University of Amsterdam

Masters Thesis

Agent-Based Modeling of Microbial and Metabolite Interactions in Early Oral Biofilms

Author:
Xiaoqing Han

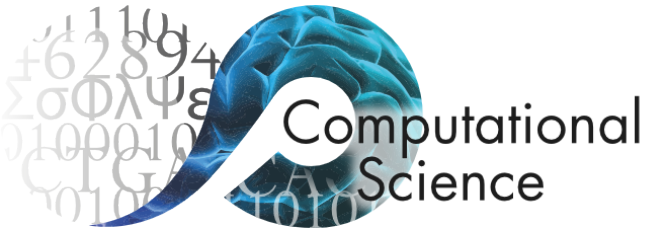
Examiner:
Vivek Sheraton Muniraj
Supervisor:
Shivam Kumar
Assessor:
Susanne Pinto

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Computational Science

in the

Computational Science Lab
Informatics Institute

August 2025

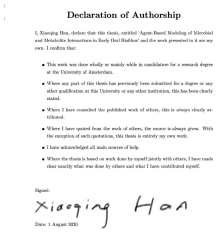

Computational
Science

Declaration of Authorship

I, Xiaoqing Han, declare that this thesis, entitled 'Agent-Based Modeling of Microbial and Metabolite Interactions in Early Oral Biofilms' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University of Amsterdam.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:



Date: 27 August 2020

"What I cannot create, I do not understand."

Richard P. Feynman

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science
Informatics Institute

Master of Science in Computational Science

Agent-Based Modeling of Microbial and Metabolite Interactions in Early Oral Biofilms

by Xiaoqing Han

This study investigated the dynamics of oral microbiota and metabolites associated with early childhood caries. Raw 16S rRNA sequencing data from CF and SECC children were processed using a custom k-mer-based pipeline to generate genus-level taxonomic profiles. The analysis revealed significant enrichment of *Streptococcus* and *Veillonella* in SECC samples, consistent with their roles in acid production, biofilm formation, and lactic acid metabolism, while other genera were reduced, supporting the ecological plaque hypothesis. Based on these profiles, a three-dimensional agent-based model was constructed using reaction-diffusion in CompuCell3D and FiPy to simulate interactions between metabolites and two bacterial species, *S. mutans* and *V. parvula*, capturing biofilm formation, agent migration, and metabolite-bacteria interactions. Our initial simulations find the mean EPS production is high in the co-culture compared to the monoculture of the S. mutants. Limitations include genus-level resolution, modeling of only two taxa, and omission of environmental factors such as spatial gradients, saliva composition, and pH dynamics. Future work will focus on pipeline optimization, model expansion to additional taxa and environmental conditions, and clinical validation to enhance predictive capacity for caries progression and preventive strategies.

Acknowledgements

First, I would like to sincerely thank Vivek. From the very beginning of my thesis, he guided me patiently and professionally, helping me clarify my research direction and achieve meaningful results. Beyond academic guidance, Vivek also supported me through monthly meetings and friendly lab gatherings, which provided opportunities to discuss ideas, receive feedback from team members, and make new friends. He also offered valuable advice on my PhD applications. These experiences greatly enhanced my research skills and made my master's journey a truly enjoyable experience.

I would also like to give special thanks to Shivam. Since I joined the team, he has provided detailed and patient guidance in every aspect of my work. From sharing and discussing literature, to answering questions about research skills, guiding my research direction, providing feedback on results, and substantial help with the modeling and simulation parts of my thesis, his support has been comprehensive. Our almost weekly meetings helped me stay on track and steadily improve my abilities. My thesis could not have been completed without his dedication and continuous support.

I would like to thank Susanne for her careful guidance. Whenever I had questions or sought help during my research, she provided clear and detailed answers, along with encouragement and positive feedback. In particular, her thoughtful feedback helped me identify issues in my work and rewrite my code to resolve a problem that had stalled my progress for a long time. Her rigorous approach and support greatly facilitated my research development.

In addition, I would like to especially thank Roland. Although he was not my thesis advisor, he provided invaluable support throughout my studies and research. In particular, he trained me in presentation skills, helping me make significant progress in effectively communicating my research. He also offered substantial help with the CompuCell3D (CC3D) part of my thesis. His enthusiasm, professionalism, and patience greatly benefited both my academic development and personal growth.

Finally, I would like to express my heartfelt gratitude to my family and friends. Throughout the long and challenging journey of my thesis, they have provided me with understanding, encouragement, and support. I am especially grateful to my mother, who has always regarded me as her pride; her love and trust have been my greatest motivation to persevere.

I would also like to thank my idol Xin Liu. In the midst of a busy and sometimes monotonous life, her performances have been a source of joy for me. She shines in the

world of art and consistently conveys positivity, serving as a role model and source of motivation for me.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

ECC	**E**arly **C**hildhood **C**aries
SECC	**S**evere **E**arly **C**hildhood **C**aries
CF	**C**aries **F**ree
EPS	**E**xtracellular **P**olymeric **S**ubstances
SCFAs	**S**short-**c**hain **F**atty **A**cids
NGS	**N**ext-**G**eneration **S**equencing
WGS	**W**hole **G**enome **S**equencing
PCR	**P**olymerase **C**hain **R**eaction
rRNA	**R**ibosomal **R**NA
OTUs	**O**perational **T**axonomic **U**nits
ASVs	**A**mplicon **S**equence **V**ariants
PDE	**P**artial **D**ifferential **E**quation
ABM	**A**gent-**B**ased **M**odel
CC3D	**C**ompu**C**ell**3**D
SRA	**S**equence **R**ead **A**rchive
EBI	**E**uropean **B**ioinformatics **I**nstitute
ENA	**E**uropean **N**ucleotide **A**rchive
HOMD	**H**uman **O**ral **M**icrobiome **D**atabase
SM	***S. m**utans*
VP	***V. p**arvula*
Su	***Su**crose*
La	***La**ctate*
Ac	***Ac**etate*
Pr	***Pr**opionate*
BH	***B**enjamini-**H**ochberg*

FDR ***F**alse **D**iscovery **R**ate*

Chapter 1

Introduction

Dental caries is one of the most prevalent chronic diseases, affecting over 530 million children worldwide [1, 2]. Early childhood caries (ECC) is defined as the presence of one or more decayed, missing (due to caries), or filled primary tooth surfaces in children under six years of age [3]. Severe ECC (SECC) is characterized by smooth-surface caries in children under three years of age, or by four to six or more affected surfaces depending on age [3]. ECC can cause persistent pain, chewing and sleep problems, impairing mood, learning, and daily life. Inadequate nutrition resulting from ECC can also hinder growth and long-term physical and mental health [3].

The development of dental caries is influenced by several major factors: the host (saliva and tooth structure), the oral microbiome (dental plaque), the immune system, the substrate (diet), and time [4, 5]. Given the multifactorial etiology of dental caries, in-depth analysis of the dynamic changes in oral microbial communities in early childhood, especially the interaction mechanism between key bacteria and metabolites, is of great significance for formulating precise prevention and intervention strategies.

The establishment of the oral microbiome begins at birth and undergoes dynamic changes throughout early childhood, forming a complex microbial ecosystem [6]. This is a selective rather than purely random process, shaped by host genetic factors, mode of delivery, and nutritional sources [7]. Vaginal delivery promotes colonization by maternal vaginal microbiota dominated by *Lactobacillus* and *Prevotella* species, while caesarean delivery favors skin-associated bacteria such as *Staphylococcus* and *Corynebacterium* [8]. Breastfeeding further shapes the early oral microbiome by introducing beneficial bacteria and providing human milk oligosaccharides that promote the growth of health-associated microorganisms [7]. As children age, the oral microbial community gradually matures and stabilizes, but remains more susceptible to environmental perturbations compared to adult microbiomes [9].

The oral environment encompasses diverse ecological niches, including saliva, hard tissues, and soft tissues, which provide a wide range of aerobic and anaerobic habitats for microorganisms [10]. More than 700 bacterial species have been identified in the oral cavity [11]. Under healthy conditions, the oral microbiota maintains a dynamic balance with the host, forming a mutually beneficial symbiotic relationship [12]. However, this balance is relatively fragile and can be disrupted by external and systemic factors such as dietary habits, psychological stress, tobacco use, and systemic diseases [13], leading to microbial dysbiosis and thereby creating an environment conducive to the development of dental caries and other oral diseases [12].

On the other hand, host defense mechanisms further shape microbial colonization. Saliva acts not only as a key immune barrier but also regulates microbial adhesion and colonization [14]. It contains antimicrobial components such as hydrogen peroxide, lactoferrin, and lysozyme, which effectively protect against invading microbes, while antimicrobial peptides exert broad-spectrum bactericidal effects by disrupting microbial membranes [14]. Nevertheless, certain bacteria, such as *Streptococcus mutans*, have evolved adaptive strategies to overcome these host defenses and achieve persistent colonization [15]. One important mechanism involves the utilization of the acquired enamel pellicle, which is enriched in proline-rich proteins that serve as adhesion receptors and help bacteria resist salivary clearance [14].

In the diverse oral microbiota, members of the *Streptococcus* and *Veillonella* genera are residents of the healthy oral community and are also involved in caries development [16]. Among them, *Streptococcus mutans* is a major cariogenic bacterium, efficiently metabolizing dietary carbohydrates to produce lactic acid and other short-chain acids, lowering local pH and promoting enamel demineralization [15]. *S. mutans* also secretes extracellular polysaccharides (EPS) to form a biofilm matrix, providing structural support for the microbial community and enhancing resistance to environmental stress and antimicrobial agents, thereby facilitating plaque formation and maturation [16, 17].

Veillonella parvula, an early colonizer of dental plaque, cannot directly ferment sugars but utilizes lactic acid produced by streptococci as its primary carbon source, converting it into weak acids such as acetic and propionic acid, which may help mitigate acid accumulation and protect enamel [16]. At the genus level, *Veillonella* species can also produce nitrite from nitrate, contributing to oral and general health [18]. Notably, recent evidence shows that *V. parvula* engages in synergistic interactions with *S. mutans*, manifested as enhanced biofilm formation and metabolic modulation, thereby potentially contributing to the pathogenesis of dental caries [16]. This synergistic interaction is illustrated in Figure 1.1.

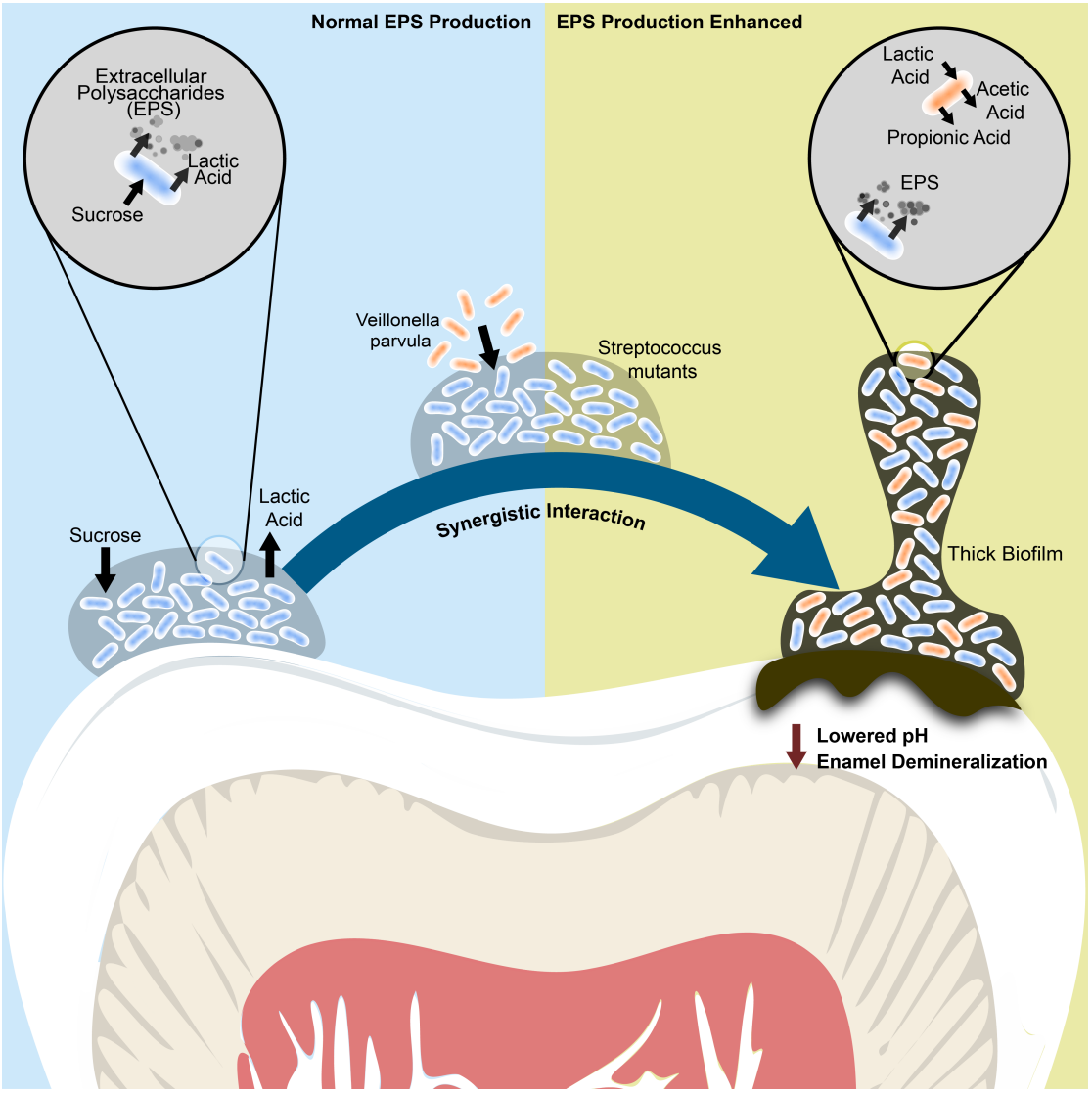


FIGURE 1.1: Schematic of the synergistic interaction between *S. mutans* and *V. parvula*. *S. mutans* metabolizes sucrose to produce lactic acid and EPS, while *V. parvula* converts lactic acid into acetic and propionic acids. Co-aggregation of *V. parvula* enhances EPS secretion, leading to a thicker and more mature biofilm.

Although the mechanistic roles of *S. mutans* and *Veillonella* in caries development are partly understood, current research lacks comprehensive computational frameworks that integrate spatiotemporal dynamics and metabolic networks to model their complex interactions [19–21].

To address these gaps, this study aims to systematically elucidate the interaction mechanisms of *S. mutans* and *V. parvula* in pediatric caries development through the integration of 16S rRNA sequencing data with computational modeling. Due to the taxonomic resolution limitations of 16S rRNA sequencing, *Streptococcus* and *Veillonella* genera are used as representative proxies for the target species in the data analysis [22].

The research strategy involves two key components: first, identification of these high-abundance genera through 16S rRNA gene sequencing data and bioinformatics analysis; second, a mechanistic Agent-based Model was developed to dynamically simulate the spatiotemporal distribution of these two critical species, EPS production leading to biofilm formation, and their metabolic exchange networks, providing insights into the biological processes associated with caries development. This computational modeling framework allows exploration and illustration of the dynamic interactions between these bacteria and their metabolites, and can inform the design of subsequent *in vitro* experiments.

Chapter 2

Literature review

2.1 Oral Biofilm Formation

Among dysbiosis-associated pathologies, dental caries exemplifies how ecological imbalances manifest at the structural level of biofilms [23]. A biofilm is a complex structure formed by a microbial community and its secreted mucilaginous polymers [17]. These microorganisms transition from a planktonic (free-floating) state to an aggregated form that adheres to and colonizes both living and non-living solid surfaces [24, 25]. The mucilaginous polymers secreted by microorganisms, known as EPS, accumulate to form a matrix that envelops the microbial community [17]. Composed mainly of nucleic acids, proteins, lipids, and polysaccharides, the EPS matrix provides structural support while shielding the community from environmental stresses, antibiotics, and toxic agents [17, 25]. The EPS matrix, by altering diffusion in combination with microbial metabolic activities, gives rise to gradients of oxygen, pH, nutrients, and metabolites within biofilms, thereby generating spatial and chemical heterogeneity [26]. EPS is essential in surface adhesion, cell recognition, biofilm formation and structure [27]. Dental plaque, forming on natural teeth and dental prostheses, represents a classic example of an oral biofilm [17]. It is influenced by multiple oral environmental and host factors, such as host immunity, pH, enzymes, saliva, and antibiotics [11].

The formation of oral biofilm follows a cyclical process, which can be summarized into several stages, as illustrated in Figure 2.1.

1. **Acquired pellicle formation.** Salivary glycoproteins such as α-amylase and proline-rich proteins spontaneously adsorb to the clean enamel surface via physicochemical interactions (long-range van der Waals forces, medium-range hydrophobic interactions, and short-range hydrogen bonding), forming the acquired pellicle

[28].

2. **Reversible attachment.** Planktonic bacteria establish reversible adhesion through nonspecific physicochemical interactions with the pellicle, often mediated by cell surface structures such as pili and fimbriae [28]. These transient contacts allow early colonizers, such as *Streptococcus* and *Actinomyces*, to remain temporarily attached [29].

3. **Irreversible attachment.** As adhesion stabilizes, some bacteria recognize specific binding proteins in the pellicle (e.g., α-amylase, proline-rich proteins) through cell-surface adhesins, entering the stage of irreversible attachment [27, 28]. At this point, pioneer bacteria begin secreting EPS, which anchor them more firmly to the surface [28].

4. **Initial colonization and biofilm formation.** Pioneer colonizers also provide new binding sites, either directly through their own surface molecules or indirectly via adsorbed salivary proteins, thereby facilitating the adhesion of secondary colonizers [28]. Later bacterial species such as *Veillonella* and *Fusobacterium* recognize polysaccharide or protein receptors on the early colonizers, leading to coaggregation and the development of characteristic spatial arrangements such as "corncob" or "bristle-brush" structures, while activate quorum sensing signaling pathways [28, 29].

5. **Biofilm maturation.** Mature oral biofilms exhibit a highly organized three-dimensional architecture, characterized by multiple layers, porous structures, and water channels that enable nutrient exchange [28].

6. **Cell dispersion.** Partial biofilm dispersion occurs, either actively or passively, releasing microbial cells to colonize new surfaces and perpetuate the biofilm life cycle [29].

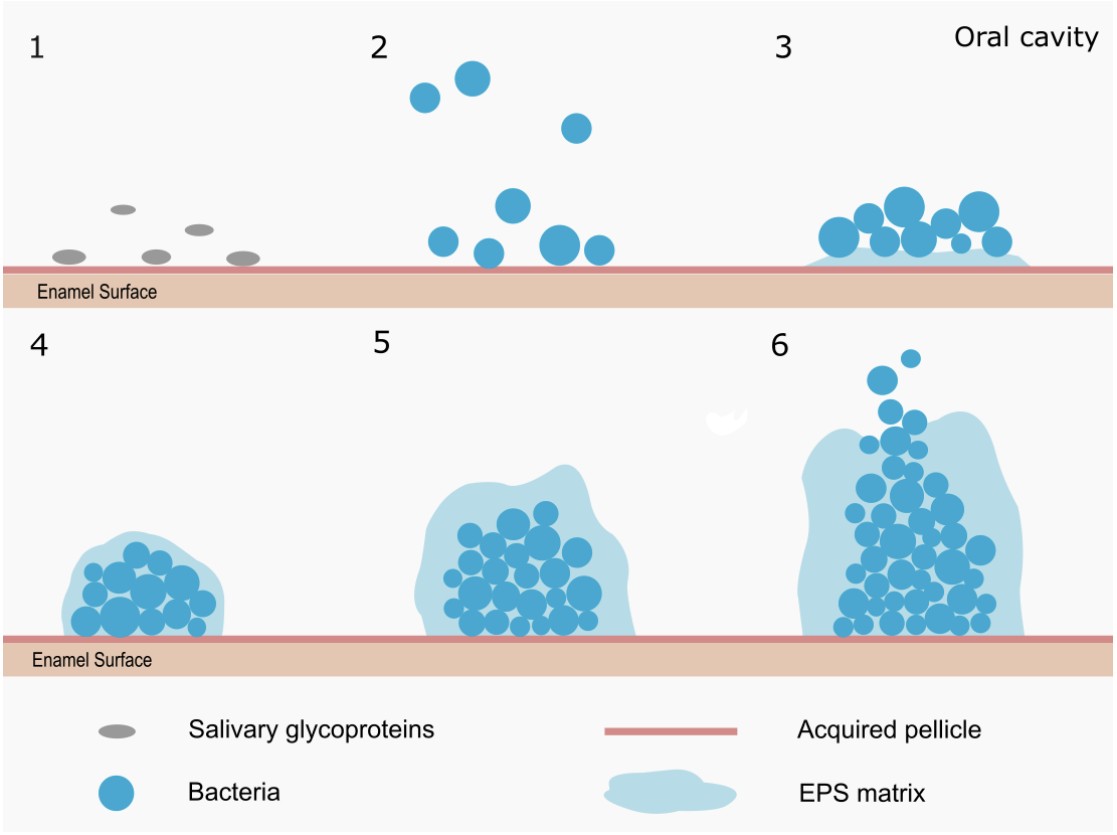


FIGURE 2.1: A simplified schematic illustration of oral biofilm formation. The process involves six stages: (1) acquired pellicle formation on the enamel surface, (2) reversible bacterial attachment, (3) irreversible attachment with EPS secretion, (4) initial colonization, (5) biofilm maturation, and (6) partial dispersion of cells for recolonization.

From Wake et al. (2016) studies, they found that the population of bacteria within biofilm and the thickness of biofilms increase in two steps based on an *in situ* model of dental biofilms. Under aerobic conditions, the number of viable bacteria increased rapidly within the first 12 hours, then slowed down, and increased significantly again between 48 to 72 hours before stabilizing [11]. Biofilm thickness began to increase after 24 hours and expanded at 48 hours, with dead cells visible in the lower layer and viable cells in the upper layer, and then reached its maximum at 72 hours before decreasing slightly [11]. They also analyzed the composition of the biofilm at different time points. Gram-positive cocci were detected within the first 8 hours, after which the biofilm was covered by a thick matrix-like structure. Within 16 hours, *Streptococci* accounted for over 20% of the total bacterial count. After 48 hours, obligate anaerobes such as *Fusobacterium*, *Prevotella*, and *Porphyromonas* became the dominant bacteria [11]. During oral biofilm formation, the initial facultative anaerobic community was gradually replaced by a Gram-negative anaerobic community.

2.2 Key Bacteria in Early Childhood Caries

In the absence of fermentable carbohydrates, mutualistic bacteria usually have an advantage; however, when sugars are frequently supplied, EPS synthesis and the formation of an acidic microenvironment disrupts this eubiosis, giving acid-producing bacteria and other acid-resistant bacteria a competitive advantage, thereby promoting the establishment of caries-related microbiota [26].

Clinical studies consistently demonstrate that *S. mutans* is the primary etiological bacterium in dental caries because of its three core cariogenic properties: (i) synthesis of large amounts of EPS from carbohydrates, (ii) metabolism of carbohydrates to organic acids, and (iii) the ability to thrive under environmental stress, especially at low pH [15]. It is a Gram-positive bacterium that relies exclusively on glycolysis for energy production [15]. Among dietary sugars, sucrose is considered the most cariogenic [30]. These processes establish a niche favorable to the growth of other acidogenic and aciduric microorganisms [15].

Within the first 24 hours of colonization, *Streptococcus* species comprise 60-90% of the supragingival plaque biomass [31]. *Streptococcus sobrinus* represents another significant risk factor for dental caries in children, exhibiting synergistic relationships with *S. mutans* that the exacerbate cariogenic potential [32]. *Bifidobacterium spp.* and *Lactobacillus spp.* are frequently associated with carious lesion progression [33].

Recent studies have also identified *Scardovia wiggsiae* as significantly associated with early childhood caries, with an abundance that is much greater in SECC than in caries-free children [34]. Furthermore, several other species including *Streptococcus salivarius*, *Streptococcus parasanguinis*, *Slackia exigua*, *Parascardovia denticolens*, and species of *Porphyromonas* have all been reported as associated with dental caries [33].

The genus *Veillonella* is a group of Gram-negative bacteria comprising 16 species, eight of which are commonly found in the human oral cavity [35]. These species account for approximately 5% of the initial plaque biomass [31]. They co-aggregate with *Streptococci* and *Actinomycetes* and exhibit metabolic properties typical of strict anaerobes, utilizing lactate produced by saccharolytic bacteria as a primary carbon and energy source to generate short-chain fatty acids (SCFAs) such as acetate and propionate [35]. This metabolic specialization explains the colocalization of *Veillonella* and *Streptococci* in dental plaque [35]. In addition to lactate, *Veillonella* species can metabolize malate, pyruvate, fumarate, and oxaloacetate [35].

Among these species, *Veillonella parvula* is highly abundant in plaque and cavities of ECC and exhibits significant functional activity; its abundance and metabolic activity

correlate with caries severity and specific gene expression changes [36–39]. These findings consistently indicate that *V. parvula* may play an important role in the development and progression of ECC.

2.3 Microbial Community Analysis

Traditional bacterial identification involves enrichment on agar-based media, followed by isolation, cultivation, and biochemical characterization [40]. However, these methods have several limitations: they are time-consuming, biased toward readily culturable taxa, and incapable of capturing full bacterial community diversity [40]. The advent of next-generation sequencing (NGS) has addressed many of these issues by enabling direct sequencing of microbial DNA or RNA from environmental samples [40, 41].

Two primary NGS-based methodologies are commonly used for microbial identification: whole genome sequencing (WGS) and amplicon sequencing [41]. WGS, also known as shotgun sequencing, captures all genetic material in a sample, while amplicon sequencing amplifies specific DNA or RNA regions via polymerase chain reaction (PCR) prior to sequencing [41]. For bacteria, 16S rRNA gene sequencing is most commonly used, targeting conserved and hypervariable regions (V1-V9) within the 16S rRNA gene [41]. Conserved regions enable universal primer binding, while hypervariable regions allow taxonomic discrimination [41].

Both methods have distinct advantages and disadvantages. WGS achieves higher species-level resolution and improved diversity detection, enabling functional gene prediction, but is considerably more expensive [42]. In contrast, 16S rRNA sequencing offers rapid and cost-effective bacterial community composition overview, but has limited species-level identification accuracy [42].

The processing and analysis of raw 16S rRNA sequencing data involves several primary steps, including adapter and primer removal, quality- and length-based trimming of reads, merging of paired-end reads when applicable, and taxonomic assignment [43]. Taxonomic classification algorithms have relied on three approaches: operational taxonomic units (OTUs), amplicon sequence variants (ASVs), and k-mer based methods.

OTUs represent a clustering method that groups sequences at a similarity threshold; for instance, sequences with 97% similarity are clustered together as an OTU and assigned a taxonomic classification [44]. ASVs, in contrast, employ denoising as their core logic—identifying sequencing errors and correcting them to generate actual biological sequences [44].

K-mer-based methods offer a fundamentally different approach to taxonomic classification. A k-mer is a contiguous genetic sequence segment of fixed length k [45]. K-mer-based tools pre-index reference genomes into k-mers, decompose sample reads into k-mers, perform rapid exact matches, and classify sequences based on a voting scheme that aggregates across k-mer hits [46]. These methods do not require sequence alignment, providing faster classification with relatively high precision and recall [47].

The selection of optimal k-value is critical for balancing precision and computational efficiency. Longer k-mers improve taxonomic specificity but increase computational cost and reduce the number of shared k-mers among samples, whereas shorter k-mers offer broader coverage of shared reads but decrease assembly quality and classification accuracy [45]. K-mer-based techniques are widely employed in bioinformatics for diverse applications, including frequency analysis, sequence indexing, and taxonomic classification [45]. Established tools such as CLARK, Kraken 2, and Mash [45] demonstrate the robustness and utility of this approach in both industry and research settings.

2.4 Computational Modeling Approaches

Current oral microbiology research faces numerous limitations that hinder a comprehensive understanding of the pathogenesis of dental caries. Most studies still focus on static characterizations of single- or dual-species systems. For example, Izumi Mashima and Futoshi Nakazawa examined biofilms formed by co-culturing *Streptococcus gordonii, Streptococcus mutans, Streptococcus salivarius*, or *Streptococcus sanguinis* with various *Veillonella* species, and demonstrated that the presence of *Veillonella* could either enhance or inhibit biofilm formation depending on the specific species combination [31]. With the application of mathematical modeling, existing approaches often rely on simplified partial differential equation (PDE) formulations. Feng et al. developed a multidimensional PDE model for a two-species *Streptococcus* and *Veillonella* biofilm that captured both cooperative and competitive interactions [20]. This model simulated biofilm growth and spatial distribution but depended on homogenization and simplifying assumptions [20].

Although several studies have explored ABM for caries dynamics, most applications have remained proof-of-concept. Head et al. applied an ABM framework to simulate two competing bacterial populations in dental biofilms differing in nutrient uptake and acid tolerance [21]. The model revealed a tipping point between benign and pathogenic states under cyclic acid challenges, and parameter sensitivity analyses suggested non-lethal interventions that could modulate biofilm composition and inform experimental design [21]. In parallel, machine learning approaches have also been applied to dental

caries research, but these studies have primarily focused on predictive classification of disease presence or risk, rather than elucidating underlying mechanistic processes [9]. In summary, although previous studies have clarified the roles of *S. mutans* and *V. parvula* in caries, mechanistic models that integrate spatiotemporal dynamics with metabolic networks remain scarce.

2.5 Mathematical Framework for Microbial Modeling

To describe microbial growth and interactions in a quantitative manner, several mathematical functions are introduced that capture key aspects of cell proliferation and substrate utilization. These functions form the core of the modeling framework and directly define the dynamic behavior of microbial populations in this study.

Monod Kinetics: Monod kinetics is a mathematical model describing microbial growth, proposed by Jacques Monod in 1942 [48]:

$$\mu = \frac{\mu_{\max} \cdot S}{K_s + S} \tag{2.1}$$

where μ is the specific growth rate, $\mu_{\max}$ is the maximum growth rate, S is the substrate concentration, and K_s is the half-velocity constant representing the substrate concentration at which the growth rate is half of $\mu_{\max}$ [48].

Hill Equation: The Hill equation, first introduced by Hill in 1910 to describe the equilibrium between oxygen and hemoglobin [49], characterizes cooperative binding kinetics:

$$\Phi(I) = \frac{I^n}{K_I^n + I^n} \tag{2.2}$$

where I is the input concentration, $\Phi(I)$ is the fractional output response, K_I is the half-saturation constant, and n is the Hill coefficient defining curve sensitivity [49]. When $n = 1$, the response reduces to a Michaelis-Menten hyperbolic curve, indicating non-cooperative binding.

The inhibitory form of the Hill equation is expressed as [49]:

$$\Phi(I) = \frac{K_I^n}{K_I^n + I^n} = \frac{1}{1 + \left(\frac{I}{K_I}\right)^n}, \tag{2.3}$$

where $\Phi(I)$ represents fractional inhibition as a function of input concentration I.

Michaelis-Menten Kinetics: Proposed by Leonor Michaelis and Maud Leonora Menten in 1913 to describe enzyme-catalyzed reactions [50]:

$$v = \frac{V_{\max} \cdot [S]}{K_m + [S]} \tag{2.4}$$

where $[S]$ is the substrate concentration, $V_{\max}$ is the maximum reaction rate, and K_m represents the substrate concentration at which the reaction rate reaches half of V_{max} [50].

Chapter 3

Methods

To investigate the metabolite dynamics of oral bacteria, our study is divided into two main stages.

In the first stage, raw genetic sequence data were obtained from publicly available databases. A k-mer-based data preprocessing pipeline was developed, following established bioinformatics principles. The k-mer approach was adopted because it enables rapid and efficient taxonomic classification with sufficient accuracy at the genus level, which is suitable for subsequent agent-based modelling. The k-mers were stored using an SQLite database.

In the second stage, a three-dimensional simulation was performed to explore the interactions between metabolites and bacterial species. The CompuCell3D (CC3D) platform, which implements agent-based modelling, was employed in combination with the external Python package FiPy to simulate spatial diffusion processes over time. To simplify the system while capturing key dynamics, two representative bacterial species were modeled based on the genus-level abundance data associated with SECC. This approach provides a dynamic and visual representation of dental plaque formation, agent migration, and interactions between oral bacteria and key metabolites. The genus-level abundance data were used as proxies for the initial species-level abundances in the simulation.

3.1 Data Sources

This study utilized NGS dataset from the Sequence Read Archive (SRA) in the European Bioinformatics Institute, European Nucleotide Archive (EBI-ENA, https://www.ebi.ac.uk/ena). The details of the dataset used are summarized in Table 3.1. All samples were collected from supragingival plaque.

In the dataset, the query sequences were provided in FASTQ format. The reference sequences were obtained in FASTA format from the Human Oral Microbiome Database (HOMD), using the latest release (version 16.02) published in April 2025.

TABLE 3.1: Characteristics of dataset

Accession	16S Region	Forward Primer	Reverse Primer	Type	Samples
PRJNA555320[51]	V4	515F	806R	Paired	40 CF, 40 SECC

CF: caries free; ECC: early childhood caries; SECC: severe ECC.
515F: GTGCCAGCMGCCGCGGTAA, 806R: GGACTACHVGGGTWTCTAAT

3.2 16S rRNA Data Preprocessing Pipeline

The developed pipeline processes 16S rRNA sequencing data through sequential steps of read cleaning, quality control, read merging, chimera removal, taxonomic annotation, and abundance calculation. The core component is a reference k-mer database that indexes k-mers with their corresponding taxonomic labels derived from reference sequences. Query sequences are decomposed into k-mers, matched against this database, and assigned to taxa based on the highest number of k-mer matches. The pipeline is designed for parallel execution to reduce processing time and computational cost. The details and default parameters of the pipeline are summarized in the following steps and illustrated in Figure 3.1. Adapter sequences had already been removed by the sequencing system on the Illumina platform (version 1.9); therefore, the pipeline does not include an additional adapter-trimming step. To ensure the absence of adapters in the raw data, two randomly selected samples were examined using FastQC Read Quality Reports (Galaxy Version 0.74+galaxy1).

The k-mer optimization strategy in this pipeline is based on previously published methods such as Kraken [52], Kraken2 [53], and CLARK [54]. Unique k-mers are stored in a SQLite database to facilitate efficient indexing and query, and duplicate k-mers are removed to reduce redundancy. Candidate k values are systematically evaluated based on taxonomic profile similarity, stability across samples, and computational runtime, rather than relying on default settings. This procedure adapts k-mer optimization to the context of 16S rRNA taxonomic classification.

1. **Primer trimming.** Universal or study-specific primers are trimmed from the raw reads. For paired-end sequences, both forward and reverse primers are removed; for single-end sequences, only the forward primer is removed. Reads without detectable primers are discarded. Primer matching allows a maximum 10 bp sliding window and up to 15% mismatches.

2. **Quality control.** Reads shorter than 100 bp, containing more than 5% ambiguous bases (N), or failing to meet the Q30 standard (99% base call accuracy) are discarded.

3. **Read merging.** For valid paired-end reads, reads with matching headers are merged into single sequences using an overlap-based algorithm, requiring a minimum 20 bp overlap and allowing up to 5% mismatches within the overlapping region. Successfully merged sequences are subsequently subjected to a second round of quality filtering. The minimum merged sequence length is set to 200 bp, and the proportion of ambiguous bases is limited to 1%. Single-end reads skip this step.

4. **K-mer database construction.** A reference database is constructed by extracting all possible k-mers of defined length from reference sequences. For different k values tested, corresponding tables are created in the SQLite database to store the mappings between k-mers and taxonomic labels. A composite primary key on the (kmer, label) columns was implemented to prevent the insertion of duplicate kmer-label pairs. All tables are indexed to optimize query efficiency.

5. **K-mer optimization.** The range of candidate k values for testing was determined with reference to the default settings of several widely used k-mer based tools, which typically employ k values between 21 and 35 (e.g., Kraken (k=31) [52], Kraken2 (k=35) [53], CLARK (k=31) [54], and Mash (k=21) [55]). In addition, k=15 was included to examine performance at smaller k value. A subset of samples is randomly selected for evaluation. Each candidate k is assessed in three steps: first, by calculating the pairwise Jaccard similarity of the resulting taxonomic profiles; second, by evaluating the stability of the k value across all selected samples; and third, by considering computational runtime. The k-value that provides the highest similarity, greatest stability, and acceptable runtime is selected for subsequent analysis.

6. **Chimera removal.** To eliminate chimeric sequences, each sequence is split into two halves at the midpoint. The k-mer hits for each half are counted, and the dominant taxon for each half is assigned based on the highest number of k-mer matches. A sequence is classified as non-chimeric if both halves share the same taxon label; otherwise, it is considered chimeric and removed.

7. **K-mer counting and taxonomic annotation.** Non-chimeric sequences are directly classified by exact k-mer matches against the k-mer database. Taxonomic assignment is determined by majority voting of k-mer matches to identify the dominant taxon.

8. **Abundance computation and downstream analysis.** Both absolute and relative abundances are calculated for each sample. Genera with zero counts in more than 5% of the samples are removed from further analysis. The top 10 most abundant genera are identified at both the sample and group levels. Microbial community compositions are visualized using bar plots, heatmaps, and box plots. Statistical comparisons of key genera (e.g., *Streptococcus* and *Veillonella*) between groups are performed using the Mann-Whitney U test or the Kruskal-Wallis test, as appropriate, with p-values for the top genera adjusted for multiple testing using the Benjamini-Hochberg (BH) false discovery rate (FDR) method.

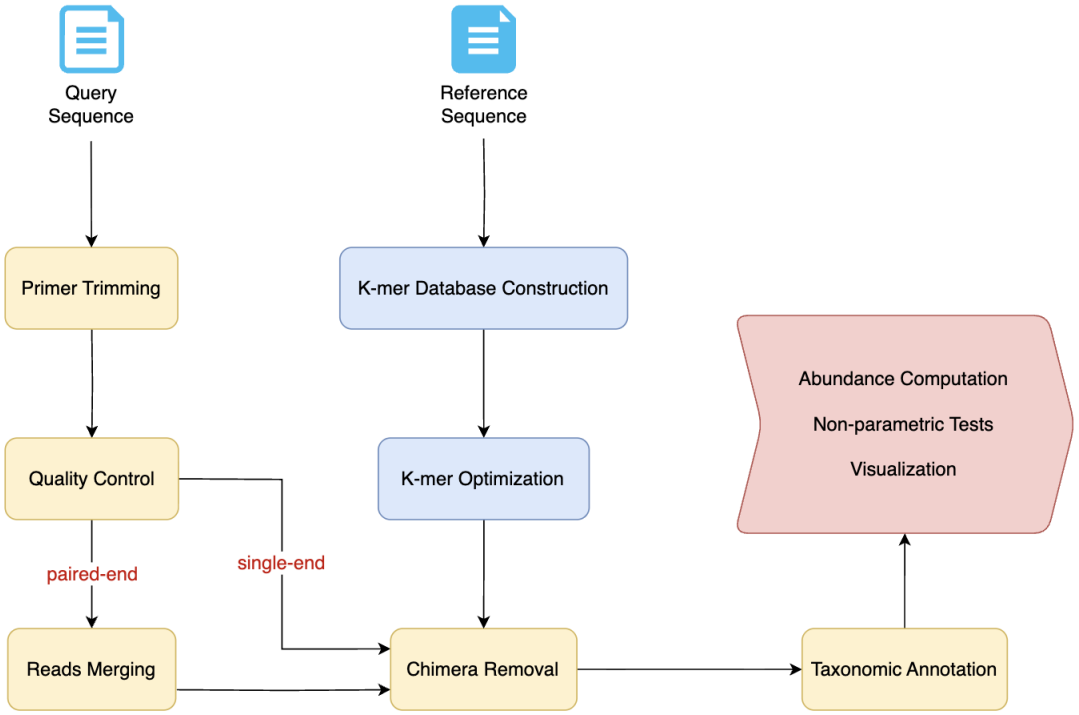


FIGURE 3.1: The 16S rRNA data preprocessing pipeline. Yellow boxes indicate steps related to query sequences, blue boxes represent steps involving reference sequences, and the red box denotes the analyses and outputs generated by the pipeline.

3.3 Simulation

Agent-Based Modeling (ABM) is a powerful computational approach for simulating the actions and interactions of individual agents, providing insights into complex real-world systems [56]. In this study, a hybrid modeling framework was employed to simulate the spatiotemporal dynamics of a dual-species oral biofilm composed of *S. mutans* and *V. parvula*. In this framework, bacteria and metabolites were modeled as agents, while

the oral cavity was represented as a three-dimensional lattice environment. Bacterial growth and movement were tracked at each time step, along with substrate production and consumption, which were visualized on the grid.

The simulation system integrates ABM with continuum modeling, consisting of three core components:

CC3D: Provides a Cellular Potts Model (CPM) engine to manage the physical representation of cells on the 3D lattice, including cell shape, volume, and adhesion-based interactions. Bacterial dynamics are modeled using time-dependent ordinary differential equations, with growth following Monod kinetics and inhibition described by a Hill function.

FiPy: A Python-based partial differential equation solver for reaction-diffusion equations, simulating metabolite concentrations (e.g., sucrose, lactate) across the environment. Substrate dynamics follow Michaelis-Menten kinetics.

Custom Stepper: Serves as a bridge between CC3D and FiPy, coordinating data extraction, transmission, and state updates between bacterial and metabolite dynamics.

This integrated framework enables detailed visualization and analysis of bacterial-metabolite interactions within the 3D oral biofilm over time. Figure 3.2 shows the illustration of the workflow.

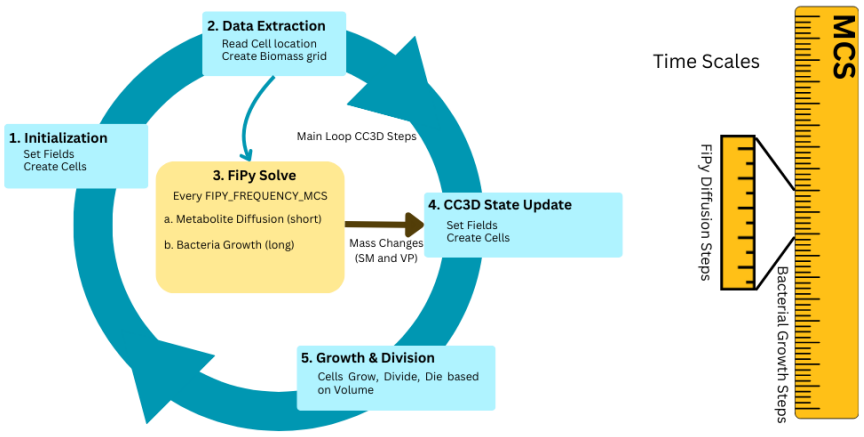


FIGURE 3.2: This diagram illustrates the typical workflow for a numerical simulation using CompuCell3D and FiPy. On the right, the timeline highlights the different time scales of the physical processes being modeled.

3.3.1 Biological mechanisms

1. **Bacterial characteristics.** *S. mutans* is a primary colonizer that consumes sucrose to produce lactate and EPS. *V. parvula* is a secondary colonizer that utilizes lactate from *S. mutans* and converts it into acetate and propionate.

2. **Cell growth, division and death.** Internal cell mass dynamics are converted to cell volume assuming a constant bacterial density. Cell division was triggered when internal mass exceeded a threshold, and cells were removed when their volume fell below a minimum. Bacteria undergo first-order decay, representing natural cell death in the simulation.

3.3.2 Physical mechanisms

1. **Simulation domain setup.** Simulations were performed on a 50×50×14 voxel three-dimensional lattice with periodic boundary conditions. Here each voxel is $1\mu m^3$ volume. Each of these voxel contains 10 bacteria. Adhesion energies between cell types were specified to capture aggregation behavior: *V. parvula* aggregates most strongly (Energy=5), *S. mutans* moderately (Energy=10), and *S. mutans* and *v. parvula* interaction is weakest (Energy=15). These energy differences influence self-clustering and spatial sorting of cells within the lattice.

2. **Cellular Potts Model.** The system evolved by minimizing the total energy (Hamiltonian), comprising two components: volume constraint energy, which enforces cells to maintain their target volume, and adhesion energy, which governs self-aggregation and interspecies interactions. Cell movement and shape changes on the lattice arise from the combined effects of these energy terms and are implemented using the CPM, where stochastic lattice-site copy attempts are accepted or rejected based on the resulting changes in total energy.

3.3.3 Chemical mechanisms

1. **Reaction-diffusion equations.** The system comprises five chemical fields: sucrose, lactate, acetate, propionate, and EPS, which are modeled using standard reaction-diffusion equations solved with FiPy. The diffusion of each chemical species is governed by its diffusion coefficient, while reaction terms are determined by the metabolic activity and biomass distribution of the cells.

2. **Time scale separation.** A multi-scale time-stepping strategy was employed. At the fastest scale, cell shape and position were updated at each Monte Carlo step (MSC) within the CPM. At the medium scale, the reaction-diffusion PDE system was solved to capture changes in metabolite concentrations. At the slowest scale, cell mass, volume, and division events were updated following the PDE solution, reflecting the slower dynamics of biological processes.

3.3.4 Initial and boundary conditions

1. **Initial conditions.** Cells were initially seeded randomly at the bottom of the simulation domain according to the experiment type (e.g., "mixed", "separated", "sm_only"). The chemical environment was initialized with a linear sucrose gradient, with the lowest concentration at the bottom and the highest at the top, while all other metabolites were set to zero initially.

2. **Boundary conditions.** Boundary conditions were applied as follows: the top surface (z = max) imposed a fixed-value (Dirichlet) condition for sucrose to simulate continuous nutrient supply from saliva; the bottom surface (z = 0) used zero-flux (Neumann) conditions for all chemical fields to represent the impermeable tooth surface; and the side boundaries were set to default zero-flux conditions.

The main simulation loop consisted of three stages. First, chemical concentrations and cell positions were extracted to construct the biomass density field. Second, a transient simulation of metabolite dynamics was performed, and cell growth was calculated based on the resulting metabolite concentrations. Finally, the chemical fields and cell states were updated, including changes in cell mass, volume, division, and death, reflecting the interplay between metabolism and cellular processes.

During the simulations, the following metrics were monitored in real time: cell population dynamics, total cell volume and mass, average metabolite concentrations, and spatial distribution patterns. All simulations were performed using a fixed random seed to ensure reproducibility.

Figure 3.3 illustrates the metabolic interactions between *S.mutans* (SM) and *V.parvula* (VP) as represented in the model. SM consumes sucrose (Su) to produce lactate (La), which VP then metabolizes to generate acetate (Ac) and propionate (Pr). VP enhances EPS production, while EPS promotes SM growth. The acids La, Ac, and Pr collectively inhibit SM via a Hill-function term. Key kinetic parameters are shown for each process, including growth rates, substrate affinities, and degradation rates. Diffusion terms account for spatial metabolite distribution. The model captures cross-feeding dynamics, EPS-mediated cooperation, and acid-dependent growth regulation.

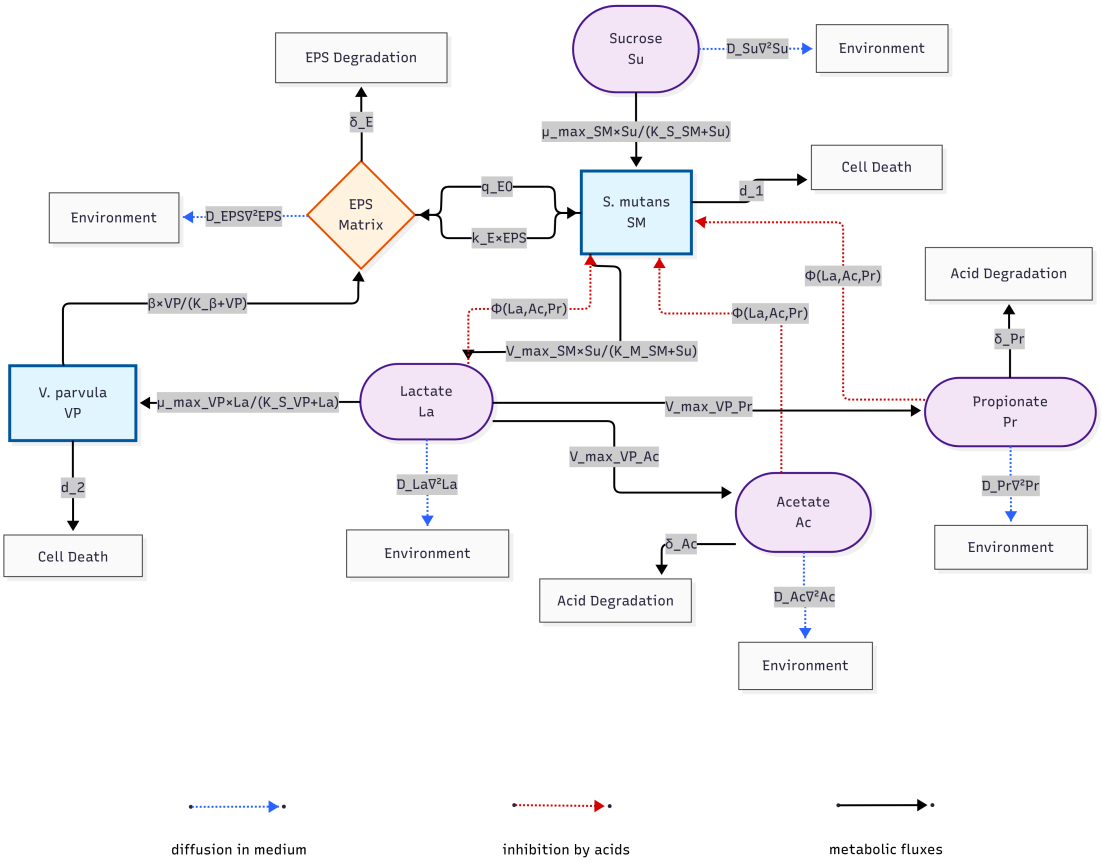


FIGURE 3.3: Computational network of the *S. mutans* and *V. parvula* metabolic system. Blue rectangles represent bacterial species, purple oval represents substrates, and orange diamond represents biofilm matrix.

3.3.5 Spatial-Temporal Dynamics and Equations

As substrates diffuse in the 3D grid, substrate concentration is represented as $S(x, y, z, t)$, changing with time and space. The general substrate dynamics are described by:

$$\frac{\partial S}{\partial t} = D_S \nabla^2 S + \sum_i P(S, A_i) - \sum_j C(S, A_j) \tag{3.1}$$

where D_S is the diffusion coefficient, the $\nabla^2 S$ is the Laplace operator. $\sum P(S, A_i)$ represents substrate production from all cells, and $\sum C(S, A_j)$ represents substrate consumption from all cells.

Incorporating Michaelis-Menten kinetics, the full equation becomes:

$$\frac{\partial S}{\partial t} = D_S \nabla^2 S + \sum_i \frac{V_{\max,i} S}{K_{m,i} + S} A_i - \sum_j \frac{V_{\max,j} S}{K_{m,j} + S} A_j \tag{3.2}$$

For bacterial population dynamics, assuming population N and death rate d:

$$\frac{dN}{dt} = \mu \cdot N - d \cdot N \tag{3.3}$$

Incorporating Monod kinetics:

$$\frac{dN}{dt} = \left(\frac{\mu_{\max} S}{K_s + S}\right) N - d \cdot N \tag{3.4}$$

Definitions and units of all parameters are summarized in Table 3.2 and Table 3.3.

The inhibitory effects of La, Ac, and Pr on bacterial growth are modeled using a modified Hill function:

$$\Phi(La, Ac, Pr) = \frac{1}{1 + \frac{La}{K_{I,La}} + \frac{Ac}{K_{I,Ac}} + \frac{Pr}{K_{I,Pr}}}$$

where $K_{I,La}, K_{I,Ac}, K_{I,Pr}$ are the half-inhibition concentrations for lactic acid, acetate, and propionate, respectively. Each term represents the relative inhibitory strength of the corresponding acid. The additive form assumes independent and cumulative inhibitory effects. The Hill coefficient n is set to 1, indicating no cooperativity between acid molecules and the bacterial acid-sensing mechanism.

The temporal changes in SM and VP populations are described by:

$$\frac{dSM}{dt} = \mu_{\max,SM} \frac{Su}{K^{Su}_{S,SM} + Su} \cdot SM \cdot \Phi(La, Ac, Pr) - d_1 SM + k_E \cdot EPS \cdot SM$$

$$\frac{dVP}{dt} = \mu_{\max,VP} \frac{La}{K^{La}_{S,VP} + La} \cdot VP - d_2 VP$$

The first term in each equation represents substrate-limited bacterial growth based on Monod kinetics. For *S.mutans*, growth is further modulated by the acid inhibition function $\Phi(La, Ac, Pr)$, natural death at rate d_1, and EPS-mediated growth enhancement. For *V.parvula*, the equation includes substrate-limited growth and natural death at rate d_2.

The spatial and temporal dynamics of metabolites are described by reaction-diffusion equations.

Sucrose:

$$\frac{\partial Su}{\partial t} = D_{Su} \nabla^2 Su - \frac{V^{Su}_{\max,SM} \cdot Su}{K^{Su}_{M,SM} + Su} \cdot SM \cdot \Phi(La, Ac, Pr)$$

Sucrose diffuses through the medium and is consumed by *S. mutans* following Michaelis-Menten kinetics, with consumption rate modulated by the acid inhibition function.

Lactate:

$$\frac{\partial La}{\partial t} = D_{La}\nabla^2 La + Y_{La/Su} \cdot \frac{V^{Su}_{\max,SM} \cdot Su}{K^{Su}_{M,SM} + Su} \cdot SM - \frac{V^{La}_{\max,VP} \cdot La}{K^{La}_{M,VP} + La} \cdot VP$$

Lactic acid diffuses, is produced by *S. mutans* from sucrose metabolism at yield $Y_{La/Su}$, and is consumed by *V. parvula*.

EPS:

$$\frac{\partial EPS}{\partial t} = D_{EPS}\nabla^2 EPS + \left(q_{E0} + \beta \frac{VP}{K_\beta + VP}\right) SM - \delta_E EPS$$

EPS diffuses slowly, is produced by *S. mutans* at basal rate $q_E 0$ and enhanced by *V. parvula* presence, and degrades at rate δ_E.

Acetate:

$$\frac{\partial Ac}{\partial t} = D_{Ac}\nabla^2 Ac + \frac{V^{Ac}_{\max,VP} \cdot La}{K^{Ac}_{M,VP} + La} \cdot VP - \delta_{Ac} Ac$$

Acetate diffuses, is produced by *V. parvula* via lactate metabolism, and degrades at rate δ_{Ac}.

Propionate:

$$\frac{\partial Pr}{\partial t} = D_{Pr}\nabla^2 Pr + \frac{V^{Pr}_{\max,VP} \cdot La}{K^{Pr}_{M,VP} + La} \cdot VP - \delta_{Pr} Pr$$

Propionate diffuses, is produced by *V. parvula* from lactate metabolism, and degrades at rate δ_{Pr}.

3.3.6 Variable Definitions and Parameters

TABLE 3.2: Definition of state variables in the model

Symbol	**Description**
SM	*S. mutans*
VP	*V. parvula*
Su	Sucrose concentration
EPS	EPS concentration
La	Lactate concentration
Ac	Acetate concentration
Pr	Propionate concentration

TABLE 3.3: Kinetic parameters and units

Parameter	**Description**	**Units**
$\mu_{\max,SM}$	Max growth rate of *S. mutans*	h^{-1}
$\mu_{\max,VP}$	Max growth rate of *V. parvula*	h^{-1}
$K_{S,SM}^{Su}$	Half-saturation constant for sucrose	mmol·L^{-1}
$K_{S,VP}^{La}$	Half-saturation constant for lactate	mmol·L^{-1}
$Y_{La/Su}$	Lactate yield from sucrose	g lactate·g^{-1} sucrose
$Y_{La/VP}$	Biomass yield of VP from lactate	g biomass·g^{-1} lactate
β	Max EPS enhancement by VP	mmol·$gDCW^{-1}$·h^{-1}
K_β	VP half-saturation for EPS enhancement	g·L^{-1}
K_I	Lactate inhibition constant for SM	mmol·L^{-1}
n	Hill coefficient (acid inhibition)	unitless
d_1, d_2	Death rates of SM and VP	h^{-1}
k_E	EPS-dependent growth enhancement	L·g^{-1}·h^{-1}
δ_E	EPS degradation rate	h^{-1}

Chapter 4

Experiments and results

In this experiment, a total of 80 raw sequencing datasets (80 samples) in FASTQ format were downloaded from HOMD, amounting to 2.2 GB of sequencing data. The samples consisted of a caries free group (CF, n = 40) and a severe early childhood caries group (SECC, n = 40).

A 16S rRNA data preprocessing pipeline was then developed, comprising primer trimming, quality control, sequence merging, k-mer database construction, k-mer value selection, chimera removal, taxonomic annotation, and the computation of sample abundances. The pipeline was implemented using custom Python scripts and provided a reliable foundation for subsequent microbial community analyses. It is flexible, supporting both single- and paired-end reads, testing of different primers, and adjustable parameters at various steps (e.g., quality thresholds, k-mer values) to optimize processing and ensure reproducible results. In this study, all parameters were applied under strict standards. The total runtime of pipeline across all samples was about 3 hours.

4.1 K-mer Optimization

To determine the optimal k-mer value for database construction and subsequent taxonomic assignment, candidate k values (15, 21, 25, 31, and 35) were evaluated across ten randomly selected samples per group (random seed = 42). For each k, the mean Jaccard similarity of the top 10 taxa, the standard deviation of k-mer performance across samples, and computational runtime were recorded. Across all samples, the number of detected taxa fluctuated slightly with varying k values but remained generally stable, with no abnormal k-mer curves observed, representative results are shown in Table 4.2. As k increased, specificity improved, confirming the stability of the selection.

The overall testing runtime was approximately 1 hour, with an average sequencing depth of 64,707 reads. The optimal k-mer value for both groups was 25, as it achieved the highest Jaccard similarity (0.946), the lowest standard deviation (0.090), and the shortest runtime (28.96 s). Detailed results for all k values are provided in Table 4.1.

TABLE 4.1: Overall Performance Metrics for Different K Values

k	Average Score	Standard Deviation	Average Time (s)
15	0.9250	0.1086	47.47
21	0.9447	0.0976	32.68
25	**0.9462**	**0.0903**	**28.96**
31	0.9273	0.1083	30.85
35	0.9205	0.1092	31.57

TABLE 4.2: Representative k-mer optimization results

Sample	**Reads**	**Taxon count at Different k**				
		15	21	25	31	35
CF4	43,728	247	236	**234**	235	232
CF6	70,176	268	258	**265**	259	257
CF..	–	–	–	–	–	–
SECC30	66,703	253	240	**239**	241	241
SECC17	57,800	255	252	**248**	247	248
SECC..	–	–	–	–	–	–

Optimal k values: k =25 for all samples.

4.2 Data Preprocessing

The raw sequences were generated on the Illumina platform (version 1.9), with adapters already removed by the sequencing system. To confirm this, two random samples were examined using FastQC Read Quality Reports (Galaxy Version 0.74+galaxy1), which verified the absence of adapters in the raw sequences.

- **Quality control of raw data.** A total of 16,071,370 paired-end reads (8,035,685 reads per strand) were obtained across all samples. After trimming primer and quality control, 12,927,489 reads (76.42%) were retained. Strand-specific filtration removed 26.83% of R1 (2,155,918 reads) and 20.33% of R2 (1,633,305 reads).
- **Read merging and chimera removal.** Header-matched read pairs were merged with 88.55% efficiency, yielding 5,206,401 merged reads from 5,879,767 valid pairs. Post-merge quality control retained 5,202,944 reads (99.93%). the length of merged sequences are range from 233 bp to 338 bp. Chimera removal produced 5,202,881 high-quality sequences, almost 100% of reads were retained in the final dataset.

- **Filtering summary.** The complete workflow eliminated 67.63% of original reads (10,868,489 filtered), with detailed metrics in Table 4.3.

TABLE 4.3: Statistics of data preprocessing

Processing Step	**R1**	**R2**	**Total**
Raw Sequencing	8,035,685	8,035,685	16,071,370
QC Filtering			
Kept reads	5,879,767	6,402,380	12,282,147
Filtered reads	2,155,918	1,633,305	3,789,223
Pass rate	73.17%	79.67%	76.42%
Read Merging			
Merged reads	–	–	5,206,401
Merge rate	–	–	88.55%
Post-Merge QC			
Kept reads	–	–	5,202,944
Filtered reads	–	–	3457
Pass rate	–	–	99.93%
Chimera Removal			
Final reads	–	–	5,202,881
Filtered reads	–	–	63
Retention rate	–	–	100%

Note: Merge rate calculated against valid pairs (5,879,767). All percentages represent proportions of previous step's output.

4.3 Abundance Analysis Results

The genus-level abundance patterns in each group are shown in Figure 4.1, highlighting the top 10 most abundant genera across the CF and SECC groups. Taxonomic profiling revealed distinct microbial composition patterns between the two study groups.

For the CF group, the five most abundant genera were *Neisseria*, *Actinomyces*, *Streptococcus*, *Corynebacterium* and *Selenomonas*, representing the core oral microbiome components in healthy children. For the SECC group, the dominant genera were *Neisseria*, *Selenomonas*, *Streptococcus*, *Veillonella* and *Actinomyces*, indicating a shift in microbial community structure associated with severe caries development.

When compared with published findings, our results showed high consistency in the core bacterial genera identified. Previous studies reported *Actinomyces*, *Neisseria*, and *Corynebacterium* as the most abundant genera in caries-free children, whereas *Veillonella*, *Neisseria*, and *Streptococcus* predominated in the S-ECC group [51]. Our analysis similarly detected these genera, with *Neisseria* consistently emerging as a dominant

genus in both groups, supporting the reproducibility of genus-level oral microbiome signatures. Observed differences in abundance rankings were noted, which may reflect population-specific variation or methodological factors. These results indicate that our pipeline reliably detects the core genera associated with health and disease states.

The relative abundances of the top genera were compared between CF and SECC groups. While *Neisseria* showed a slight decrease and *Actinomyces* and *Corynebacterium* declined more obviously in SECC, the abundances of *Streptococcus*, *Selenomonas*, and *Veillonella* increased. These trends are consistent with previous reports, which observed increased abundances of *Streptococcus*, *Veillonella*, *Prevotella*, and *Selenomonas*, and decreased abundances of *Actinomyces* and *Leptotrichia* [51].

The Mann-Whitney U test was applied to *Streptococcus* and *Veillonella*, both showing significant differences between CF and SECC ($p < 0.05$). Boxplots (Figure 4.2) were generated for these genera, highlighting their marked increase in SECC, particularly for *Veillonella*, consistent with previous reports [51]. This pipeline enabled accurate characterization of microbial communities and emphasized the important roles of *Streptococcus* and *Veillonella* in dental caries.

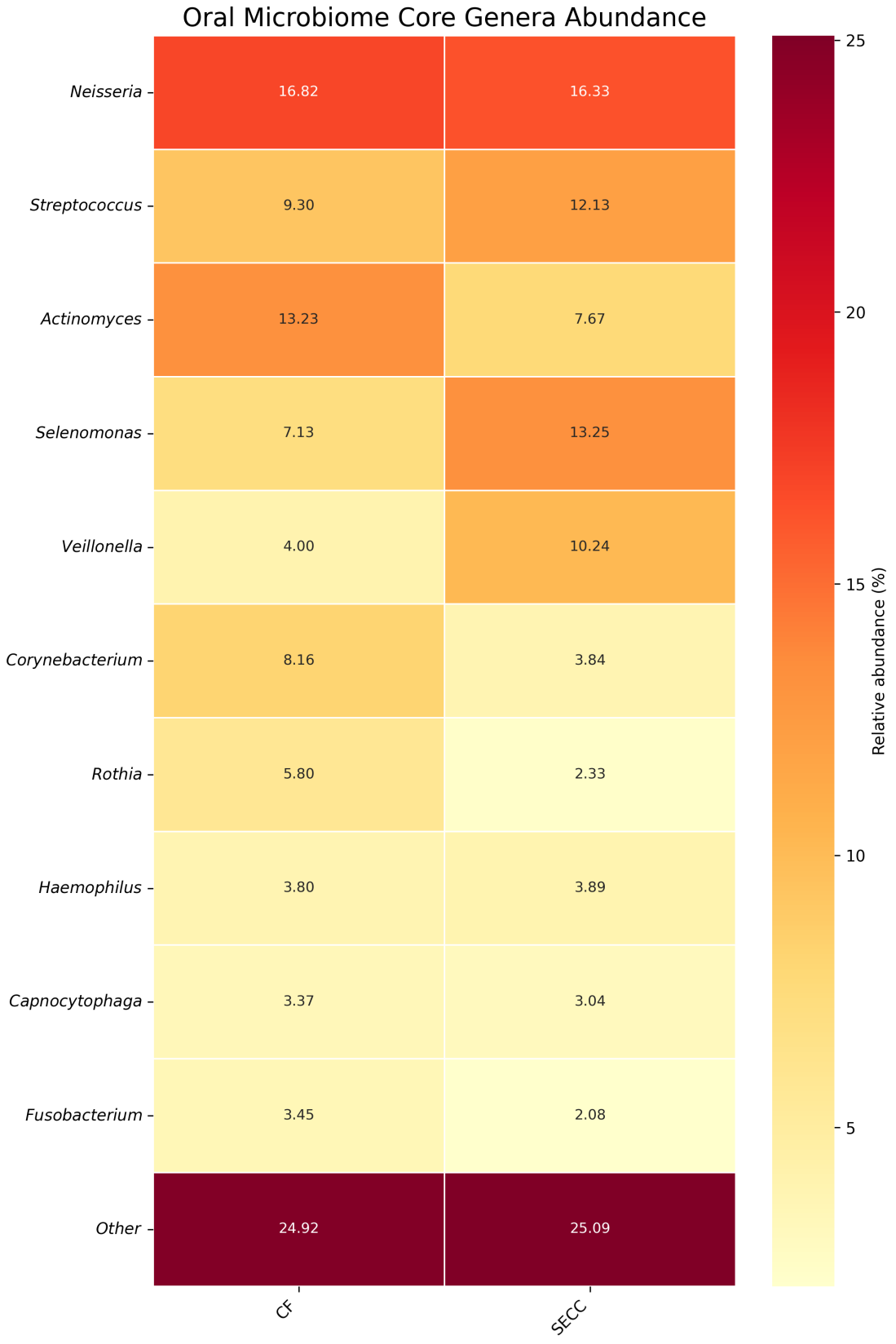


FIGURE 4.1: Heatmap of the top 10 most abundant genera across CF and SECC groups. Distinct clustering of microbial profiles is observed between groups, with higher relative abundance of *Streptococcus* and *Veillonella* in SECC.

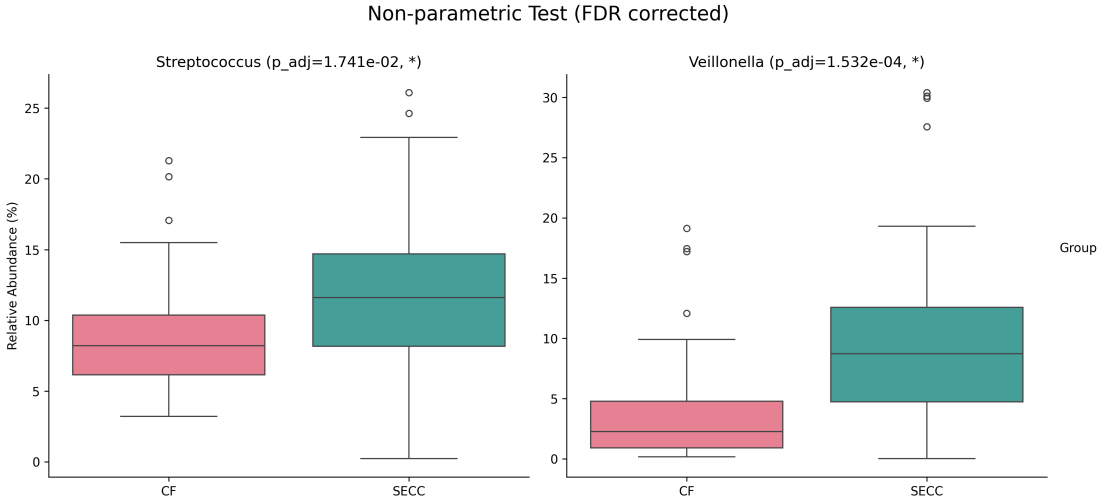
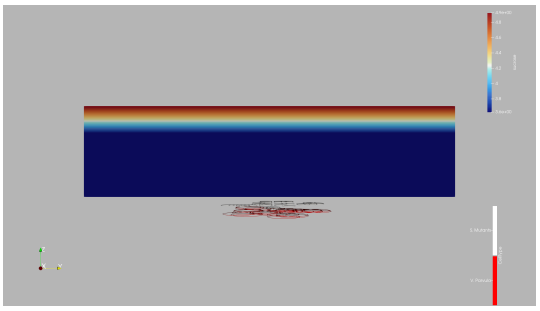


FIGURE 4.2: Boxplots comparing the relative abundances of *Streptococcus* and *Veillonella* between CF and SECC groups. Mann-Whitney U tests revealed significantly higher abundances of both genera in the SECC group, particularly *Veillonella*.

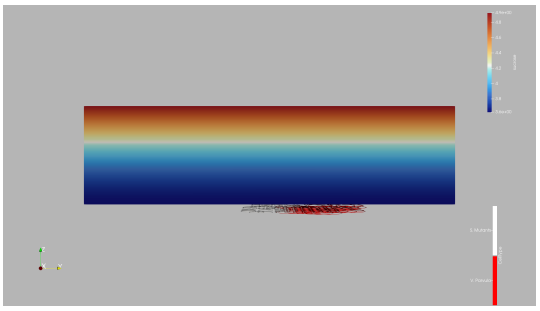
These abundance distributions provided the quantitative basis for initializing the agent-based model, allowing simulation of biofilm development under different ecological conditions.

4.4 CC3D Simulation

The simulation begins with two bacterial species, *S. mutans* (white) and *V. parvula* (red), positioned at the bottom of the domain in spatially separated colonies. The initial bacterial quantities in the model were set based on the relative abundances calculated after data processing. To validate that the simulation accurately reflects the model dynamics, it was initiated with a sucrose gradient, as depicted in Fig. 4.3a. A Dirichlet boundary condition was applied at the top boundary to maintain the sucrose concentration at a constant 5.0 μM. Figure 4.3b illustrates the state of the gradient at a subsequent time point.



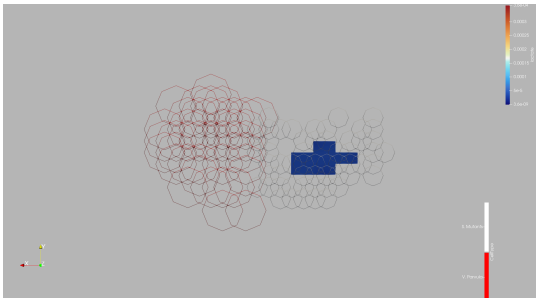
(A) Initial sucrose gradient (t=0). A constant sucrose concentration of 5.0 μM is maintained at the top boundary.



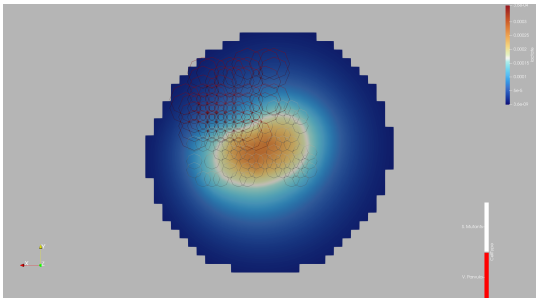
(B) Fully developed sucrose gradient at $t > 0$, showing diffusion and consumption effects.

FIGURE 4.3: Sucrose gradient simulation: (a) initial state, (b) state after diffusion and consumption.

The diffusion of sucrose initiates lactate production by *S. mutans*. Figure 4.4a illustrates the initial state from a top-down perspective, with lactate appearing around the bacterial colonies (represented in white). As shown in Figure 4.4b, the lactate concentration increases and disperses throughout the entire domain after one hour of simulation.



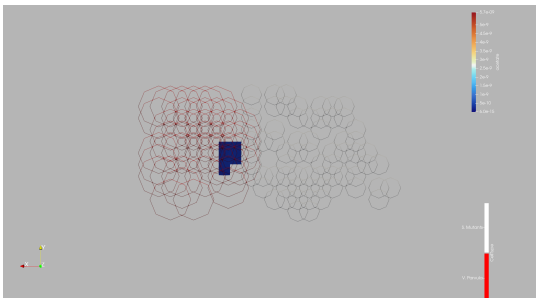
(A) Initial phase of lactate production. Lactate concentration begins to increase near *S. mutans* colonies (white).



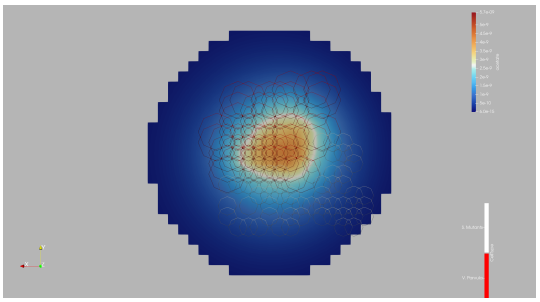
(B) Lactate gradient after one hour, showing dispersion and accumulation.

FIGURE 4.4: Lactate production dynamics by *S. mutans*: (a) initial phase, (b) after one hour of simulation.

The diffusion of lactate stimulates acetate production by *V. parvula* (shown in red). Acetate production is initiated at the interface with the *S. mutans* colonies (Figure 4.5a). Figure 4.5b shows the acetate concentration after one hour, by which time it has dispersed throughout the simulation domain.



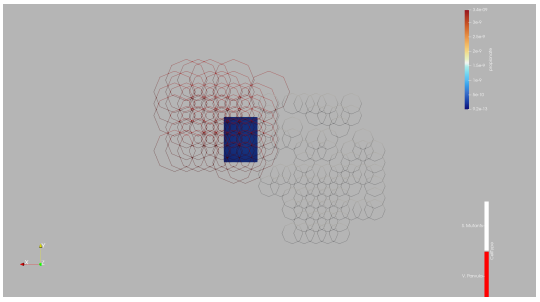
(A) Initiation of acetate production by *V. parvula* (red) upon stimulation by lactate. Production is localized to the interface between the two species.



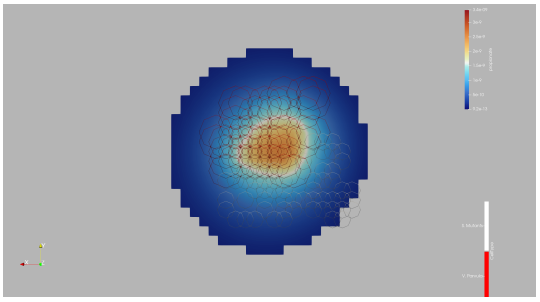
(B) Acetate concentration after one hour, showing widespread distribution due to metabolic cross-feeding.

FIGURE 4.5: Acetate production dynamics by *V. parvula*: (a) initial state, (b) after one hour of simulation.

The diffusion of lactate stimulates propionate production by *V. parvula* (shown in red). Propionate production is initiated at the interface with the *S. mutans* colonies (Figure 4.6a). Figure 4.6b shows the propionate concentration after one hour, by which time it has dispersed throughout the simulation domain.



(A) Initiation of propionate production by *V. parvula* (red) upon stimulation by lactate. Production is localized to the interface between the two species.

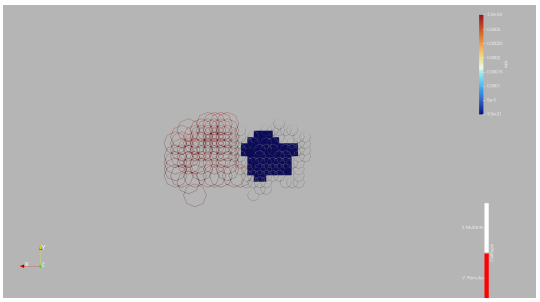


(B) Propionate concentration after one hour, showing widespread distribution due to metabolic cross-feeding.

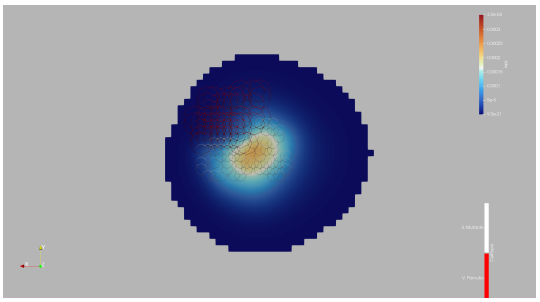
FIGURE 4.6: Propionate production dynamics by *V. parvula*: (a) initial state, (b) after one hour of simulation.

Next, a comparison was made between EPS production by *S. mutans* with and without the presence of *V. parvula*. Figure 4.7a shows the initial production of EPS by *S. mutans* (white). In Figure 4.7b, the EPS has spread and is distributed around the *S. mutans* colonies.

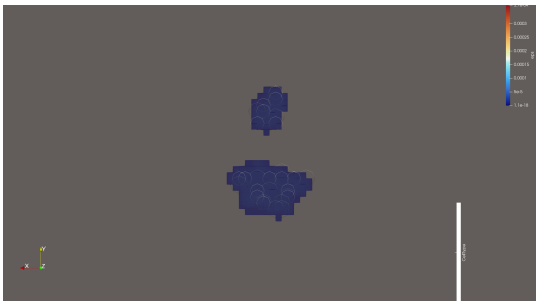
For the *S. mutans* monoculture, initial EPS production is depicted in Figure 4.7c, followed by a significant increase in EPS concentration shown in Figure 4.7d.



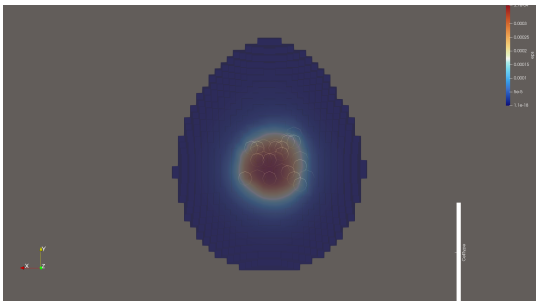
(A) EPS production by *S. mutans* in co-culture: initial state.



(B) EPS production by *S. mutans* in co-culture: after growth.



(C) EPS production by *S. mutans* monoculture: initial state.



(D) EPS production by *S. mutans* monoculture: after growth.

FIGURE 4.7: EPS production dynamics: (a,b) co-culture with *V. parvula*, (c,d) *S. mutans* monoculture.

Our initial findings suggest that the presence of V. parvula increases mean EPS production. However, because these biological processes are inherently stochastic, a single observation is not conclusive. Therefore, multiple future simulations are required to statistically confirm this effect.

Chapter 5

Discussion

5.1 Pipeline Performance

This study developed a custom pipeline for processing 16S rRNA sequencing data, designed to ensure high-quality analysis through steps including quality control, read merging, and chimera removal. A key advantage of this pipeline is flexibility: it allows full parameter customization, supports selection of different reference databases, accommodates both single-end and paired-end sequencing data, and removes variable primer sequences. In addition, the pipeline generates all necessary files and logs at each step, ensuring full traceability of the analysis.

Building on this framework, a k-mer optimization module was introduced based on result consistency. The module evaluates multiple candidate k-mer lengths by performing k-mer searches against a k-mer database and genus-level annotation, then selects the optimal k by comparing classification results, achieving Jaccard similarities of 0.9462 (k=25) for all samples. This approach contrasts with existing 16S sequence analysis tools (e.g., QIIME2 or DADA2) [57, 58] and standard k-mer classifiers (e.g., Kraken 2 or CLARK) [45], which typically rely on fixed k-mer settings. Relevant studies have shown that different k-mer lengths can significantly affect classification accuracy and resolution, and using multiple k-mer sizes performs particularly well in strain-level classification [59]. In addition, tools such as KITSUNE have attempted to systematically optimize k-mer lengths in phylogenetic analyses, indicating that k-mer selection based on informational features has a theoretical basis [60].

During the optimization process, a notable computational feature regarding runtime was observed: contrary to conventional expectations, runtime decreased with increasing k-mer length. This occurs because the pipeline queries each sample's unique k-mers

against a pre-constructed database, and longer k-mers are more specific, resulting in fewer candidate matches per query. Consequently, although the theoretical number of k-mers grows with k, the actual number of k-mers requiring verification is smaller, leading to faster computation.

Overall, the results-driven k-mer optimization implemented in our Python-based pipeline enhances the robustness and accuracy of taxonomic annotation, as validated by the consistent detection of core oral microbiome genera. Its portability and adaptability make it suitable for diverse research scenarios, providing a reliable tool for downstream microbiome studies and computational modeling. Notably, the pipeline employs SQLite to construct and manage the k-mer database, which enables efficient handling of multiple candidate k values. By organizing k-mer information in a relational database, the pipeline supports rapid retrieval, reduces redundant computation, and ensures systematic traceability of k-mer tests. This database-centric design offers a level of flexibility and scalability that is generally absent in conventional sequence analysis pipelines, which typically rely on static file-based data management.

5.2 Simulation Performance

This study developed a hybrid ABM framework integrating CC3D and FiPy to simulate the growth and metabolic interactions of a dual-species oral biofilm (*S. mutans* and *V. parvula*) in a three-dimensional environment. The simulation revealed the spatiotemporal dynamics of the bacteria–metabolite network, providing mechanistic insights into biofilm formation and microbial community interactions.

The model successfully reproduced the classical cross-feeding mechanism, with *S. mutans* consuming sucrose to produce lactate, which V. parvula subsequently utilized to generate acetate and propionate. This metabolic coupling captured spatial gradient effects of lactate metabolism and highlighted the ecological role of *V. parvula* as a secondary colonizer, demonstrating the importance of metabolic complementarity for community homeostasis. Initial simulation of co-culture simulation also showed enhanced EPS production, suggesting that *V. parvula* may indirectly support *S. mutans* growth and colonization through EPS-mediated aggregation, consistent with observed increases in biofilm thickness and structural complexity under co-culture conditions [16].

The integration of cellular mechanics (CPM) with biochemical processes (reaction–diffusion equations) demonstrates the feasibility and power of multi-scale modeling for complex microbial systems. The ABM framework provides a quantitative and spatially resolved

understanding of microbial behavior under caries-associated conditions, reproduces observed abundance patterns, and enables prediction of dynamic responses to environmental perturbations, such as nutrient fluctuations or metabolite accumulation. These capabilities offer a foundation for designing targeted *invitro* experiments and exploring potential strategies to modulate microbial interactions, ultimately supporting the development of personalized prevention and treatment approaches for pediatric caries.

5.3 Limitations

This study has several methodological limitations that should be acknowledged. First, the 16S rRNA sequencing data provide primarily genus-level resolution, preventing the identification of functionally distinct species or strains within genera and limiting the interpretation of specific metabolic capabilities or virulence factors. Second, the k-mer optimization and taxonomic classification strategies, including random subsampling and database-driven unique k-mer queries, may introduce selection biases or affect abundance estimates, particularly for low-abundance taxa. Third, findings were validated against only one published dataset, limiting generalizability across different populations or methodological contexts.

Regarding the 3D oral biofilm simulation, the model incorporates only two bacterial taxa, which simplifies the microbial community and may overlook interactions present in more diverse biofilms. Several key assumptions further limit the model's realism. First, the metabolic kinetic parameters are derived from literature values or approximate estimates and lack precise experimental validation under the specific culture conditions. Second, EPS is represented as a diffusible chemical substance, whereas in reality it forms a viscous, non-diffusible matrix that affects local viscosity and diffusion. Third, chemical, cellular mechanical, and growth processes are assumed to operate on strictly separated time scales, which may introduce inaccuracies under rapidly changing conditions. Fourth, the boundary conditions are idealized: the top surface assumes an infinite nutrient supply, while the bottom and side surfaces are zero-flux, differing from the dynamic flow and diffusion occurring in the oral environment. Finally, the model does not account for dynamic pH changes, immune responses, or fluid dynamics, all of which may substantially influence bacterial distribution and metabolite accumulation *invivo*.

Chapter 6

Conclusion and future work

This study investigated the metabolite dynamics of oral bacteria in two main stages. In the first stage, raw genetic sequence data were obtained from publicly available databases and processed using a custom k-mer-based pipeline. This approach enables rapid and efficient genus-level taxonomic profiling suitable for initializing agent-based simulations, with k-mers stored in an SQLite database for subsequent analysis and parameter optimization. Analysis of the oral microbiota of CF and SECC children using this pipeline revealed significant microbial shifts associated with caries development. Specifically, *Streptococcus* and *Veillonella* were significantly enriched in SECC samples, ranking among the top five most abundant genera. This observation aligns with established mechanisms of dental caries: *Streptococcus* drives acid production and biofilm formation, while *Veillonella* metabolizes lactic acid into short-chain fatty acids, modulating the acidic microenvironment. Some genera showed reduced abundance in SECC, consistent with the ecological plaque hypothesis, which posits that caries results from shifts in microbial community composition rather than the proliferation of a single pathogenic species [61].

In the second stage, a three-dimensional agent-based simulation was performed to explore interactions between metabolites and bacterial species. The CompuCell3D (CC3D) platform, combined with the Python package FiPy, simulated spatial diffusion of metabolites over time. To capture key dynamics while simplifying the system, two representative bacterial species, *S. mutans* and *V. parvula*, were modeled based on genus-level abundance data associated with SECC. This setup enables dynamic visualization of biofilm formation, agent migration, and metabolite-bacteria interactions, with initial cell distributions informed by sequencing-derived abundance profiles.

Despite these achievements, several limitations exist. The analysis was restricted to the genus level, and the ABM simulation included only two taxa, limiting representation of

the full microbial diversity and potential interactions within oral biofilms. The model does not account for environmental factors such as the spatial distribution of teeth, saliva composition, or gradients of pH and nutrients, restricting its representation of the local microenvironment where caries occur. While most previous studies investigate microbial dynamics at the species level, genus-level modeling is appropriate here because 16S rRNA data do not reliably resolve species distinctions [22, 62] and still allow capture of functional ecological interactions within biofilms [63].

Future work will focus on three directions: (1) pipeline enhancement, (2) model expansion, and (3) clinical validation. Pipeline enhancement may involve stratified k-mer optimization based on sequencing depth and community complexity, validation across diverse datasets, and integration of metagenomic data to achieve species-level resolution and functional annotation. Model expansion could include additional key genera (*Actinomyces, Corynebacterium, Neisseria*), finer taxonomic resolution, and incorporation of environmental factors such as spatial heterogeneity and metabolite diffusion, enabling more accurate simulation of biofilm dynamics and metabolic interactions. Clinical validation would leverage the predictive capacity of the model to explore caries progression trajectories and assess the effects of targeted interventions, reducing reliance on long-term wet-lab experiments and providing a cost- and time-efficient framework for early diagnosis and preventive strategies.

Chapter 7

Ethics and Data Management

I acknowledge that the thesis adheres to the ethical code (https://student.uva.nl/en/topics/ethics-in-research) and research data management policies (https://rdm.uva.nl/en) of UvA and IvI.

The following table lists the data used in this thesis (including source codes). I confirm that the list is complete and the listed data are sufficient to reproduce the results of the thesis. If a prohibitive non-disclosure agreement is in effect at the time of submission "NDA" is written under "Availability" and "License" for the concerned data items.

Short description (max. 10 words)	**Availability (e.g., URL, DOI)**	**License (e.g., MIT, GPL, Creative Commons)**
The supragingival plaque microbiota of pre-school children.	https://github.com/XiaoqingHan/16S-rRNA-Pipeline-hxq.git	GPL

Sample ID	**Age (months)**	**Sex**	**Health/Caries**	**FASTQ Accession**	**Group**
SAMN12315832	69	f	h	SRR9712102	CF1
SAMN12315851	50	m	h	SRR9712229	CF10
SAMN12315776	51	m	h	SRR9712232	CF11
SAMN12315774	42	f	h	SRR9712234	CF12
SAMN12315818	63	f	h	SRR9712243	CF13
SAMN12315825	36	f	h	SRR9712107	CF14
SAMN12315826	46	m	h	SRR9712108	CF15
SAMN12315824	50	f	h	SRR9712110	CF16
SAMN12315804	65	m	h	SRR9712149	CF17
SAMN12315805	28	f	h	SRR9712150	CF18
SAMN12315840	35	f	h	SRR9712211	CF19
SAMN12315829	18	m	h	SRR9712103	CF2
SAMN12315838	65	f	h	SRR9712213	CF20
SAMN12315837	34	m	h	SRR9712214	CF21
SAMN12315844	43	m	h	SRR9712222	CF22
SAMN12315780	63	f	h	SRR9712236	CF23
SAMN12315775	71	f	h	SRR9712231	CF24
SAMN12315816	63	f	h	SRR9712249	CF25

Sample ID	Age (months)	Sex	Health/Caries	FASTQ Accession	Group
SAMN12315831	51	f	h	SRR9712101	CF26
SAMN12315827	20	f	h	SRR9712105	CF27
SAMN12315823	50	m	h	SRR9712109	CF28
SAMN12315835	34	m	h	SRR9712216	CF29
SAMN12315830	44	m	h	SRR9712104	CF3
SAMN12315834	36	f	h	SRR9712217	CF30
SAMN12315833	25	m	h	SRR9712218	CF31
SAMN12315841	37	f	h	SRR9712220	CF32
SAMN12315852	42	f	h	SRR9712230	CF33
SAMN12315843	57	m	h	SRR9712221	CF34
SAMN12315847	33	f	h	SRR9712225	CF35
SAMN12315848	65	m	h	SRR9712226	CF36
SAMN12315773	56	f	h	SRR9712233	CF37
SAMN12315779	51	m	h	SRR9712235	CF38
SAMN12315822	41	f	h	SRR9712241	CF39
SAMN12315784	34	m	h	SRR9712186	CF4
SAMN12315821	44	m	h	SRR9712242	CF40
SAMN12315836	62	m	h	SRR9712215	CF5
SAMN12315842	67	f	h	SRR9712219	CF6
SAMN12315845	28	m	h	SRR9712223	CF7
SAMN12315846	38	f	h	SRR9712224	CF8
SAMN12315849	40	m	h	SRR9712227	CF9
SAMN12315828	68	f	c	SRR9712106	SECC1
SAMN12315801	30	f	c	SRR9712123	SECC10
SAMN12315795	35	f	c	SRR9712125	SECC11
SAMN12315807	47	f	c	SRR9712144	SECC12
SAMN12315809	40	f	c	SRR9712146	SECC13
SAMN12315810	62	f	c	SRR9712147	SECC14
SAMN12315806	34	f	c	SRR9712151	SECC15
SAMN12315812	44	f	c	SRR9712153	SECC16
SAMN12315785	40	f	c	SRR9712183	SECC17
SAMN12315783	28	m	c	SRR9712185	SECC18
SAMN12315850	70	m	c	SRR9712228	SECC19
SAMN12315793	40	f	c	SRR9712127	SECC2
SAMN12315778	45	m	c	SRR9712238	SECC20
SAMN12315781	37	m	c	SRR9712239	SECC21
SAMN12315782	31	f	c	SRR9712240	SECC22
SAMN12315820	53	m	c	SRR9712245	SECC23
SAMN12315815	59	m	c	SRR9712250	SECC24
SAMN12315802	38	f	c	SRR9712122	SECC25
SAMN12315796	40	f	c	SRR9712124	SECC26
SAMN12315794	52	m	c	SRR9712126	SECC27
SAMN12315799	70	f	c	SRR9712129	SECC28
SAMN12315798	41	m	c	SRR9712130	SECC29
SAMN12315808	47	f	c	SRR9712145	SECC3
SAMN12315797	44	m	c	SRR9712131	SECC30
SAMN12315786	33	f	c	SRR9712184	SECC31
SAMN12315789	57	m	c	SRR9712187	SECC32
SAMN12315790	52	f	c	SRR9712188	SECC33
SAMN12315788	47	f	c	SRR9712190	SECC34
SAMN12315791	31	m	c	SRR9712191	SECC35
SAMN12315839	52	m	c	SRR9712212	SECC36
SAMN12315817	33	m	c	SRR9712244	SECC37
SAMN12315819	40	m	c	SRR9712246	SECC38

Sample ID	Age (months)	Sex	Health/Caries	FASTQ Accession	Group
SAMN12315814	39	m	c	SRR9712247	SECC39
SAMN12315800	56	f	c	SRR9712128	SECC4
SAMN12315777	40	f	c	SRR9712237	SECC40
SAMN12315803	33	f	c	SRR9712148	SECC5
SAMN12315811	49	f	c	SRR9712152	SECC6
SAMN12315787	52	f	c	SRR9712189	SECC7
SAMN12315792	51	f	c	SRR9712192	SECC8
SAMN12315813	64	f	c	SRR9712248	SECC9

TABLE 7.1: Summary of 80 supragingival plaque samples (40 CF, 40 SECC) downloaded from HOMD.

Bibliography

[1] Selwitz RH, Ismail AI, and Pitts NB. Dental caries. *The Lancet*, 369(9555):51–59, 1 2007. URL https://doi.org/10.1016/S0140-6736(07)60031-2.

[2] Saikia A, Aarthi J, Muthu MS, Patil SS, Anthonappa RP, Walia T, Shahwan M, Mossey P, and Dominguez M. Sustainable development goals and ending ecc as a public health crisis. *Front Public Health*, 10(931243), 10 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9624450/.

[3] Anil S and Anand PS. Early childhood caries: Prevalence, risk factors, and prevention. *Front Pediatr*, 5(157), 7 2017. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5514393/#abstract1.

[4] Gupta P, Gupta N, Pawar AP, Birajdar SS, Natt AS, and Singh HP. Role of sugar and sugar substitutes in dental caries: a review. *ISRN Dent*, 2013(519421), 12 2013. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3893787/.

[5] Alarcón-Sánchez MA, Becerra-Ruiz JS, Avetisyan A, and Heboyan A. Activity and levels of tnf-, il-6 and il-8 in saliva of children and young adults with dental caries: a systematic review and meta-analysis. *BMC Oral Health*, 24(816), 7 2024. URL https://doi.org/10.1186/s12903-024-04560-8.

[6] Sampaio-Maia B and Monteiro-Silva F. Acquisition and maturation of oral microbiome throughout childhood: An update. *Dent Res J*, 11(3):291–301, 5 2014. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4119360/.

[7] Butler CA, Adams GG, Blum J, Byrne SJ, Carpenter L, Gussy MG, Calache H, Catmull DV, Reynolds EC, and Dashper SG. Breastmilk influences development and composition of the oral microbiome. *J Oral Microbiol*, 14(1):2096287, 7 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9272919/#abstract1.

[8] M.G. Dominguez-Bello, E.K. Costello, M. Contreras, M. Magris, N. Fierer G. Hidalgo, and R. Knight. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad.*

Sci. U.S.A., 107(26):11971–11975, 5 2010. URL https://doi.org/10.1073/pnas.1002601107.

[9] Lif Holgerson P, Esberg A, Sjödin A, E.West C, and Johansson I. A longitudinal study of the development of the saliva microbiome in infants 2 days to 5 years compared to the microbiome in adolescents. *Sci Rep*, 10(9629), 6 2020. URL https://www.nature.com/articles/s41598-020-66658-7.

[10] Li X, Liu Y, Yang X, Li C, and Song Z. The oral microbiota: Community composition, influencing factors, pathogenesis, and interventions. *Front Microbiol*, 13 (895537), 4 2022. URL https://pubmed.ncbi.nlm.nih.gov/35572634/.

[11] Wake N, Asahi Y, Noiri Y, Hayashi M, Motooka D, Nakamura S, Gotoh K, Miura J, Machi H, Iida T, and Ebisu S. Temporal dynamics of bacterial microbiota in the human oral cavity determined using an in situ model of dental biofilms. *npj Biofilms Microbiomes*, 2(16018), 8 2016. URL https://doi.org/10.1038/npjbiofilms.2016.18.

[12] Allan Radaic and Yvonne L. Kapila. The oralome and its dysbiosis: New insights into oral microbiome-host interactions. *Computational and Structural Biotechnology Journal*, 19:1335–1360, 2 2021. URL https://doi.org/10.1016/j.csbj.2021.02.010.

[13] Giordano-Kelhoffer B, Lorca C, March Llanes J, Rábano A, Del Ser T, Serra A, and Gallart-Palau X. Oral microbiota, its equilibrium and implications in the pathophysiology of human diseases: A systematic review. *Biomedicines*, 10(8): 1803, 7 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9405223/.

[14] Vila T, Rizk AM, Sultan AS, and Jabra-Rizk MA. The power of saliva: Antimicrobial and beyond. *PLoS Pathog*, 15(11):e1008058, 11 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6855406/.

[15] Lemos JA, Palmer SR, Zeng L, Wen ZT, Kajfasz JK, Freires IA, Abranches J, and Brady LJ. The biology of streptococcus mutans. *Microbiol Spectr*, 7(1), 1 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6615571/.

[16] Liu S, Chen M, Wang Y, Zhou X, Peng X, Ren B, Li M, and Cheng L. Effect of veillonella parvula on the physiological activity of streptococcus mutans. *Arch Oral Biol*, 109(104578), 1 2020. URL https://www.sciencedirect.com/science/article/pii/S000399691930562X?via%3Dihub.

[17] Rath S, Bal SCB, and Dubey D. Oral biofilm: Development mechanism, multidrug resistance, and their effective management with novel techniques. *Rambam Maimonides medical journal*, 12(1):e0004, 1 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7835112/.

[18] Wicaksono DPWashio J, Abiko Y, Domon H, and Takahashi N. Nitrite production from nitrate and its link with lactate metabolism in oral veillonella spp. *Appl Environ Microbiol*, 86(e01255):20, 10 2020. URL https://doi.org/10.1128/AEM.01255-20.

[19] Wei Y, Zhang Y, Zhuang Y, Tang Y, Nie H, Haung Y, Liu T, Yang W, Yan F, and Zhu Y. Veillonella parvula acts as a pathobiont promoting the biofilm virulence and cariogenicity of streptococcus mutans in adult severe caries. *Microbiol Spectr*, 12(11):e0431823, 11 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11537095/.

[20] Feng D, Neuweiler I, Nogueira R, and Nackenhorst U. Modeling of symbiotic bacterial biofilm growth with an example of the streptococcus-veillonella sp. system. *Bull Math Biol*, 83(5):48, 3 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7990864/#Sec2.

[21] Head DA, Marsh PD, and Devine DA. Non-lethal control of the cariogenic potential of an agent-based model for dental plaque. *PLoS One*, 9(8):e105012, 8 2014. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4140729/?utm_source=chatgpt.com.

[22] Janda JM and Abbott SL. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J Clin Microbiol*, 45(9), 9 2007. URL https://doi.org/10.1128/jcm.01228-07.

[23] Alex M Valm. The structure of dental plaque microbial communities in the transition from health to dental caries and periodontal disease. *J Mol Biol*, 431(16):2957–2969, 7 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6646062/.

[24] Berger D, Rakhamimova A, Pollack A, and Loewy Z. Oral biofilms: Development, control, and analysis. *High Throughput*, 7(3):24, 8 2018. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6163956/.

[25] Liu X, Yao H, Zhao X, and Ge C. Biofilm formation and control of foodborne pathogenic bacteria. *Molecules*, 28(6):2432, 3 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10058477/.

[26] Bowen WH, Burne RA, Wu H, and Koo H. Oral biofilms: Pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends Microbiol*, 26(3):229–242, 3 2018. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5834367/.

[27] Kreve S and Reis ACD. Bacterial adhesion to biomaterials: What regulates this attachment? a review. *Jpn Dent Sci Rev*, 57:85–96, 11 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8215285/#abs0005.

[28] Huang R, Li M, and Gregory RL. Bacterial interactions in dental biofilm. *Virulence*, 2(5):435–44, 9 2011. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3322631/#sec2.

[29] Abebe GM. Oral biofilm and its impact on oral health, psychological and social interaction. *Int J Oral Dent Health*, 7(127), 3 2021. URL https://www.clinmedjournals.org/articles/ijodh/international-journal-of-oral-and-dental-health-ijodh-7-127.php?jid=ijodh.

[30] Du Q, Fu M, Zhou Y, Cao YP, Guo TW, Zhou Z, Li MY, Peng X, Zheng X, Li Y, Xu X, He JZ, and Zhou XD. Sucrose promotes caries progression by disrupting the microecological balance in oral biofilms: an in vitro study. *Sci Rep*, 10(2961), 2020. URL https://doi.org/10.1038/s41598-020-59733-6.

[31] Izumi Mashima and Futoshi Nakazawa. The influence of oral veillonella species on biofilms formed by streptococcus species. *Anaerobe*, 28:54–61, 8 2014. URL https://doi.org/10.1016/j.anaerobe.2014.05.003.

[32] Rupf S, Merte K, Eschrich K, and Kneist S. Streptococcus sobrinus in children and its influence on caries activity. *European Archives of Paediatric Dentistry*, 1:17–22, 3 2006. URL https://link.springer.com/article/10.1007/BF03320810.

[33] Ribeiro AA and Paster BJ. Dental caries and their microbiomes in children: what do we do now? *J Oral Microbiol*, 15(1):2198433, 4 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10088930/.

[34] Sucheta Prabhu Matondkar, Chandrashekar Yavagal, Manohar Kugaji, and Kishore G. Bhat. Quantitative assessment of scardovia wiggsiae from dental plaque samples of children suffering from severe early childhood caries and caries free children. *Anaerobe*, 62(102110), 4 2020. URL https://doi.org/10.1016/j.anaerobe.2019.102110.

[35] Zhou P, Manoil D, Belibasakis GN, and Kotsakis GA. Veillonellae: Beyond bridging species in oral biofilm ecology. *Front Oral Health*, 2(774115), 10 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8757872/#s5.

[36] Anderson M, Grindefjord M, Dahllöf G, Dahlén G, and Twetman S. Oral microflora in preschool children attending a fluoride varnish program: a cross-sectional study. *BMC Oral Health*, 16:130, 12 2016. URL https://doi.org/10.1186/s12903-016-0325-6.

[37] NI Chalmers, K Oh, Hughes CV, N Pradhan, Kanasi E, Ehrlich Y, Dewhirst FE, and Tanner ACR. Pulp and plaque microbiotas of children with severe early childhood caries. *J Oral Microbiol*, 2(7):25951, 2015. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4317471/.

[38] Tanner AC, Mathney JM, Kent RL, Chalmers NI, Hughes CV, Loo CY, Pradhan N, Kanasi E, Hwang J, Dahlan MA, Papadopolou E, and Dewhirst FE. Cultivable anaerobic microbiota of severe early childhood caries. *J Clin Microbiol*, 49(4):1464–74, 4 2011. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3122858/.

[39] Do T, Sheehy EC, Mulli T, Hughes F, and Beighton D. Transcriptomic analysis of three veillonella spp. present in carious dentine and in the saliva of caries-free individuals. *Front Cell Infect Microbiol*, 5(25), 3 2015. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4374535/.

[40] Gupta S, Mortensen MS, Schjørring S, Trivedi U, Vestergaard G, Stokholm J, Bisgaard H, Krogfelt KA, and Sørensen SJ. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Commun Biol*, 2(291), 8 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6683184/.

[41] Wensel CR, Pluznick JL, Salzberg SL, and Sears CL. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J Clin Invest*, 132(7):e154944, 4 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8970668/.

[42] Ranjan R, Rani A, Metwally A, McGee HS, and Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochem Biophys Res Commun*, 469(4):967–977, 1 2016. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4830092/#S22.

[43] Dacey DP and Chain FJJ. Concatenation of paired-end reads improves taxonomic classification of amplicons for profiling microbial communities. *BMC Bioinformatics*, 22(493), 10 2021. URL https://doi.org/10.1186/s12859-021-04410-2.

[44] Gallert C Jeske JT. Microbiome analysis via otu and asv-based pipelines-a comparative interpretation of ecological data in wwtp systems. *Bioengineering (Basel)*, 9(4):146, 3 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9029325/#abstract1.

[45] Moeckel C, Mareboina M, Konnaris MA, Chan CSY, Mouratidis I, Montgomery A, Chantzi N, Pavlopoulos GA, and Georgakopoulos-Soares I. A survey of k-mer methods and applications in bioinformatics. *Comput Struct Biotechnol J*, 23:2289–2303, 5 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11152613/#sec0010.

[46] Lindgreen S, Adair KL, and Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep*, 6(19233), 1 2016. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4726098/#abstract1.

[47] MacDonald ML, Polson SW, and Lee KH. k-mer-based metagenomics tools provide a fast and sensitive approach for the detection of viral contaminants in biopharmaceutical and vaccine manufacturing applications using next-generation sequencing. *msphere*, 6(2), 4 2021. URL https://journals.asm.org/doi/10.1128/msphere.01336-20.

[48] Jacques Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3:371–394, 10 1949. URL 10.1146/annurev.mi.03.100149.002103.

[49] Somvanshi, Pradeep R, and KV Venkatesh. Hill equation. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*. Springer, New York, NY, 2013. doi: 10.1007/978-1-4419-9863-7_946.

[50] Leonor Michaelis and Maud Leonora May Menten. Die kinetik der invertinwirkung. *Biochemische Zeitschrift*, 49(333–369):352, 1913. URL https://doi.org/10.1007/BF01328285.

[51] de Jesus VC, Shikder R, Oryniak D, Mann K, Alamri A, Mittermuller B, Duan K, Hu P, Schroth RJ, and Chelikani P. Sex-based diverse plaque microbiota in children with severe caries. *J Dent Res*, 99(6):703–712, 6 2020. URL https://journals.sagepub.com/doi/10.1177/0022034520908595?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed.

[52] Wood DE and Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*, 15(3):3, 3 2014. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4053813/.

[53] Wood DE, Lu J, and Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*, 20(257), 11 2019. URL https://doi.org/10.1186/s13059-019-1891-0.

[54] Ounit R, Wanamaker S, Close TJ, and Lonardi S. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(236), 3 2015. URL https://doi.org/10.1186/s12864-015-1419-2.

[55] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, and Phillippy AM. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol*, 17(132), 6 2016. URL https://doi.org/10.1186/s13059-016-0997-x.

[56] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci*, 99:7280–7287, 5 2002. URL https://doi.org/10.1073/pnas.082080899.

[57] Estaki M, Jiang L, Bokulich NA, González A McDonald D, Kosciolek T, Martino C, Zhu Q, Birmingham A, Vázquez-Baeza Y, Dillon MR, Bolyen E, Caporaso JG, and Knight R. Qiime 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current Protocols in Bioinformatics*, 70(e100), 4 2020. URL https://pubmed.ncbi.nlm.nih.gov/32343490/.

[58] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, and Holmes SP. Dada2: High-resolution sample inference from illumina amplicon data. *Nat Methods*, 13(7):581–3, 7 2016. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4927377/.

[59] Bussi Y, Kapon R, and Reich Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLOS ONE*, 16 (10):e0258693, 10 2021. URL https://doi.org/10.1371/journal.pone.0258693.

[60] Pornputtapong N, Acheampong DA, Patumcharoenpol P, Jenjaroenpun P, Wongsurawat T, Jun S-R, Yongkiettrakul S, Chokesajjawatee N, and Nookaew I. Kitsune: A tool for identifying empirically optimal k-mer length for alignment-free phylogenomic analysis. *Frontiers in Bioengineering and Biotechnology*, 8, 9 2020. URL https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2020.556413.

[61] Marsh PD. Microbial ecology of dental plaque and its significance in health and disease. *Advances in Dental Research*, 8(2):263–271, 7 1994. URL https://journals.sagepub.com/doi/10.1177/08959374940080022001.

[62] Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, and Weinstock GM. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nat Commun*, 10(5029), 11 2019. URL https://doi.org/10.1038/s41467-019-13036-1.

[63] Marsh PD. Dental plaque as a biofilm and a microbial community - implications for health and disease. *BMC Oral Health*, 6(1):S14, 6 2006. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2147593/.