and limited iron conditions support high growth rates, emphasizing iron's metabolic importance. Additionally, the presence of diverse iron acquisition systems in the model supports iron critical role in survival and virulence.

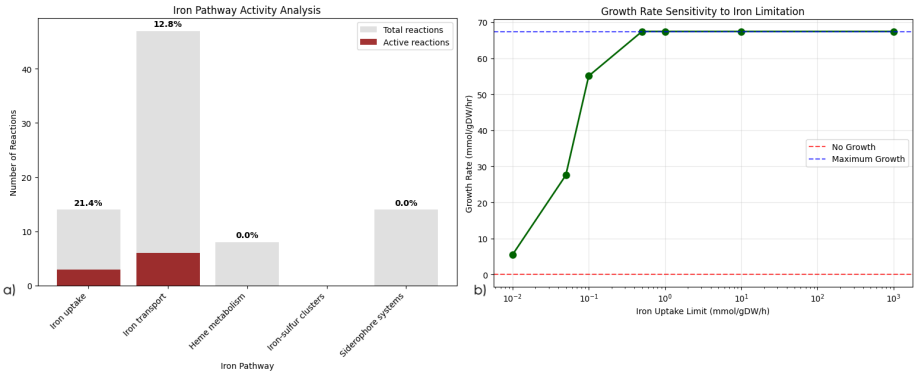

Figure 10: Detailed flux analysis, sensitivity curves, and pathway activity, using COBRApy. a) Bar graph compares the total and active reactions across five iron-related metabolic pathways: Iron Uptake, Iron Transport, Heme Metabolism, Iron-Sulfur Clusters, and Siderophore Systems. Grey bars represent the total number of reactions, while red bars indicate the subset of reactions that are active under the studied conditions. Percentages above each bar denote the proportion of active reactions, highlighting differential pathway utilization and potential regulatory prioritization. b) This plot illustrates the relationship between iron uptake limits and cellular growth rate. The x-axis shows iron uptake limits (mmol/gDW/hr) on a logarithmic scale, while the y-axis indicates growth rate (mmol/gDW/hr).

Identifying this potential limitation, 70 reactions and 40 metabolites related to iron across all three cellular compartments (cytoplasm, periplasm, and extracellular space) are used as the new constraints. The refined code indicates iron is required in biomass composition, for *F. nucleatum* cannot produce biomass without iron. FBA under different iron availability conditions is then performed. The model predicts that complete iron deprivation results in the arrest of growth. The metabolic consequences of iron limitation show that although iron is essential for bacterial growth, it only requires a limited amount of iron for maximum growth ($< 1\ mmol/gDW/h$). Under iron-limited conditions, the model shows a compensatory activation of multiple siderophore systems to maximize scavenging, although this still results in an 18.3% reduction in growth rate. In addition, the bacterium possesses complex iron acquisition mechanisms that involve direct iron transport (ferrous iron (Fe^{2+} transport, ferric iron (Fe^{3+}) transport, and ABC transporters), siderophore-mediated iron acquisition (enterobactin system, salmochelin system, and staphyloferrin systems), and heme iron uptake.

Furthermore, this bacterium demonstrates a strong preference for ferric iron (Fe^{3+}) over ferrous iron (Fe^{2+}). An examination of iron transport systems in *F. nucleatum* using the metabolic model generated by Carvme from annotated Prokka through an AI-assisting tool, Biomni. The results suggested that 77.7% of total iron transport within the model is dedicated to Fe^{3+} uptake, with direct ferric iron uptake flux being 0.526, ferripyoverdine complex formation flux being 0.526, and Fe^{3+}-pyoverdine exchange flux being 0.526. The direct ferrous iron transport flux is only 0.453.

Ferrous iron (Fe^{2+}) is usually more soluble over a larger pH range than ferric iron (Fe^{3+}), which hydrolyzes and precipitates as iron(III) hydroxide at near-neutral and higher pH values. The preference of *F. nucleatum* for Fe^{3+} may represent a metabolic adaptation to obtain iron in acidic niches, typically seen in tumor microenvironments.

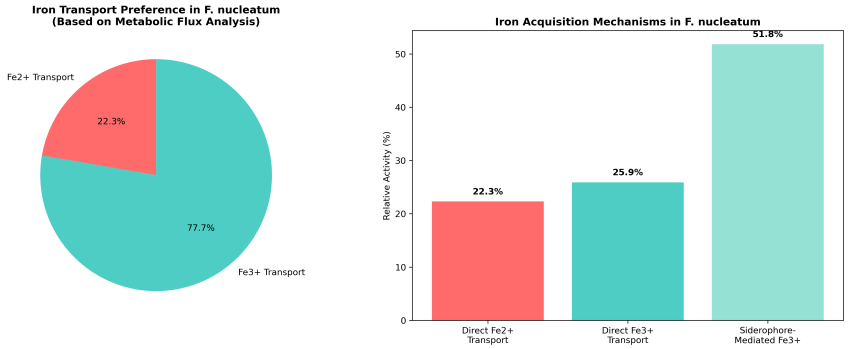

Figure 11: Examination of iron transport systems in *F. nucleatum* using the metabolic model generated by Carvme from annotated Prokka through an AI-assisting tool, Biomni.(Huang et al., 2025)

F. nucleatum and cancer

F.nucleatum, though not the direct cause of cancer, is frequently found in cancer tissue as well as associated with promoting tumor growth and accelerating metastasis. Literature review suggests *F. nucleatum* has a symbiotic relationship with the cancer cells, promoting tumor progression, supporting colonization in the cells, and altering the immune response within the TME. High levels of *F. nucleatum* establish a positive correlation with the number of tumor-promoting immune cells, including tumor-associated macrophages (TAMs), tumor-associated neutrophils (TANs), dendritic cells (DCs), and myeloid-derived suppressor cells (MDSCs), and a decrease in the function of T-cells and natural killer (NK) cells. *F. nucleatum* contains virulence factors, such as Fap2, FadA, RadD, and FomA, which enable it to interact with cancer cells, modulating key signaling pathways, altering cytokine secretion, and promoting cancer progression. This bacterium can also secrete outer membrane vesicles that facilitate metastasis by promoting autophagy in oral cancer (Rubinstein et al., 2013; Willenbockel et al., 2024).

Metabolic analysis has revealed that iron is essential for F. nucleatum. Through genomic analysis of *F. nucleatum* strain ATCC 25586, direct cytotoxicity virulence factors are identified (shlB_3, shlB_2, shlB_1, tlyA). Iron acquisition by *F. nucleatum* through hemolysis is purpose to be responsible for *F. nucleatum* modulation of the hippo pathway, which allows *F. nucleatum* to play a role in cancer progression and therapy resistance (Han et al., 2020; Millot et al., 2016; Wang et al., 2024).

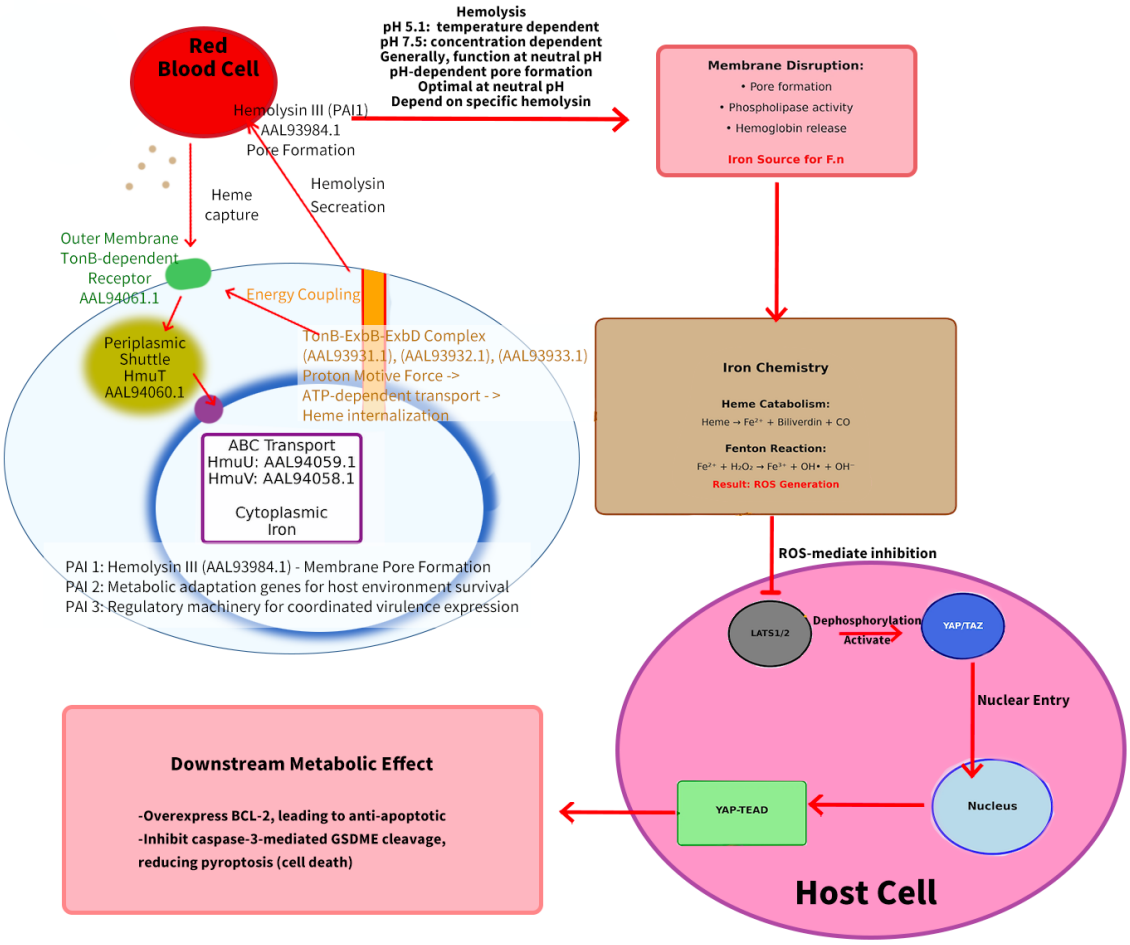

Figure 12: *F. nucleatum*'s iron acquisition through a hemolysin-mediated mechanism can modulate the Hippo signaling pathway. Hemolysin secretion and pore formation are pH-dependent, with optimal activity near neutral pH (Kasschau et al., 1995). Contributions from three predicted PAIs are noted, and key proteins involved in this pathway, including TonB-dependent receptors, HmuTUV ABC transporters, and the TonB-ExbB-ExbD energy transduction system, with their respective protein IDs, are labeled (Braun et al., 2023; Richard et al., 2019; Silale and van den Berg, 2023). Intracellular heme catabolism liberates iron, which participates in the Fenton reaction, generating reactive oxygen species (ROS) (Xuan et al., 2025; Zhao, 2019). ROS inhibit LATS1/2 activity, leading to YAP/TAZ dephosphorylation and nuclear entry. Activated YAP-TEAD signaling drives anti-apoptotic responses by upregulating BCL-2 and suppressing caspase-3–mediated GSDME cleavage, thereby reducing host cell pyroptosis(Amanda et al., 2024; Chen et al., 2017; Wang et al., 2024).

Hemolysis is generally associated with iron overload and excess hemoglobin in the bloodstream. Hemolysins lyse red blood cells, releasing iron, indirectly increasing iron concentration in blood (Millot et al., 2016). Free iron undergoes Fenton reactions, generating reactive oxygen species (ROS) such as hydroxyl radicals $Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH \cdot + OH^-$, causing oxidative damage to lipids, proteins, and DNA (Xuan et al., 2025; Zhao, 2019).

Iron-induced oxidative stress inhibits LATS1/2 kinases, which dephosphorylate and nuclear translocation YAP and TAZ, the downstream effectors of the Hippo pathway. YAP/TAZ overactivation can promote pluripotency, tumor initiation, and chemotherapy resistance, which are common signs in cancer cells. They upregulate multidrug transporters (e.g., ABCB1, ABCC1, ABCG2) and anti-apoptotic proteins (e.g., Bcl-xL, cIAP1, Survivin), while suppressing p53 via MAPK/NF-ϰB signaling. Furthermore, YAP/TAZ activate TEAD-mediated transcription, and BCL-2 is a downstream target of the

YAP-TEAD1 transcriptionally active complex (Amanda et al., 2024; Moroishi et al., 2015; Nguyen and Yi, 2019; Xuan et al., 2025). BCL-2 regulates apoptosis, and its overexpression is found in over 50% of oral squamous cell carcinoma (OSCC) tissues. The balance between pro-apoptotic (e.g., BAK, BAD) and anti-apoptotic (e.g., BCL-2, MCL-1) proteins determines tumor cell survival (Chen et al., 2017). *F. nucleatum*'s iron-mediated activation of YAP/TAZ may thus contribute to OSCC progression and resistance to anti-tumor therapies.

Further annotation of the Prokka-annotated genome was performed using eggNOG-mapper (see supplementary file `FN.annotations.xlsx`) (Huerta-Cepas et al., 2018). The result revealed the presence of the adhesin protein FadA (Fusobacterium adhesin A), but didn't direct the identification of Fap2 or RadD in their classic forms. However, related proteins with similar functions were identified.

Two FadA proteins were found in the dataset (`EJMLJHKB_00032` and `EJMLJHKB_00869`), both annotated as "Adhesion protein FadA" with the FadA protein family domain (PFAMs: FadA); several other proteins with the FadA binding domain were also identified. FadA can activate the β-catenin and Wnt pathway via the binding to E-cadherin on the surface of CRC cells, causing relocation. Under stress conditions, FadA increase pathogenicity via conformational changes (Chen et al., 2022; Rubinstein et al., 2013).

While direct Fap2 was not found, several galactose-related proteins were found, including galactose kinase (`EJMLJHKB_00595`) and galactose-1-phosphate uridyl transferase (`EJMLJHKB_00596`), along with trimeric autotransporter adhesins (`EJMLJHKB_01075` and `EJMLJHKB_01330`) which have structural homology with Fap2. Fibroblast activation protein 2 (Fap2), an outer membrane protein from F. nucleatum, is a galactose-sensitive hemagglutinin and adhesive protein that facilitate bacterial co-aggregation and host cell adhesion through binding to D-galactose-β (1-3)-N-acetyl-D-galactosamine (Gal-GalNAc) on host cell surfaces. Fap2 can bind to inhibitory receptors on immune cells, preventing immune recognition and facilitating immune evasion (Coppenhagen-Glazer et al., 2015; Ganesan et al., 2019). Although RadD is not directly present, multiple DNA repair and recombination proteins were identified, including RecG helicase (`EJMLJHKB_00161`), DNA topoisomerases (`EJMLJHKB_00613`, `EJMLJHKB_00614`), transcription-repair coupling factor (`EJMLJHKB_00637`), and Holliday junction resolvase RuvC (EJMLJHKB_00821). RadD-related proteins function in DNA repair and recombination pathways. The identified protein also has a role in both of these pathways.

Cancer cells generally create a more acidic microenvironment than normal tissues, which favors their progression (Tafech and Stéphanou, 2024). Understanding how *F. nucleatum* responds to changes in environmental pH, providing critical insights into whether this pathogenic bacterium thrives in cancer environments and how its metabolic outputs could influence cancer progression.

Based on experimental literature, *F. nucleatum* demonstrates optimal growth at pH 7.4, a relatively neutral pH. The bacterium exhibits significantly reduced metabolic activity, with glucose consumption decreasing to less than 50% and cell yield being halved at pH 6.8. Under moderate acidic conditions (pH 5.8), growth is substantially reduced, while under slightly alkaline conditions (pH 7.8), growth fails (Rogers et al., 1991). While Rogers's team (1991) found a narrow optimal range, other studies indicate a broader survival capacity. *F. nucleatum* subsp. polymorphum demonstrates the ability to form biofilms at alkaline pH 8.2, indicating subspecies-specific alkaline adaptation capabilities(Chew et al., 2012). However, *F. nucleatum* ATCC 25586 (subsp. nucleatum), the strain commonly found in oral cavities, probably exhibits tolerance to alkaline conditions, as *F. nucleatum* is found to survive in root canal systems at pH 9.0 after $Ca(OH)_2$ treatment (Zilm and Rogers, 2007). Furthermore, a key literature finding suggests *F. nucleatum* possesses extreme acid tolerance, due to its cell membrane containing erucic acid, which is regulated by the FnFabM protein, giving it the ability to resist the extreme acidic conditions of the stomach (pH 1.5) when translocated into the intestine (Li et al., 2025).

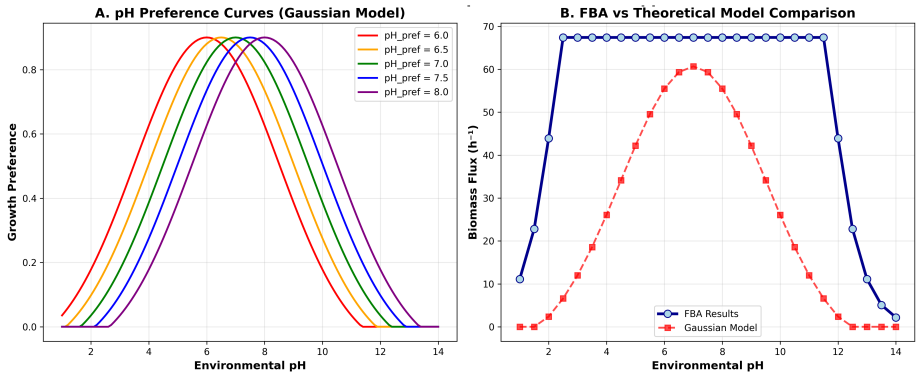

Figure 13: This figure presents an analysis of *F. nucleatum* growth across environmental pH conditions using Gaussian preference functions and differential equation modeling, visualizing the effect of pH on the growth and metabolic activity of *F.nucleatum*.

pH-dependent growth modeling employed Gaussian preference functions shows *F. nucleatum* can maintain growth across a wide pH range (1.0-14.0) and maintain a maximum growth rate of 67.4165 mmol/gDW/hr at a pH of 2.5-9.5. The Gaussian Model analysis suggested optimal growth around pH 7.5, which aligns with the literature.

In addition, 786 pH-sensitive reactions are identified, with a predominance of acid-producing reactions (149 reactions) compared to base-producing reactions (52 reactions). At the optimal growth rate, the acid production flux is 2040, and the base production flux is 190. The major difference suggests that though *F. nucleatum* prefers moderate pH conditions, it actively lowers its environmental pH through acid production.

Tumor microenvironments are generally more acidic compared to normal tissue. Under this acidic condition, *F. nucleatum* in intestinal tissue undergoes significant proteomic remodeling, with approximately 70% of the bacterium's outer membrane vesicle (OMV) protein expression altered. In addition, several virulence factors—including type 5a secreted autotransporters (T5aSSs) and MORN domain-containing proteins—are upregulated in low pH environments, potentially enhancing the bacterium's pathogenicity in cancer-associated niches (Zhang et al., 2023). On the other hand, oral cancers are also characterized by an acidic tumor microenvironment, where the pH of OSCC cells has been reported to range between 6.56 and 6.97 (Liu et al., 2024). The *F. nucleatum* strain in this study appears to be insensitive to environmental pH in terms of growth, suggesting a degree of physiological resilience or adaptation.

Conclusion

As the amount of molecular sequence data of pathogens in the public domain grows, so does the range of bioinformatics and machine learning methods that operate on them through functional, structural and evolutionary considerations. A number of developments in both bioinformatics and machine learning have enabled methodology to keep pace and are producing the promised advances in the understanding of the molecular mechanisms of pathogenicity. Here, we show how a variety of methodologies from both fields enables to provide further details into the pathogenic mechanisms of the bacterium *F. nucleatum* which is often associated with oral and colon cancer. The main contribution of this work is the suggestions of experimental design to refute the bioinformatics and machine learning findings.

References

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/s0022-2836(05)80360-2. URL http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Bella Amanda, Rangga Pragasta, Haris Cakrasana, Arifa Mustika, Zakiyatul Faizah, and Delvac Oceandy. The hippo signaling pathway, reactive oxygen species production, and oxidative stress: A two-way traffic regulation. *Cells*, 13(22):1868, November 2024. ISSN 2073-4409. doi: 10.3390/cells13221868. URL http://dx.doi.org/10.3390/cells13221868.

Nelson E. Arenas, Luz M. Salazar, Carlos Y. Soto, Carolina Vizcaíno, Manuel E. Patarroyo, Manuel A. Patarroyo, and Arley Gómez. Molecular modeling and in silico characterization of mycobacterium tuberculosis tlya: Possible misannotation of this tubercle bacilli-hemolysin. *BMC Structural Biology*, 11:16, 2011. doi: 10.1186/1472-6807-11-16.

Franco Bagnoli and Pietro Liò. Selection, mutations and codon usage in a bacterial model. *Journal of Theoretical Biology*, 173(3):271–281, 1995.

Claire Bertelli, Matthew R Laird, Kelly P Williams, Britney Y Lau, Gemma Hoad, Geoffrey L Winsor, and Fiona SL Brinkman. Islandviewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, 45(W1):W30–W35, May 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx343. URL http://dx.doi.org/10.1093/nar/gkx343.

J. Boone. A topology-based machine learning model decisively outperforms flux balance analysis in predicting metabolic gene essentiality. *ArXiv.org*, 2025. URL https://arxiv.org/abs/2507.20406. Accessed via arXiv.

Tye O. Boynton, Lauren E. Daugherty, Tamara A. Dailey, and Harry A. Dailey. Identification of escherichia coli hemg as a novel, menadione-dependent flavodoxin with protoporphyrinogen oxidase activity. *Biochemistry*, 48(29):6705–6711, July 2009. ISSN 1520-4995. doi: 10.1021/bi900850y. URL http://dx.doi.org/10.1021/bi900850y.

Volkmar Braun, Anna C. Ratliff, Herve Celia, and Susan K. Buchanan. Energization of outer membrane transport by the exbb exbd molecular motor. *Journal of Bacteriology*, 205(6), June 2023. ISSN 1098-5530. doi: 10.1128/jb.00035-23. URL http://dx.doi.org/10.1128/jb.00035-23.

Xiyan Chen, Weiting Gu, Qi Wang, Xucheng Fu, Ying Wang, Xin Xu, and Yong Wen. C-myc and bcl-2 mediate yap-regulated tumorigenesis in oscc. *Oncotarget*, 9(1):668–679, December 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.23089. URL http://dx.doi.org/10.18632/oncotarget.23089.

Yuanxin Chen, Zhijie Huang, Zhengming Tang, Yisheng Huang, Mingshu Huang, Hongyu Liu, Dirk Ziebolz, Gerhard Schmalz, Bo Jia, and Jianjiang Zhao. More than just a periodontal pathogen –the research progress on fusobacterium nucleatum. *Frontiers in Cellular and Infection Microbiology*, 12, February 2022. ISSN 2235-2988. doi: 10.3389/fcimb.2022.815318. URL http://dx.doi.org/10.3389/fcimb.2022.815318.

Zhenzhen Chen and Leaf Huang. Fusobacterium nucleatum carcinogenesis and drug delivery interventions. *Advanced Drug Delivery Reviews*, 209:115319, June 2024. ISSN 0169-409X. doi: 10.1016/j.addr.2024.115319. URL http://dx.doi.org/10.1016/j.addr.2024.115319.

Jactty Chew, Peter S Zilm, Janet M Fuss, and Neville J Gully. A proteomic investigation of fusobacterium nucleatum alkaline-induced biofilms. *BMC Microbiology*, 12(1), September 2012. ISSN 1471-2180. doi: 10.1186/1471-2180-12-189. URL http://dx.doi.org/10.1186/1471-2180-12-189.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, March 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL http://dx.doi.org/10.1093/bioinformatics/btp163.

S. Coppenhagen-Glazer, A. Sol, J. Abed, R. Naor, X. Zhang, Y. W. Han, and G. Bachrach. Fap2 of fusobacterium nucleatum is a galactose-inhibitable adhesin involved in coaggregation, cell adhesion, and preterm birth. *Infection and Immunity*, 83(3):1104–1113, March 2015. ISSN 1098-5522. doi: 10.1128/iai.02838-14. URL http://dx.doi.org/10.1128/IAI.02838-14.

Anne de Jong, Hilco Pietersma, Martijn Cordes, Oscar P Kuipers, and Jan Kok. Pepper: a web-server for prediction of prokaryote promoter elements and regulons. *BMC Genomics*, 13(1), July 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-299. URL http://dx.doi.org/10.1186/1471-2164-13-299.

Philippe Delepelaire. Bacterial abc transporters of iron containing compounds. *Research in Microbiology*, 170(8):345–357, November 2019. ISSN 0923-2508. doi: 10.1016/j.resmic.2019.10.008. URL http://dx.doi.org/10.1016/j.resmic.2019.10.008.

Kumar Ganesan, Songhe Guo, Sundaz Fayyaz, Ge Zhang, and Baojun Xu. Targeting programmed fusobacterium nucleatum fap2 for colorectal cancer therapy. *Cancers*, 11(10):1592, October 2019. ISSN 2072-6694. doi: 10.3390/cancers11101592. URL http://dx.doi.org/10.3390/cancers11101592.

Hu Han, Yan Li, Wan Qin, Lu Wang, Han Yin, Beibei Su, and Xianglin Yuan. Fusobacterium nucleatumfacilitates cetuximab resistance in colorectal cancer via the pi3k/akt and jak/stat3 pathways. December 2020. doi: 10.1101/2020.12.21.423755. URL http://dx.doi.org/10.1101/2020.12.21.423755.

Yiping W Han. Fusobacterium nucleatum: a commensal-turned pathogen. *Current Opinion in Microbiology*, 23:141–147, February 2015. ISSN 1369-5274. doi: 10.1016/j.mib.2014.11.013. URL http://dx.doi.org/10.1016/j.mib.2014.11.013.

Kah Yan How, Keang Peng Song, and Kok Gan Chan. Porphyromonas gingivalis: An overview of periodontopathic pathogen below the gum line. *Frontiers in Microbiology*, 7, February 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00053. URL http://dx.doi.org/10.3389/fmicb.2016.00053.

Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, December 2008. ISSN 1750-2799. doi: 10.1038/nprot.2008.211. URL http://dx.doi.org/10.1038/nprot.2008.211.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical ai agent. June 2025. doi: 10.1101/2025.05.30.656746. URL http://dx.doi.org/10.1101/2025.05.30.656746.

Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, November 2018. ISSN 1362-4962. doi: 10.1093/nar/gky1085. URL http://dx.doi.org/10.1093/nar/gky1085.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Margaret R. Kasschau, Mary P. Byam-Smith, Deborah S. Gentry, and Fay N. Watson. Influence of ph and temperature on hemolysis by adult schistosoma mansoni membranes. *Journal of Experimental Zoology*, 271(4):315–322, March 1995. ISSN 1097-010X. doi: 10.1002/jez.1402710409. URL http://dx.doi.org/10.1002/jez.1402710409.

Brandon A. Keith, John C.S. Harding, and Matthew E. Loewen. Mutational analysis of tlya from brachyspira hampsonii reveals two key residues conserved in pathogenic bacteria responsible for oligomerization and hemolytic activity. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1866(2):130045, February 2022. ISSN 0304-4165. doi: 10.1016/j.bbagen.2021.130045. URL http://dx.doi.org/10.1016/j.bbagen.2021.130045.

Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001. doi: 10.1006/jmbi.2000.4315.

B. D. Lee. Python implementation of codon adaptation index. *Journal of Open Source Software*, 3(30): 905, 2019. doi: 10.21105/joss.00905. URL https://cai.readthedocs.io/en/latest/api.html.

Xiaocong Li, Shipeng Zhang, Huafang Sheng, Yan Zhen, Buling Wu, Zhuang Li, Dingqiang Chen, and Hongwei Zhou. Oral fusobacterium nucleatum resists the acidic ph of the stomach due to membrane erucic acid synthesized via enoyl-coa hydratase-related protein fnfabm. *Journal of Oral Microbiology*, 17(1), January 2025. ISSN 2000-2297. doi: 10.1080/20002297.2025.2453964. URL http://dx.doi.org/10.1080/20002297.2025.2453964.

Serene Lim, Wan-Hsin Hsueh, Ni-Hung Wu, Claire Hodges, Christopher Vuong, Kynzi Smith, Jenn-Ren Hsiao, Jeffrey S Chang, Jang-Yang Chang, Jenn-Wei Chen, and I-Hsiu Huang. Characterization of clinical fusobacterium nucleatum isolates from oral squamous cell carcinoma patients. *ArXiv.org*, January 2025. doi: 10.1101/2025.01.08.631950. URL http://dx.doi.org/10.1101/2025.01.08.631950.

Shiyu Liu, Jing Chen, Xuedong Zhou, Yu Hao, Yawen Zong, Yangyang Shi, Xiao Guo, Qi Han, Mingyun Li, Bolei Li, and Lei Cheng. A ph-sensitive smart monomer prevents oral cancer progression. *Advanced NanoBiomed Research*, 4(4), January 2024. ISSN 2699-9307. doi: 10.1002/anbr.202300119. URL http://dx.doi.org/10.1002/anbr.202300119.

H. Louvel, S. Bommezzadri, N. Zidane, C. Boursaux-Eude, S. Creno, A. Magnier, Z. Rouy, C. Médigue, I. Saint Girons, C. Bouchier, and M. Picardeau. Comparative and functional genomic analyses of iron transport and regulation in leptospira spp. *Journal of Bacteriology*, 188(22):7893–7904, November 2006. ISSN 1098-5530. doi: 10.1128/jb.00711-06. URL http://dx.doi.org/10.1128/JB.00711-06.

Xiaomei Ma, Tianyong Sun, Jiannan Zhou, Mengfan Zhi, Song Shen, Yushang Wang, Xiufeng Gu, Zixuan Li, Haiting Gao, Pingping Wang, and Qiang Feng. Pangenomic study of fusobacterium nucleatum reveals the distribution of pathogenic genes and functional clusters at the subspecies and strain levels. *Microbiology Spectrum*, 11(3), June 2023. ISSN 2165-0497. doi: 10.1128/spectrum.05184-22. URL http://dx.doi.org/10.1128/spectrum.05184-22.

Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, June 2018. ISSN 1362-4962. doi: 10.1093/nar/gky537. URL http://dx.doi.org/10.1093/nar/gky537.

Alexandra K. McGregor, Anson C.K. Chan, Megan D. Schroeder, Long T.M. Do, Gurpreet Saini, Michael E.P. Murphy, and Kirsten R. Wolthers. A new member of the flavodoxin superfamily from fusobacterium nucleatum that functions in heme trafficking and reduction of anaerobilin. *Journal of Biological Chemistry*, 299(7):104902, July 2023. ISSN 0021-9258. doi: 10.1016/j.jbc.2023.104902. URL http://dx.doi.org/10.1016/j.jbc.2023.104902.

Sarah Millot, Constance Delaby, Boualem Moulouel, Thibaud Lefebvre, Nathalie Pilard, Nicolas Ducrot, Cécile Ged, Philippe Lettéron, Lucia de Franceschi, Jean Charles Deybach, Carole Beaumont, Laurent Gouya, Hubert De Verneuil, Saïd Lyoumi, Hervé Puy, and Zoubida Karim. Hemolytic anemia repressed hepcidin level without hepatocyte iron overload: lesson from günther disease model. *Haematologica*, 102(2):260–270, November 2016. ISSN 1592-8721. doi: 10.3324/haematol.2016.151621. URL http://dx.doi.org/10.3324/haematol.2016.151621.

Marwan Mansoor Ali Mohammed, Veronika Kuchařová Pettersen, Audun H. Nerland, Harald G. Wiker, and Vidar Bakken. Protein interactions between the oral bacteriafusobacterium nucleatumandporphyromonas gingivalisin biofilm and planktonic culture. November 2020. doi: 10.1101/2020.11.23.393850. URL http://dx.doi.org/10.1101/2020.11.23.393850.

Toshiro Moroishi, Hyun Woo Park, Baodong Qin, Qian Chen, Zhipeng Meng, Steven W. Plouffe, Koji Taniguchi, Fa-Xing Yu, Michael Karin, Duojia Pan, and Kun-Liang Guan. A yap/taz-induced feedback mechanism regulates hippo pathway homeostasis. *Genes amp; Development*, 29(12):1271–1284,

June 2015. ISSN 1549-5477. doi: 10.1101/gad.262816.115. URL http://dx.doi.org/10.1101/gad.262816.115.

Chan D.K. Nguyen and Chunling Yi. Yap/taz signaling and resistance to cancer therapy. *Trends in Cancer*, 5(5):283–296, May 2019. ISSN 2405-8033. doi: 10.1016/j.trecan.2019.02.010. URL http://dx.doi.org/10.1016/j.trecan.2019.02.010.

Akash Patel, Monica Patel, and Joseph Glowacki. Fusobacterium nucleatum: More than just an oral anaerobe. *Cureus*, July 2022. ISSN 2168-8184. doi: 10.7759/cureus.26989. URL http://dx.doi.org/10.7759/cureus.26989.

Jessica Queen, Zam Cing, Hana Minsky, Asmita Nandi, Taylor Southward, Jacqueline Ferri, Madison McMann, Thevambiga Iyadorai, Jamuna Vadivelu, April Roslani, Mun Fai Loke, Jane Wanyiri, James R. White, Julia L. Drewes, and Cynthia L. Sears. Fusobacterium nucleatum is enriched in invasive biofilms in colorectal cancer. *npj Biofilms and Microbiomes*, 11(1), May 2025. ISSN 2055-5008. doi: 10.1038/s41522-025-00717-7. URL http://dx.doi.org/10.1038/s41522-025-00717-7.

Md. Aejazur Rahman, Parveen Sobia, Ved Prakash Dwivedi, Aakansha Bhawsar, Dhiraj Kumar Singh, Pawan Sharma, Prashini Moodley, Luc Van Kaer, William R Bishai, and Gobardhan Das. Mycobacterium tuberculosis tlya protein negatively regulates t helper (th) 1 and th17 differentiation and promotes tuberculosis pathogenesis. *Journal of Biological Chemistry*, 290(23):14407–14417, June 2015. ISSN 0021-9258. doi: 10.1074/jbc.m115.653600. URL http://dx.doi.org/10.1074/jbc.M115.653600.

Christoph Ratzke and Jeff Gore. Modifying and reacting to the environmental ph can drive bacterial interactions. *PLOS Biology*, 16(3):e2004248, March 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2004248. URL http://dx.doi.org/10.1371/journal.pbio.2004248.

Kaylie L. Richard, Brittni R. Kelley, and Jeremiah G. Johnson. Heme uptake and utilization by gram-negative bacterial pathogens. *Frontiers in Cellular and Infection Microbiology*, 9, March 2019. ISSN 2235-2988. doi: 10.3389/fcimb.2019.00081. URL http://dx.doi.org/10.3389/fcimb.2019.00081.

A. H. Rogers, P. S. Zilm, N. J. Gully, A. L. Pfennig, and P. D. Marsh. Aspects of the growth and metabolism of fusobacterium nucleatum atcc 10953 in continuous culture. *Oral Microbiology and Immunology*, 6(4):250–255, August 1991. ISSN 1399-302X. doi: 10.1111/j.1399-302x.1991.tb00486.x. URL http://dx.doi.org/10.1111/j.1399-302x.1991.tb00486.x.

Mara Roxana Rubinstein, Xiaowei Wang, Wendy Liu, Yujun Hao, Guifang Cai, and Yiping W. Han. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating e-cadherin/β-catenin signaling via its fada adhesin. *Cell Host amp; Microbe*, 14(2):195–206, August 2013. ISSN 1931-3128. doi: 10.1016/j.chom.2013.07.012. URL http://dx.doi.org/10.1016/j.chom.2013.07.012.

Nitchakan Samainukul, Aung Khine Linn, Mohammad Bagher Javadi, Somsri Sakdee, Chanan Angsuthanasombat, and Gerd Katzenmeier. Importance of the cys124cys128 intermolecular disulfide bonding for oligomeric assembly and hemolytic activity of the helicobacter pylori tlya hemolysin. *Biochemical and Biophysical Research Communications*, 514(2):365–371, June 2019. ISSN 0006-291X. doi: 10.1016/j.bbrc.2019.04.096. URL http://dx.doi.org/10.1016/j.bbrc.2019.04.096.

Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, March 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu153. URL http://dx.doi.org/10.1093/bioinformatics/btu153.

Mukul Sharma and Pushpendra Singh. Role of tlya in the biology of uncultivable mycobacteria. *Combinatorial Chemistry amp; High Throughput Screening*, 25(10):1587–1594, August 2022. ISSN 1386-2073. doi: 10.2174/1386207325666220111150923. URL http://dx.doi.org/10.2174/1386207325666220111150923.

Brad T Sherman, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, 50(W1):W216–W221, March 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac194. URL http://dx.doi.org/10.1093/nar/gkac194.

Augustinas Silale and Bert van den Berg. Tonb-dependent transport across the bacterial outer membrane. *Annual Review of Microbiology*, 77(1):67–88, September 2023. ISSN 1545-3251. doi: 10.1146/annurev-micro-032421-111116. URL http://dx.doi.org/10.1146/annurev-micro-032421-111116.

Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1113. URL http://dx.doi.org/10.1093/nar/gkae1113.

Alaa Tafech and Angélique Stéphanou. On the importance of acidity in cancer cells and therapy. *Biology*, 13(4):225, March 2024. ISSN 2079-7737. doi: 10.3390/biology13040225. URL http://dx.doi.org/10.3390/biology13040225.

Sergi Vallvé. Gc_skew-plot: Plot a gc skew to look for the origin of replication of a bacterial genome. https://github.com/sgvallve/GC_skew-plot, 2023. Accessed: 2025-08-09.

Ni Wang, Lu Zhang, Xiao-Xu Leng, Yi-Le Xie, Zi-Ran Kang, Li-Cong Zhao, Lin-Hong Song, Cheng-Bei Zhou, and Jing-Yuan Fang. Fusobacterium nucleatum induces chemoresistance in colorectal cancer by inhibiting pyroptosis via the hippo pathway. *Gut Microbes*, 16(1), March 2024. ISSN 1949-0984. doi: 10.1080/19490976.2024.2333790. URL http://dx.doi.org/10.1080/19490976.2024.2333790.

Hanna F Willenbockel, Birte Dowerg, and Thekla Cordes. Multifaceted metabolic role of infections in the tumor microenvironment. *Current Opinion in Biotechnology*, 89:103183, October 2024. ISSN 0958-1669. doi: 10.1016/j.copbio.2024.103183. URL http://dx.doi.org/10.1016/j.copbio.2024.103183.

Yixian Wu, Songhe Guo, Fangfang Chen, Yiqiu Li, Yuying Huang, Wanli Liu, and Ge Zhang. Fn-dps, a novel virulence factor of fusobacterium nucleatum, disrupts erythrocytes and promotes metastasis in colorectal cancer. *PLOS Pathogens*, 19(1):e1011096, January 2023. ISSN 1553-7374. doi: 10.1371/journal.ppat.1011096. URL http://dx.doi.org/10.1371/journal.ppat.1011096.

Wei Xuan, Dandan Song, Jianghua Hou, and Xiuping Meng. Regulation of hippo-yap1/taz pathway in metabolic dysfunction-associated steatotic liver disease. *Frontiers in Pharmacology*, 16, January 2025. ISSN 1663-9812. doi: 10.3389/fphar.2025.1505117. URL http://dx.doi.org/10.3389/fphar.2025.1505117.

Xuqiang Zhang, Yuxin Wang, Ruochen Fan, Liying Zhang, Zhuting Li, Yanmei Zhang, Wei Zheng, Lulu Wang, Baoquan Liu, and Chunshan Quan. Quantitative proteomic analysis of outer membrane vesicles from fusobacterium nucleatum cultivated in the mimic cancer environment. *Microbiology Spectrum*, 11(4), August 2023. ISSN 2165-0497. doi: 10.1128/spectrum.00394-23. URL http://dx.doi.org/10.1128/spectrum.00394-23.

Zhongwei Zhao. Iron and oxidizing species in oxidative stress and alzheimer's disease. *AGING MEDICINE*, 2(2):82–87, June 2019. ISSN 2475-0360. doi: 10.1002/agm2.12074. URL http://dx.doi.org/10.1002/agm2.12074.

Peter S. Zilm and Anthony H. Rogers. Co-adhesion and biofilm formation by fusobacterium nucleatum in response to growth ph. *Anaerobe*, 13(3–4):146–152, June 2007. ISSN 1075-9964. doi: 10.1016/j.anaerobe.2007.04.005. URL http://dx.doi.org/10.1016/j.anaerobe.2007.04.005.