Building Predictive Understanding of Microbial Ecology by Bridging Microbial Growth Kinetics and Microbial Population Dynamics

Zhang Cheng [a], Weibo Xia [a], Sean McKelvey [a,b], Qiang He [c], Yuzhou Chen [d,*], Heyang Yuan [a,*].

[a] Department of Civil & Environmental Engineering, Temple University, 1947 N. 12th Street, Philadelphia, PA 19122, U.S.

[b] Philadelphia Water Department, ADRESS, Philadelphia, PA ZIP CODE, U.S.

[c] Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Knoxville, TN 37996, U.S.

[d] Department of Computer & Information Sciences, Temple University, 1947 N. 12th Street, Philadelphia, PA 19122, U.S.

Intended for ***TBD***

Type of contribution***: Research Article***

*** Corresponding author**

Yuzhou Chen: yuzhou.chen@temple.edu

Heyang Yuan: heyang.yuan@temple.edu

Abstract

Modeling microbial communities can provide predictive insights into microbial ecology, but current modeling approaches suffer from inherent limitations. In this study, a novel modeling approach was proposed to address those limitations based on the intrinsic connection between the growth kinetics of guilds and the dynamics of individual microbial populations. To implement the modeling approach, 466 samples from four full-scale activated sludge systems were retrieved from the literature. The raw samples were processed using a data transformation method that not only increased the dataset size by three times but also enabled quantification of population dynamics. Most of the 42 family-level core populations showed overall dynamics close to zero within the sampling period, explaining their resilience to environmental perturbation. Bayesian networks built with environmental factors, perturbation, historical abundance, population dynamics, and mechanistically derived microbial kinetic parameters classified the core populations into heterotrophic and autotrophic guilds. Topological data analysis was applied to identify keystone populations and their time-dependent interactions with other populations. The data-driven inferences were validated directly using the Microbial Database for Activated Sludge (MiDAS) and indirectly by predicting population abundance and community structure using artificial neural networks. The Bray-Curtis similarity between predicted and observed communities was significantly higher with microbial kinetic parameters than without parameters (0.70 vs. 0.66), demonstrating the accuracy of the modeling approach. Implemented based on engineered systems, this modeling approach can be generalized to natural systems to gain predictive understandings of microbial ecology.

Keywords: Microbial ecology; Microbial population dynamics; Microbial growth kinetics; Mechanistic modeling; Data-driven modeling.

1. Introduction

Microorganisms play critical roles in various ecosystems by interacting with each other and forming microbial communities (Falkowski et al. 2008, Rittmann and McCarty 2012, Yatsunenko et al. 2012). How individual microorganisms function and interact within communities has always been a core question in microbial ecology (Coyte et al. 2015, Fuhrman 2009, Fuhrman et al. 2015, Stams and Plugge 2009, Torsvik and Øvreås 2002).

Our understanding of microbial ecology has been significantly advanced by the development of experimental and computational approaches (Ju and Zhang 2015b, Vanwonterghem et al. 2014). These include marker gene amplicon sequencing (Caporaso et al. 2012), metagenomic and transcriptomic sequencing (Quince et al. 2017, Stewart et al. 2018), multivariate statistical analysis (Buttigieg and Ramette 2014), network analysis (Weiss et al. 2016), and bioinformatics tools (Douglas et al. 2020, Langille et al. 2013), etc. The massive amount of data and knowledge accumulated over the past decades has motivated the development of modeling approaches for a predictive understanding of the functions and interactions within communities (Larsen et al. 2012a, Lopatkin and Collins 2020). Two distinct approaches have been used to model microbial communities: mechanistic modeling and data-driven modeling (Widder et al. 2016, Yao et al. 2022).

Mechanistic modeling can provide insight into the interactions between microbial populations by simulating their growth kinetics (Kumar et al. 2019, Song et al. 2014). Originally used to describe the growth of a single organism on a given substrate (Kovárová-Kovar and Egli 1998, Monod 1942), microbial growth kinetics has been extended to simulate the overall growth of a guild (i.e.,

a group of populations collectively responsible for a function) on a class of substrates (e.g., organic carbon) (Veshareh and Nick 2021). For example, communities in activated sludge system (an engineered system widely used for wastewater treatment) are modeled by simulating the growth kinetics of four guilds: organic-degrading heterotrophs, ammonia-oxidizing autotrophs, denitrifying bacteria, and phosphate-accumulating organisms (Gujer et al. 1995, Gujer et al. 1999, Henze et al. 1987, Henze et al. 1999). This extension has made mechanistic modeling a robust approach for understanding microbial ecology at the guild level (Batstone et al. 2002, Bouskill et al. 2012, Jin and Roden 2011, Wieder et al. 2015, Wieder et al. 2013). However, the growth kinetics of guilds provide little information about the functions and interactions of individual populations observed in the community. Inferring the functions and interactions of individual populations by simulating their growth kinetics remains challenging due to the lack of knowledge about their growth behavior (Ansari et al. 2021, Rinke et al. 2013).

Data-driven modeling can be used to learn the functions and interactions of microbial populations from their abundance (Kumar et al. 2019, Larsen et al. 2015). This is achieved by capturing the latent relationship between populations and their environments using statistical, probabilistic, or machine learning methods (Ghannam and Techtmann 2021, Larsen et al. 2012a, Mowbray et al. 2021). For example, Bayesian networks, a probabilistic graphical model that can reveal causal relationships between variables via directed acyclic graphs (Uusitalo 2007), have been built to infer microbial interactions and their effects on community structure (Lax et al. 2014, Metcalf et al. 2016, Yuan et al. 2017). Despite its potential, the application of data-driven modeling in microbial ecology is limited by the challenges of data collecting and processing. Because population abundance is difficult and expensive to collect, data-driven models are typically built

71 with small datasets of fewer than a hundred samples (Kuang et al. 2016, Larsen et al. 2012b, Lesnik
72 et al. 2020, Lesnik and Liu 2017, Staley et al. 2014, Yuan et al. 2017). Moreover, population
73 abundance is not adequately processed to capture the variability of microbial functions and
74 interactions at different temporal scales (Ruan et al. 2006, Xia et al. 2011). These challenges can
75 severely compromise the robustness of data-driven inference (Althnian et al. 2021, Ghannam and
76 Techtmann 2021).
77
78 Here, a new modeling approach is proposed to leverage mechanistic and data-driven modeling to
79 learn the functions and interactions of individual populations from the growth kinetics of guilds.
80 The mathematical foundation of the proposed approach is revealed by the following derivation. In
81 mechanistic modeling, the total abundance of a guild (X) is simulated as first-order kinetics
82 (without considering flow for simplicity):

$$\frac{dX}{dt} = (\mu - b)X \quad \text{Eq. 1}$$

83 where μ and b are the growth and decay rates, respectively. The integration of Eq. 1 yields the
84 explicit expression of the specific growth rate of the guild within any given time span (Δt):

$$\mu - b = \frac{1}{\Delta t} \ln \frac{X}{X^*} \quad \text{Eq. 2}$$

85 where X^* is the historical abundance the guild. The specific growth rate μ can be further expressed
86 as a function of substrate concentration using the Monod equation (Monod 1942, 1949):

$$\mu = \mu_{max} \frac{S}{K + S} \quad \text{Eq. 3}$$

87 where μ_{max} is the maximum specific growth rate, K is the Monod constant, and S is the substrate
88 concentration. Assuming that the guild is composed of n populations, the abundance of the guild
89 is the sum of the abundances of all individual populations (X_i) associated with the guild:

$$X = \sum_{i=1}^{n} X_i \quad \text{Eq. 4}$$

Combining Eqs. 2-4 yields:

$$\mu_{max} \frac{S}{K+S} - b = \frac{1}{\Delta t} \ln \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^*} \quad \text{Eq. 5}$$

where X_i^* is the historical abundance of the i-th population within the guild. Eq. 5 indicates that, within any given time span Δt, the growth kinetics of a guild (numerically represented by its microbial kinetic parameters) are related to the dynamics of individual populations (numerically represented by $1/\Delta t \cdot \ln X_i/X_i^*$). Such an intrinsic connection can be learned using data-driven approaches to infer if a population contributes to the function of the guild and how it interacts with other populations within the guild.

In this study, the proposed modeling approach was implemented based on activated sludge systems. This engineered system has been applied worldwide to treat wastewater, and the Activated Sludge Model No.1 (ASM1) is a well-established mechanistic model that can be readily used to derive the microbial kinetic parameters of guilds (Henze et al. 2000). Meanwhile, activated sludge communities have been extensively studied, and the functional populations in the communities are well curated (McIlroy et al. 2017, Nierychlo et al. 2020, Saunders et al. 2016, Wu et al. 2019). Therefore, activated sludge systems represent an ideal level of complexity for linking microbial growth kinetics to microbial population dynamics. To implement the modeling approach, time series data from four full-scale activated sludge systems were retrieved from the literature. A data transformation method was developed to augment the data and capture the temporal variability of microbial functions and interactions as required by Eq. 5. After data transformation and processing, Bayesian networks were trained to infer functions, and inferences were validated with the

Microbial Database for Activated Sludge (MiDAS) (Dueholm et al. 2022). Interactions were further inferred using topological data analysis. Inferences were also validated indirectly by predicting community structure. Implemented based on activated sludge systems, the modeling approach is expected to be broadly applicable to other natural and engineered ecosystems to provide predictive insight into the functions and interactions within communities.

2. Materials and Methods

2.1 Data collection and preprocessing

A total of 466 samples were retrieved from three studies to develop the modeling approach (Jiang et al. 2018, Peces et al. 2022, Sun et al. 2021). These studies were selected because they comprehensively reported the environmental factors of four full-scale activated sludge systems in the form of time series (Table 1) and properly deposited sequencing data of 16s rRNA gene amplicon in the National Center for Biotechnology Information database. The diversity of the samples was analyzed using principal coordinate analysis (PCoA) based on Bray-Curtis distance.

Table 1. Overview of the four full-scale activated sludge systems

	Hong Kong	Beijing-1	Beijing-2	Aalborg
Location	(22°N, 114°E)	(40°N, 117°E)	(40°N, 110°E)	(57°N, 10°E)
Q (m^3/day)	2.6×10^5	1.0×10^6	1.0×10^5	1.4×10^5
Sampling Period	2013 - 2014	2015 - 2016	2015 - 2016	2017 - 2020
Number of samples	260	42	42	122
BOD (mg/L)	110 - 360	180 - 400	230 - 1300	60 - 430
Temperature (°C)	7 - 31	15 - 27	15 - 29	8 - 22
SRT (day)	9 - 19	~10	~17	10 - 28
MLVSS (mg/L)	1400 - 3300	1500 - 3300	3500 - 9900	1900 - 3900

Q: treatment capacity; BOD: biochemical oxygen demand; SRT: sludge retention time; MLVSS: mixed-liquid volatile suspended solid.

Sequencing data were retrieved from the database and preprocessed as previously described (Cheng et al. 2021b). Briefly, the QIIME2 pipeline was loaded with DADA2 for denoising the reads (Bolyen et al. 2019). The Silva database (LTPs132_SSU.arb for 16s rDNA, updated in June 2018) was used for sequence classification (Quast et al. 2013). The raw sequences were classified using A classifier that was trained on OTU sequences at 97% identity from the database. The communities in the four activated sludge systems were compared with those from other activated sludge systems using Bray-Curtis distance-based PCoA (Gómez-Acata et al. 2017, Ju and Zhang 2015a, Liu et al. 2017, Ouyang et al. 2016, Saunders et al. 2016, Yuan et al. 2019b, Zhang et al. 2017).

Core populations were selected at the family level based on average relative abundance and occurrence frequency across all 466 samples (Ling et al. 2016, Yuan et al. 2019b). More details about core population selection can be found in the Supporting Information (SI) Methods. Redundancy analysis (RDA) was performed on relative abundance and environmental factors. Absolute abundances of core populations were calculated by multiplying relative abundances by volatile suspended solids, a wastewater quality parameter commonly used to represent biomass concentration (Rittmann and McCarty 2012, Saunders et al. 2016).

2.2 Data transformation

The 466 samples were randomly divided into training (380 samples) and test (86 samples) sets. Both sets were transformed using the method outlined in Figure 1. First, the time span (Δt) was varied from 7 to 14 days based on the assumption that they could adequately reflect the dynamics of individual populations and the growth kinetics of guilds (Weissman et al. 2021). At each time

span, historical and current samples were drawn from the raw datasets to form transformed samples. Taking $\Delta t = 7$ d as an example (Figure 1), samples at $t = 1$ (historical) and $t = 7$ (present) were combined to form the first transformed sample, samples at $t = 2$ (historical) and $t = 8$ (present) were combined to form the second transformed sample, and so on. Transformation is repeated for the four activated sludge systems and eight time spans, followed by combining all transformed samples. The 380 raw samples in the training set and 86 samples in the test set were converted into 936 and 134 transformed samples, respectively.

Following the transformation, several new features were derived to capture the variability of microbial functions and interactions (Figure 1). First, environmental perturbations, key drivers of population dynamics (Griffin and Wells 2016, Vuono et al. 2014, Yuan et al. 2019a), were quantified as the change in environmental factors normalized to the time span. Second, population dynamics was quantified as the natural logarithm of present versus historical absolute abundances normalized to the time span (SI Method). RDA was performed on environmental perturbation and population dynamics. Finally, microbial kinetic parameters of guilds were derived by inputting the environmental factors within a time span into the ASM1 (SI Methods) (Cheng et al. 2024, Cheng et al. 2021a). The ASM1 was chosen over other activated sludge models because the selected studies only reported data related to heterotrophic organic removal and autotrophic ammonium oxidation. The derived parameters included the maximum specific growth rate μ_{max}, Monod constant K, biomass yield Y, and decay rate b of heterotrophs and autotrophs.

2.3 Model construction

Bayesian networks were built to learn the intrinsic connection between growth kinetics and population dynamics (Eq. 5). The first step of network construction was to scale the values of variables, including time span, historical environmental factors, historical population abundances, environmental perturbation, microbial kinetic parameters and population dynamics, to between 0 and 1 (Bishop 1995):

$$\bar{v}_j = \frac{v_j - v_{min}}{v_{max} - v_{min}} \quad \text{Eq. 6}$$

where $\bar{v}_j$ and v_j are the normalized and actual values of variable v in sample j, respectively; v_{min} and v_{max} are the minimum and maximum values of variable v, respectively. Next, network structure learning was performed using the R package "bnlearn" with a hill-climbing algorithm and the following assumptions (Scutari et al. 2014): 1) each variable is normally distributed, and variables are multivariate normal; 2) stochastic dependencies are assumed to be linear; 3) time span, historical environmental factors, historical population abundances and environmental perturbation are root parent nodes and cannot depend on any other nodes; 4) for any given population, its dynamics strictly depends on its historical abundance. A final Bayesian network was obtained by averaging 100 bootstrap replicates. Arc strengths representing the frequency of presence in all networks were exported with a threshold calculated by a built-in function in "bnlearn". Based on the Bayesian inferences, topological data analysis was performed to further capture the microbial interactions within the heterotrophic and autotrophic guilds (SI Methods).

2.4 Inference validation

The inferred functions (indicated by the dependencies between kinetic parameters and population abundance/dynamics) from the Bayesian network were validated by searching individual populations in the MiDAS. Within each family, the three most abundant genera with a relative

196 abundance reaching 0.1% were mapped to their corresponding functions in the MiDAS, with
197 particular emphasis on their contribution to heterotrophic organic degradation and autotrophic
198 ammonia oxidation (Dueholm et al. 2022).
199
200 As an additional validation step, artificial neural networks (ANNs) were trained to predict
201 population abundance and community structure. To prepare for ANN training, the root parent
202 nodes of population dynamics were identified using a depth-first search algorithm (Heineman et
203 al. 2008). ANNs were then trained with root parent nodes and population dynamics as inputs and
204 outputs, respectively. The hyperparameters of the ANNs were optimized via grid search using the
205 R package Neuralnet (Wright 2019). To perform the optimization, the transformed training set
206 (936 transformed samples) was first divided into 13 subsets (72 transformed samples in each
207 subset). A series of ANNs were then trained with the combinations of different number of hidden
208 layers (2 - 5) and number of nodes (12, 18, 24). The prediction performance of each ANN was
209 evaluated with relative root mean square error (RMSE) based on 13-fold cross-validation (Anguita
210 et al. 2012, Roberts et al. 2017):

$$relative\ RMSE = \frac{\sqrt{\frac{\sum(\hat{v}_j - v_j)^2}{j}}}{v_{max}} \quad \text{Eq. 7}$$

211 where $\hat{v}_j$ is the predicted value of variable v in sample j. After hyperparameter optimization, the
212 ANNs with optimal configurations were tested on 134 transformed samples. The ANN outputs
213 (i.e., population dynamics) were converted back to population abundance using historical
214 abundance and time span. The Bray-Curtis similarity between the predicted and observed
215 communities was then calculated.
216

Two control models were built to compare the prediction performance. The first one was built following previous studies without transforming the data or incorporating microbial growth kinetics (Kuang et al. 2016, Larsen et al. 2012b). A second one is a null model constructed with the mean values of the variables (Harvey et al. 1983).

3. Results

3.1 Overview of the activated sludge systems

The four activated sludge systems selected for model development were operated under distinct conditions (SI Figure S1). For example, sludge retention time, a key operating factor for microbial community structure (Mansfeldt et al. 2019, Yuan et al. 2019a), varied from 10 days in the Beijing-1 system to 17 days in the Beijing-2 system. The mean temperature and dissolved oxygen were higher in Hong Kong and Beijing than in Aalborg (t-test, $p < 0.05$, SI Figure S1). Moreover, the BOD (600 mg/L) and ammonia (50 mg/L) concentrations in the influent of Beijing-2 were significantly higher than those of Hong Kong (190 and 30 mg/L), Beijing-1 (260 and 45 mg/L), and Aalborg (200 and 30 mg/L) (t-test, $p < 0.05$).

The performance of the systems was also noticeably different (SI Figure S1). Specifically, the effluent BODwas much higher in the two Beijing systems (> 30 mg/L) than in the other two systems (< 5 mg/L), whereas the effluent ammonium showed an opposite trend. On the other hand, the biomass concentration (represented by volatile suspended solids) in Beijing-2 was more than two times higher than in the other systems. As shown by PCoA (Figure 2), the two Beijing systems were significantly different despite their geographic affinity. Although samples from Beijing-1 overlapped with those from Hong Kong and Aalborg, they shifted toward different directions

during the seasonal variation (SI Figure S2). Overall, the results show a high diversity of the samples, which is desirable for building robust models.

3.2 Overview of the activated sludge communities

The communities in the four systems showed location-dependent differentiation (SI Figure S3). The communities in the two Beijing systems, despite their difference in environmental factors, clustered together and were separate from those in Hong Kong and Aalborg. Similar results were observed in a comprehensive survey of global activated sludge microbiomes (Wu et al. 2019). The communities in Hong Kong and Aalborg showed clear shifts, which could be driven by deterministic factors such as sludge retention time, temperature, and influent concentrations. As shown in SI Figure S1, sludge retention time fluctuated more dramatically in Hong Kong and Aalborg than in Beijing. This environmental factor has been shown to exert strong selective pressure and drive deterministic community assembly (Mansfeldt et al. 2019, Vuono et al. 2014). Although the communities in the three sites formed individual clusters, they belonged to an activated sludge microbiome formed by eleven systems in China, Mexico, the U.S., and Denmark, demonstrating the representativeness of the communities being modeled.

A total of 42 core populations were selected at the family level based on the following criteria: average abundance > 0.5% and occurrence > 35% across all 466 samples (SI Methods and Table S1). They accounted for an average of 70% of the total relative abundance and included common activated sludge populations such as Comamonadaceae, Rhodobacteraceae, Saprospiraceae, Nitrospiraceae, Nitrosomonadaceae, and Rhodocyclaceae (Figure 3). As revealed by RDA, Comamonadaceae, Rhodobacteraceae and Saprospiraceae were influenced by organic carbon (i.e.,

BOD) and dissolved oxygen (SI Figure S4). In comparison, Nitrospiraceae, Nitrosomonadaceae and Rhodocyclaceae showed high sensitivity to ammonia and temperature (SI Figure S4). Members of these families were frequently found abundant in activated sludge communities and were known to contribute to organic carbon degradation, ammonia/nitrite oxidation, and nitrate reduction (Dueholm et al. 2022). The core also included less well characterized populations (e.g., Bradyrhizobiaceae, Sphingomonadaceae, and Intrasporangiaceae), as well as unknown taxa due to the lack of their representative sequences in the *Silva* database (updated in June 2018).

3.3 Dynamics of the core populations

Population dynamics was expected to aggregate the periodic fluctuation in microbial growth and reflect the stability of the populations (SI Methods). As shown in Figure 4, the majority of the core populations exhibited high stability with the dynamics ranging between ±0.05 d^{-1}. Comamonadaceae, Rhodobacteraceae and Rhodocyclaceae were particularly stable, as evidenced by the close-to-zero dynamics (-0.015 d^{-1}) and narrow 95% confidence intervals (< 0.015 d^{-1}) across all 936 transformed samples. These indicated their resilience to environment perturbation and explained their ubiquity in activated sludge systems. Fluctuation was observed for Saprospiraceae (confidence interval >0.02), suggesting their weaker resilience to environmental perturbation. For nitrifying bacteria, Nitrospiraceae appeared to be more stable (-0.03 d^{-1}, 95% confidence interval 0.03) than Nitrosomonadaceae (-0.07 d^{-1}, 95% confidence interval 0.04) (t-test, $p < 0.05$).

RDA was performed to understand the relationship between population dynamics and environmental perturbation (SI Figure S5). Most of the core populations were found in the center

of the RDA plot, meaning that they were not influenced by the change in any particular environmental factor. For example, Comamonadaceae, Rhodobacteraceae and Rhodocyclaceae were very close to the origin, highlighting their resilience to perturbation. Saprospiraceae was slightly affected by the changes in sludge retention time and effluent ammonia, consistent with its dynamics (Figure 4). Similarly, Nitrospiraceae was closer to the origin, whereas Nitrosomonadaceae was more positively associated with sludge retention time and negatively associated with temperature. Populations with the high dynamics (e.g., unknown SC_I_84, OM60, Phyllobacteriaceae) were found in the outermost areas in the RDA plot and were strongly affected by environmental perturbation.

3.4 Inference of functions and interactions

Bayesian networks were built through data-driven structure learning to identify associations between growth kinetics and population dynamics. The subnetwork centered on heterotrophic growth kinetics consisted of typical heterotrophs such as Comamonadaceae, Rhodobacteraceae, and Saprospiraceae (Figure 5A). Comamonadaceae were directed to the Monod constant of aerobic heterotrophs (K_{oh}) with an arc strength of 0.61 (SI Table S2). According to the MiDAS, the dominant genera in this family were all identified as aerobic heterotrophs and could grow on complex substrates such as sugars and proteins (SI Figure S7). Rhodobacteraceae served as an alternative aerobic heterotroph, as evidenced by its strong dependence on the maximum specific growth rate of heterotrophs (u_h, arc strength 0.75). The genera of Rhodobacteraceae were abundant and could also degrade different substrates (SI Figure S7). Saprospiraceae was associated with the biomass yield of aerobic heterotrophs (Y_h) with an arc strength of 0.69 (SI Table S2). This family was dominated by an unknown genus (average relative abundance 4.9%) that could not be mapped

309 to the MiDAS. One of its potential functions is to prey on live cells and/or convert dead cells into
310 simpler organic compounds (Seguel Suazo et al. 2024) .
311
312 Topological data analysis was performed to further learn the interactions within populations
313 related to heterotrophic growth kinetics (Figure 5B and SI Figure S9). The results showed a time-
314 dependent relationship between historical abundance and population dynamics. Specifically, the
315 average number of arcs between the two categories increased from less than three at short time
316 spans ($\Delta t < 10$ days) to more than seven at long time spans ($\Delta t > 11$ days). Bradyrhizobiaceae was
317 identified as a keystone population as its historical abundance was linked to the dynamics of
318 multiple populations including Pirellulaceae at different time spans. Bradyrhizobiaceae and
319 Pirellulaceae have been reported to coexist in various environments including marine sediments
320 and soils (Liu et al. 2018, Walters et al. 2018). According to the MiDAS, the dominant genus of
321 Bradyrhizobiaceae, *Bradyrhizobium*, can perform denitrification (SI Figure S7). Within the
322 interactions learned from historical abundance, consistent patterns were observed. For example,
323 Rhodobacteraceae was associated with Sphingomonadaceae across all eight time spans, and their
324 interactions at short time spans were among the strongest, as revealed by the adjacency matrices
325 (i.e., the inverse of the Euclidean distance, SI Table S3). Similarly, the interaction between
326 Saprospiraceae and Cryomorphaceae were consistently observed in the top ten adjacency list.
327
328 In the subnetwork centered on autotrophic growth kinetics (Figure S6), Nitrospiraceae and
329 Nitrosomonadaceae were found to point to the decay and growth rates of autotrophs with an arc
330 strength of 0.64 and 0.59, respectively (SI Table S4). The genus *Nitrospira* of Nitrospiraceae is
331 known to be involved in ammonia and nitrite oxidation (SI Figure S8). In the family

Nitrosomonadaceae, an unknown genus contributed 99% of the abundance but could not be found in the MiDAS. This genus is likely an ammonium/nitrite oxidizer but warrants further investigation. Rhodocyclaceae as the most abundant population was associated with the Monod constant of ammonia oxidation (*Kn*) with an arc strength of 0.60 (SI Table S4). According to the MiDAS, the genera in this family can suppress ammonia- and nitrite-oxidizing bacteria (e.g., *Candidatus* Accumulibacter, SI Figure S8). Therefore, the presence of Rhodocyclaceae is expected to exert negative effects on nitrification. In addition, families such as Sphingomonadaceae, Intrasporangiaceae, and Caldilineaceae were associated with both the microbial kinetic parameters of heterotrophs (Figure 5A) and autotroph (Figure S6). Similar to Rhodocyclaceae, these populations are listed as aerobic heterotrophs in the MiDAS (SI Figure S8) and potentially inhibit nitrification.

Topological data analysis of the autotroph-related populations revealed a clear independence between historical abundance and population dynamics (SI Figure S10). The interactions within the autotrophic guild were also simpler than those between the heterotroph-related populations as evidenced by the greater sparsity of the arcs. Moreover, the interactions appeared to be more consistent across time spans. For example, when interactions were inferred from historical abundance, Nitrospiraceae was always associated with Rhodocyclaceae and an unknown family of the order Chromatiales, while Rhodocyclaceae interacted consistently with Moraxellaceae and OM60. In terms of the dynamics-inferred associations, Nitrosomonadaceae and Caldilineaceae actively interacted at all time spans with the adjacency among the highest at most time spans (SI Table S5). Meanwhile, Nitrosomonadaceae, Intrasporangiaceae, and OM60 formed a stable

relationship throughout the eight time spans (SI Figure S10). Their adjacencies were ranked in the top ten at time spans 7 and 9-14 (SI Table S5).

3.5 Validation of the inferences

The inferred functions and interactions were indirectly validated by predicting absolute abundance and community structure, with the expectation that adequate inferences should lead to accurate predictions. After optimization using a grid search method (SI Methods), ANNs were tested on 134 transformed samples to predict absolute abundance and community structure. A total of 5628 predictions (134 samples × 42 populations) were obtained (SI Figure S11). The fitted slope of 0.51 between observed and predicted abundances indicated that the model tended to overestimate population abundance. This was because a large number of populations were absent (i.e., zero abundance) in the test set but were predicted to be present with certain abundances. As a result, the R^2 between observed and predicted abundances was only 0.36. When the absent populations were excluded from the test set, the prediction was significantly more accurate with the slope and R^2 increasing to 0.75 and 0.56, respectively. In comparison, the ANNs trained without microbial kinetic parameters as inputs yielded more server overestimation (slope 0.47) and less accurate predictions (R^2 0.50) even when the absent populations were excluded. The difficulty in predicting the abundance of individual populations, especially those with low to zero abundances, could be attributed to the incompleteness of the dataset (e.g., lack of environmental factors such as nitrate and phosphate concentrations and their perturbation), as well as the inherently stochastic processes of microbial community assembly (Langille et al. 2013, Zhou and Ning 2017).

The predicted and observed community structures were compared using Bray-Curtis similarity. The communities predicted with microbial kinetic parameters as inputs shared a similarity of 0.70 ± 0.19 with the observed communities (Figure 6). This was comparable to the results obtained in previous studies (e.g., 0.65 for order-level communities in acid mine drainage and 0.72 for genus-level communities in bioelectrochemical systems) (Cheng et al. 2021a, Langille et al. 2013). In comparison, prediction without microbial kinetic parameters yielded a lower similarity of 0.66 ± 0.18 (t-test, $p < 0.05$). Given that over half of the raw samples were from Hong Kong, additional predictions were performed for the Hong Kong samples only. The similarity increased significantly to 0.84 ± 0.07 (with parameters) and 0.70 ± 0.20 (without parameters). The results suggest slight overfitting caused by inherent bias in the training set, as well as the benefit of including microbial kinetic parameters to mitigate such bias. To further understand the potential of the model, ANNs were trained and tested with relative abundance as input. Compared to those trained with absolute abundance, the similarity was improved to 0.74 ± 0.17 (SI Figure S12, t-test, $p < 0.05$). The improvement could be due to the elimination of the error introduced by biomass concentration. The control model (built following a previous study (Larsen et al. 2012b)) and the null model (built using average abundances) produced much less accurate predictions with a similarity of 0.60 ± 0.20 and 0.53 ± 0.19, respectively (t-test, $p < 0.05$).

4. Discussion

There is a growing and extensive interest in building a predictive understanding of microbial ecology by modeling microbial communities (Ghannam and Techtmann 2021, Kumar et al. 2019, Larsen et al. 2012a, Lopatkin and Collins 2020). Conventional mechanistic models can predict the functions and interactions of guilds but not the actual microbial populations observed in the

399 community due to the challenges of simulating the growth kinetics of individual populations
400 (Bouskill et al. 2012, Song et al. 2014). Emerging data-driven models can infer the functions and
401 interactions within the community, but the robustness of the inference can be compromised by
402 limited data availability and lack of temporal variation (Metcalf et al. 2016, Yao et al. 2022).
403
404 The present study aimed to address these limitations by integrating mechanistic and data-driven
405 modeling based on the intrinsic connection between microbial growth kinetics and microbial
406 population dynamics (Eq. 5). To implement this novel modeling approach, a comprehensive
407 literature review was conducted, and 466 raw samples from four full-scale activated sludge
408 systems in Hong Kong, Beijing, and Aalborg were retrieved (Table 1) (Jiang et al. 2018, Peces et
409 al. 2022, Sun et al. 2021). The environmental factors and microbial communities were highly
410 diverse (Figures 2 and 3), demonstrating the representativeness of the selected samples for
411 implementing the modeling approach. Using the data transformation method developed in this
412 study (Figure 1), samples from different activated sludge systems were transformed into the same
413 structure and combined, and the size of the training set was increased by three times.
414
415 In addition to data augmentation, the transformation enabled quantification of population
416 dynamics and environmental perturbation. The overall dynamics of a core population that is
417 universally abundant in different environments is expected to be close to zero (SI Methods).
418 Consistent with this assumption, 34 of the 42 core populations exhibited low dynamics within
419 $\pm 0.05\ d^{-1}$ (Figure 4) and were resilient to environmental perturbation (SI Figure S4). Although the
420 calculation of population dynamics differs from a previously developed mass balance-based
421 method that calculates net growth rates (Kim et al. 2020, Peces et al. 2022, Saunders et al. 2016),

422 the two methods are complementary in terms of identifying consistently active populations. The
423 remaining eight populations were highly dynamic and strongly affected by environmental
424 perturbation (SI Figure S6). Notably, unknown SC_I_84 with the most positive dynamics showed
425 the strongest positive association with temperature perturbation and negative association with
426 sludge retention time perturbation, whereas the populations with the most negative dynamics
427 (Nitrosomonadaceae, Phyllobacteriaceae, and OM60) showed the exact opposite. The results
428 imply the replacement of the latter by the former and highlight the potential of the data
429 transformation method for holistic analysis of microbiomes from different systems, which can lead
430 to new insights into microbial ecology (Wu et al. 2019).
431
432 Learned from the transformed data, the Bayesian network and topological associations served as
433 exploratory tools for identifying keystone populations and inferring their roles within the
434 community. For example, the family Intrasporangiaceae was found to be linked to the parameters
435 of both heterotrophs and autotrophs (SI Figure S5). Dominated by the putative phosphate-
436 accumulating organism *Tetrasphaera* (Kristiansen et al. 2013), Intrasporangiaceae was
437 consistently associated with Nitrosomonadaceae in topological data analysis (SI Figure S10).
438 Together with their dynamics and the MiDAS, it is reasonable to speculate that Intrasporangiaceae
439 and Nitrosomonadaceae have an agonistic relationship. Such inferences can be combined with
440 downstream ANN prediction of population abundance to forecast system performance. For
441 example, high abundance of suppressors such as Intrasporangiaceae and low abundance of
442 nitrifying bacteria would result in excess nitrates and nitrites, which in turn could inhibit various
443 microorganisms depending on pH, temperature, and influent concentrations (Zhou et al. 2011).
444 This inhibition could lead to significant shifts in microbial community structure and consequently

impair system performance. Predicted dominance and washout of indicator populations can alert treatment plant personnel and enable predictive control to prevent system failure.

Despite their potential to provide predictive insights into microbial ecology and engineering applications, the data transformation method and modeling approach developed in this study remain to be improved with more data, particularly time series of population abundance collected at high sampling frequencies (e.g., every 1-3 days). Several global surveys have been conducted to understand the microbiomes in activated sludge systems and anaerobic digesters (Mei et al. 2017, Wu et al. 2019), but the data are not reported as time series and cannot be incorporated into this study. The modeling approach also needs to be examined with natural ecosystems, where the growth kinetics of guilds have been simulated mechanistically (Bouskill et al. 2012, Jin and Roden 2011, Wieder et al. 2015, Wieder et al. 2013). Finally, the modeling approach can be improved by proper selection and integration of probabilistic and machine learning methods. In this study, Bayesian networks have two functions: to classify populations based on probabilistic dependencies with guild kinetic parameters, and to reduce the computational cost of downstream topological data analysis and ANN training. Topological data analysis has emerged as a powerful method in machine learning/deep learning for extracting complementary information about the observed objects (Baas et al. 2020, Wasserman 2018, Zomorodian 2012). It is applied for the first time to capture hidden spatial dependencies and genuine associations between microbial populations, and its application in microbial ecology remains to be refined.

Acknowledgment

This work was supported by the U.S. Department of Agriculture [Award No. 2020-67019-31027].

Reference

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., Abou Elwafa, A. and Kurdi, H. (2021) Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Applied Sciences 11(2), 796.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L. and Ridella, S. (2012) The 'K' in K-fold Cross Validation. ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 25-27.

Ansari, A.F., Reddy, Y.B., Raut, J. and Dixit, N.M. (2021) An efficient and scalable top-down method for predicting structures of microbial communities. Nature Computational Science 1(9), 619-628.

Baas, N.A., Carlsson, G.E., Quick, G., Szymik, M. and Thaule, M. (2020) Topological Data Analysis, Springer.

Batstone, D.J., Keller, J., Angelidaki, I., Kalyuzhnyi, S., Pavlostathis, S., Rozzi, A., Sanders, W., Siegrist, H. and Vavilin, V. (2002) The IWA anaerobic digestion model no 1 (ADM1). Water Science and Technology 45(10), 65-73.

Bishop, C.M. (1995) Neural Networks for Pattern Recognition, Oxford University Press Inc.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J.G. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37(8), 852-857.

Bouskill, N.J., Tang, J., Riley, W.J. and Brodie, E.L. (2012) Trait-based representation of biological nitrification: model development, testing, and predicted community composition. Frontiers in Microbiology 3, 364.

506 Buttigieg, P.L. and Ramette, A. (2014) A guide to statistical analysis in microbial ecology: a
507 community-focused, living review of multivariate data analyses. FEMS Microbiology Ecology
508 90(3), 543-550.

509 Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens,
510 S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G. and Knight, R.
511 (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq
512 platforms. The ISME Journal 6(8), 1621-1624.

513 Cheng, Z., Ronen, A. and Yuan, H. (2024) Hybrid Modeling of Engineered Biological Systems
514 through Coupling Data-Driven Calibration of Kinetic Parameters with Mechanistic Prediction of
515 System Performance. ACS ES&T Water 4(3), 958-968.

516 Cheng, Z., Yao, S. and Yuan, H. (2021a) Linking population dynamics to microbial kinetics for
517 hybrid modeling of bioelectrochemical systems. Water Research 202, 117418.

518 Cheng, Z., Yao, S. and Yuan, H. (2021b) Linking Population Dynamics to Microbial Kinetics for
519 Hybrid Modeling of Bioelectrochemical Systems. Water research, 117418.

520 Coyte, K.Z., Schluter, J. and Foster, K.R. (2015) The ecology of the microbiome: networks,
521 competition, and stability. Science 350(6261), 663-666.

522 Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M.,
523 Huttenhower, C. and Langille, M.G.I. (2020) PICRUSt2 for prediction of metagenome functions.
524 Nature Biotechnology 38(6), 685-688.

525 Dueholm, M.K.D., Nierychlo, M., Andersen, K.S., Rudkjøbing, V., Knutsson, S., Arriaga, S.,
526 Bakke, R., Boon, N., Bux, F., Christensson, M., Chua, A.S.M., Curtis, T.P., Cytryn, E., Erijman,
527 L., Etchebehere, C., Fatta-Kassinos, D., Frigon, D., Garcia-Chaves, M.C., Gu, A.Z., Horn, H.,
528 Jenkins, D., Kreuzinger, N., Kumari, S., Lanham, A., Law, Y., Leiknes, T., Morgenroth, E.,
529 Muszyński, A., Petrovski, S., Pijuan, M., Pillai, S.B., Reis, M.A.M., Rong, Q., Rossetti, S.,
530 Seviour, R., Tooker, N., Vainio, P., van Loosdrecht, M., Vikraman, R., Wanner, J., Weissbrodt,
531 D., Wen, X., Zhang, T., Nielsen, P.H., Albertsen, M., Nielsen, P.H. and Mi, D.A.S.G.C. (2022)
532 MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies
533 of bacterial communities in wastewater treatment plants. Nature Communications 13(1), 1908.

534 Falkowski, P.G., Fenchel, T. and Delong, E.F. (2008) The microbial engines that drive Earth's
535 biogeochemical cycles. Science 320(5879), 1034-1039.

536 Fuhrman, J.A. (2009) Microbial community structure and its functional implications. Nature
537 459(7244), 193-199.

538 Fuhrman, J.A., Cram, J.A. and Needham, D.M. (2015) Marine microbial community dynamics
539 and their ecological interpretation. Nature Reviews Microbiology 13(3), 133-146.

540 Ghannam, R.B. and Techtmann, S.M. (2021) Machine learning applications in microbial
541 ecology, human microbiome studies, and environmental monitoring. Computational and
542 Structural Biotechnology Journal 19, 1092-1107.

543 Gómez-Acata, S., Esquivel-Ríos, I., Pérez-Sandoval, M.V., Navarro-Noya, Y., Rojas-Valdez, A.,
544 Thalasso, F., Luna-Guido, M. and Dendooven, L. (2017) Bacterial community structure within
545 an activated sludge reactor added with phenolic compounds. Applied Microbiology and
546 Biotechnology 101(8), 3405-3414.

547 Griffin, J.S. and Wells, G.F. (2016) Regional synchrony in full-scale activated sludge bioreactors
548 due to deterministic microbial community assembly. The ISME Journal 11, 500.

549 Gujer, W., Henze, M., Mino, T., Matsuo, T., Wentzel, M. and Marais, G. (1995) The activated
550 sludge model No. 2: biological phosphorus removal. Water Science and Technology 31(2), 1-11.

551 Gujer, W., Henze, M., Mino, T. and Van Loosdrecht, M. (1999) Activated sludge model no. 3.
552 Water Science and Technology 39(1), 183-193.

553 Harvey, P.H., Colwell, R.K., Silvertown, J.W. and May, R.M. (1983) Null Models in Ecology.
554 Annual Review of Ecology and Systematics 14, 189-211.

555 Heineman, G.T., Pollice, G. and Selkow, S. (2008) Algorithms in a Nutshell, O'Reilly Media.

556 Henze, M., Grady Jr, C.L., Gujer, W., Marais, G. and Matsuo, T. (1987) A general model for
557 single-sludge wastewater treatment systems. Water research 21(5), 505-515.

558 Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M.C., Marais, G.v.R. and Van Loosdrecht,
559 M.C. (1999) Activated sludge model no. 2d, ASM2d. Water Science and Technology 39(1), 165-
560 182.

561 Henze, M., Gujer, W., Mino, T. and van Loosdrecht, M.C. (2000) Activated sludge models
562 ASM1, ASM2, ASM2d and ASM3, IWA publishing.

563 Jiang, X.-T., Ye, L., Ju, F., Wang, Y.-L. and Zhang, T. (2018) Toward an intensive longitudinal
564 understanding of activated sludge bacterial assembly and dynamics. Environmental Science &
565 Technology 52(15), 8224-8232.

566 Jin, Q. and Roden, E.E. (2011) Microbial physiology-based model of ethanol metabolism in
567 subsurface sediments. Journal of contaminant hydrology 125(1-4), 1-12.

568 Ju, F. and Zhang, T. (2015a) Bacterial assembly and temporal dynamics in activated sludge of a
569 full-scale municipal wastewater treatment plant. The ISME Journal 9(3), 683-695.

570 Ju, F. and Zhang, T. (2015b) Experimental design and bioinformatics analysis for the application
571 of metagenomics in environmental sciences and biotechnology. Environmental Science &
572 Technology 49(21), 12628-12640.

573 Kim, J., Mei, R., Wilson, F.P., Yuan, H., Bocher, B.T.W. and Liu, W.-T. (2020) Ecogenomics-
574 Based Mass Balance Model Reveals the Effects of Fermentation Conditions on Microbial
575 Activity. Frontiers in Microbiology 11(3115), 595036.

576 Kovárová-Kovar, K. and Egli, T. (1998) Growth kinetics of suspended microbial cells: from
577 single-substrate-controlled growth to mixed-substrate kinetics. Microbiology and Molecular
578 Biology Reviews 62(3), 646-666.

579 Kristiansen, R., Nguyen, H.T., Saunders, A.M., Nielsen, J.L., Wimmer, R., Le, V.Q., McIlroy,
580 S.J., Petrovski, S., Seviour, R.J., Calteau, A., Nielsen, K.L. and Nielsen, P.H. (2013) A metabolic
581 model for members of the genus Tetrasphaera involved in enhanced biological phosphorus
582 removal. Isme j 7(3), 543-554.

583 Kuang, J., Huang, L., He, Z., Chen, L., Hua, Z., Jia, P., Li, S., Liu, J., Li, J., Zhou, J. and Shu, W.
584 (2016) Predicting taxonomic and functional structure of microbial communities in acid mine
585 drainage. The ISME Journal 10(6), 1527-1539.

586 Kumar, M., Ji, B., Zengler, K. and Nielsen, J. (2019) Modelling approaches for studying the
587 microbiome. Nature Microbiology 4(8), 1253-1267.

588 Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A.,
589 Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G. and Huttenhower,
590 C. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene
591 sequences. Nature Biotechnology 31(9), 814-821.

592 Larsen, P., Dai, Y. and Collart, F.R. (2015) Artificial neural networks, pp. 33-43, Springer.

593 Larsen, P., Hamada, Y. and Gilbert, J. (2012a) Modeling microbial communities: Current,
594 developing, and future technologies for predicting microbial community interaction. Journal of
595 Biotechnology 160(1), 17-24.

596 Larsen, P.E., Field, D. and Gilbert, J.A. (2012b) Predicting bacterial community assemblages
597 using an artificial neural network approach. Nature Methods 9(6), 621-625.

598 Lax, S., Smith, D.P., Hampton-Marcell, J., Owens, S.M., Handley, K.M., Scott, N.M., Gibbons,
599 S.M., Larsen, P., Shogan, B.D., Weiss, S., Metcalf, J.L., Ursell, L.K., Vázquez-Baeza, Y., Van
600 Treuren, W., Hasan, N.A., Gibson, M.K., Colwell, R., Dantas, G., Knight, R. and Gilbert, J.A.
601 (2014) Longitudinal analysis of microbial interaction between humans and the indoor
602 environment. Science 345(6200), 1048-1052.

603 Lesnik, K.L., Cai, W. and Liu, H. (2020) Microbial Community Predicts Functional Stability of
604 Microbial Fuel Cells. Environmental Science & Technology 54(1), 427-436.

605 Lesnik, K.L. and Liu, H. (2017) Predicting Microbial Fuel Cell Biofilm Communities and
606 Bioreactor Performance using Artificial Neural Networks. Environmental Science & Technology
607 51(18), 10881-10892.

608 Ling, F., Hwang, C., LeChevallier, M.W., Andersen, G.L. and Liu, W.-T. (2016) Core-satellite
609 populations and seasonality of water meter biofilms in a metropolitan drinking water distribution
610 system. The ISME Journal 10(3), 582-595.

611 Liu, D., Yang, Y., An, S., Wang, H. and Wang, Y. (2018) The Biogeographical Distribution of
612 Soil Bacterial Communities in the Loess Plateau as Revealed by High-Throughput Sequencing.
613 Front Microbiol 9, 2456.

614 Liu, W., Peng, Y., Ma, B., Ma, L., Jia, F. and Li, X. (2017) Dynamics of microbial activities and
615 community structures in activated sludge under aerobic starvation. Bioresource Technology 244,
616 588-596.

617 Lopatkin, A.J. and Collins, J.J. (2020) Predictive biology: modelling, understanding and
618 harnessing microbial complexity. Nature Reviews Microbiology 18(9), 507-520.

619 Mansfeldt, C., Achermann, S., Men, Y., Walser, J.-C., Villez, K., Joss, A., Johnson, D.R. and
620 Fenner, K. (2019) Microbial residence time is a controlling parameter of the taxonomic
621 composition and functional profile of microbial communities. The ISME Journal.

622 McIlroy, S.J., Kirkegaard, R.H., McIlroy, B., Nierychlo, M., Kristensen, J.M., Karst, S.M.,
623 Albertsen, M. and Nielsen, P.H. (2017) MiDAS 2.0: an ecosystem-specific taxonomy and online
624 database for the organisms of wastewater treatment systems expanded for anaerobic digester
625 groups. Database : the journal of biological databases and curation 2017(1), bax016.

626 Mei, R., Nobu, M.K., Narihiro, T., Kuroda, K., Muñoz Sierra, J., Wu, Z., Ye, L., Lee, P.K.H.,
627 Lee, P.-H., van Lier, J.B., McInerney, M.J., Kamagata, Y. and Liu, W.-T. (2017) Operation-
628 driven heterogeneity and overlooked feed-associated populations in global anaerobic digester
629 microbiome. Water research 124, 77-84.

630 Metcalf, J.L., Xu, Z.Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E.R., Song, S.J., Amir, A.,
631 Larsen, P., Sangwan, N., Haarmann, D., Humphrey, G.C., Ackermann, G., Thompson, L.R.,
632 Lauber, C., Bibat, A., Nicholas, C., Gebert, M.J., Petrosino, J.F., Reed, S.C., Gilbert, J.A.,
633 Lynne, A.M., Bucheli, S.R., Carter, D.O. and Knight, R. (2016) Microbial community assembly
634 and metabolic function during mammalian corpse decomposition. Science 351(6269), 158-162.

635 Monod, J. (1942) Recherches sur la croissance des cultures bacteriennes.

636 Monod, J. (1949) The growth of bacterial cultures. Annual Review of Microbiology 3(1), 371-
637 394.

638 Mowbray, M., Savage, T., Wu, C., Song, Z., Cho, B.A., Del Rio-Chanona, E.A. and Zhang, D.
639 (2021) Machine learning for biochemical engineering: A review. Biochemical Engineering
640 Journal 172, 108054.

641 Nierychlo, M., Andersen, K.S., Xu, Y., Green, N., Jiang, C., Albertsen, M., Dueholm, M.S. and
642 Nielsen, P.H. (2020) MiDAS 3: An ecosystem-specific reference database, taxonomy and
643 knowledge platform for activated sludge and anaerobic digesters reveals species-level
644 microbiome composition of activated sludge. Water research, 115955.

645 Ouyang, F., Ji, M., Zhai, H., Dong, Z. and Ye, L. (2016) Dynamics of the diversity and structure
646 of the overall and nitrifying microbial community in activated sludge along gradient copper
647 exposures. Applied Microbiology and Biotechnology 100(15), 6881-6892.

Peces, M., Dottorini, G., Nierychlo, M., Andersen, K.S., Dueholm, M.K.D. and Nielsen, P.H. (2022) Microbial communities across activated sludge plants show recurring species-level seasonal patterns. ISME Communications 2(1), 18.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(D1), D590-D596.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. Nature Biotechnology 35, 833.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K. and Gies, E.A. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459), 431-437.

Rittmann, B.E. and McCarty, P.L. (2012) Environmental biotechnology: principles and applications, Tata McGraw-Hill Education.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F. and Dormann, C.F. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8), 913-929.

Ruan, Q., Dutta, D., Schwalbach, M.S., Steele, J.A., Fuhrman, J.A. and Sun, F. (2006) Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics 22(20), 2532-2538.

Saunders, A.M., Albertsen, M., Vollertsen, J. and Nielsen, P.H. (2016) The activated sludge ecosystem contains a core community of abundant organisms. The ISME Journal 10(1), 11-20.

Scutari, M., Howell, P., Balding, D.J. and Mackay, I. (2014) Multiple Quantitative Trait Analysis Using Bayesian Networks. Genetics 198(1), 129-137.

Seguel Suazo, K., Dobbeleers, T. and Dries, J. (2024) Bacterial community and filamentous population of industrial wastewater treatment plants in Belgium. Applied Microbiology and Biotechnology 108(1), 43.

Song, H.-S., Cannon, W.R., Beliaev, A.S. and Konopka, A. (2014) Mathematical Modeling of Microbial Community Dynamics: A Methodological Review. Processes 2(4), 711-752.

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B. and Sadowsky, M.J. (2014) Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. Frontiers in Microbiology 5(524).

Stams, A.J. and Plugge, C.M. (2009) Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nature Reviews Microbiology 7(8), 568-577.

682 Stewart, R.D., Auffret, M.D., Warr, A., Wiser, A.H., Press, M.O., Langford, K.W., Liachko, I.,
683 Snelling, T.J., Dewhurst, R.J., Walker, A.W., Roehe, R. and Watson, M. (2018) Assembly of 913
684 microbial genomes from metagenomic sequencing of the cow rumen. Nature Communications
685 9(1), 870.

686 Sun, C., Zhang, B., Ning, D., Zhang, Y., Dai, T., Wu, L., Li, T., Liu, W., Zhou, J. and Wen, X.
687 (2021) Seasonal dynamics of the microbial community in two full-scale wastewater treatment
688 plants: diversity, composition, phylogenetic group based assembly and co-occurrence pattern.
689 Water research 200, 117295.

690 Torsvik, V. and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to
691 ecosystems. Current Opinion in Microbiology 5(3), 240-245.

692 Uusitalo, L. (2007) Advantages and challenges of Bayesian networks in environmental
693 modelling. Ecological Modelling 203(3), 312-318.

694 Vanwonterghem, I., Jensen, P.D., Ho, D.P., Batstone, D.J. and Tyson, G.W. (2014) Linking
695 microbial community structure, interactions and function in anaerobic digesters using new
696 molecular techniques. Current Opinion in Biotechnology 27, 55-64.

697 Veshareh, M.J. and Nick, H.M. (2021) A novel relationship for the maximum specific growth
698 rate of a microbial guild. FEMS Microbiology Letters 368(12).

699 Vuono, D.C., Benecke, J., Henkel, J., Navidi, W.C., Cath, T.Y., Munakata-Marr, J., Spear, J.R.
700 and Drewes, J.E. (2014) Disturbance and temporal partitioning of the activated sludge
701 metacommunity. The ISME Journal 9(2), 425.

702 Walters, W.A., Jin, Z., Youngblut, N., Wallace, J.G., Sutter, J., Zhang, W., González-Peña, A.,
703 Peiffer, J., Koren, O. and Shi, Q. (2018) Large-scale replicated field study of maize rhizosphere
704 identifies heritable microbes. Proceedings of the National Academy of Sciences 115(28), 7368-
705 7373.

706 Wasserman, L. (2018) Topological data analysis. Annual Review of Statistics and Its
707 Application 5, 501-532.

708 Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu,
709 Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou,
710 J. and Knight, R. (2016) Correlation detection strategies in microbial data sets vary widely in
711 sensitivity and precision. The ISME Journal 10(7), 1669-1681.

712 Weissman, J.L., Hou, S. and Fuhrman, J.A. (2021) Estimating maximal microbial growth rates
713 from cultures, metagenomes, and single cells via codon usage patterns. Proceedings of the
714 National Academy of Sciences 118(12), e2016810118.

715 Widder, S., Allen, R.J., Pfeiffer, T., Curtis, T.P., Wiuf, C., Sloan, W.T., Cordero, O.X., Brown,
716 S.P., Momeni, B., Shou, W., Kettle, H., Flint, H.J., Haas, A.F., Laroche, B., Kreft, J.-U., Rainey,
717 P.B., Freilich, S., Schuster, S., Milferstedt, K., van der Meer, J.R., Grokopf, T., Huisman, J.,
718 Free, A., Picioreanu, C., Quince, C., Klapper, I., Labarthe, S., Smets, B.F., Wang, H., Isaac

Newton Institute, F. and Soyer, O.S. (2016) Challenges in microbial ecology: building predictive understanding of community function and dynamics. The ISME Journal 10(11), 2557-2568.

Wieder, W.R., Allison, S.D., Davidson, E.A., Georgiou, K., Hararuk, O., He, Y., Hopkins, F., Luo, Y., Smith, M.J., Sulman, B., Todd-Brown, K., Wang, Y.-P., Xia, J. and Xu, X. (2015) Explicitly representing soil microbial processes in Earth system models. Global Biogeochemical Cycles 29(10), 1782-1800.

Wieder, W.R., Bonan, G.B. and Allison, S.D. (2013) Global soil carbon projections are improved by modelling microbial processes. Nature climate change 3(10), 909-912.

Wright, S.F.a.F.G.a.M.N. (2019) neuralnet: Training of Neural Networks.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., Zhang, Q., Brown, M.R., Li, Z., Van Nostrand, J.D., Ling, F., Xiao, N., Zhang, Y., Vierheilig, J., Wells, G.F., Yang, Y., Deng, Y., Tu, Q., Wang, A., Acevedo, D., Agullo-Barcelo, M., Alvarez, P.J.J., Alvarez-Cohen, L., Andersen, G.L., de Araujo, J.C., Boehnke, K.F., Bond, P., Bott, C.B., Bovio, P., Brewster, R.K., Bux, F., Cabezas, A., Cabrol, L., Chen, S., Criddle, C.S., Deng, Y., Etchebehere, C., Ford, A., Frigon, D., Sanabria, J., Griffin, J.S., Gu, A.Z., Habagil, M., Hale, L., Hardeman, S.D., Harmon, M., Horn, H., Hu, Z., Jauffur, S., Johnson, D.R., Keller, J., Keucken, A., Kumari, S., Leal, C.D., Lebrun, L.A., Lee, J., Lee, M., Lee, Z.M.P., Li, Y., Li, Z., Li, M., Li, X., Ling, F., Liu, Y., Luthy, R.G., Mendonça-Hagler, L.C., de Menezes, F.G.R., Meyers, A.J., Mohebbi, A., Nielsen, P.H., Ning, D., Oehmen, A., Palmer, A., Parameswaran, P., Park, J., Patsch, D., Reginatto, V., de los Reyes, F.L., Rittmann, B.E., Noyola, A., Rossetti, S., Shan, X., Sidhu, J., Sloan, W.T., Smith, K., de Sousa, O.V., Stahl, D.A., Stephens, K., Tian, R., Tiedje, J.M., Tooker, N.B., Tu, Q., Van Nostrand, J.D., De los Cobos Vasconcelos, D., Vierheilig, J., Wagner, M., Wakelin, S., Wang, A., Wang, B., Weaver, J.E., Wells, G.F., West, S., Wilmes, P., Woo, S.-G., Wu, L., Wu, J.-H., Wu, L., Xi, C., Xiao, N., Xu, M., Yan, T., Yang, Y., Yang, M., Young, M., Yue, H., Zhang, B., Zhang, P., Zhang, Q., Zhang, Y., Zhang, T., Zhang, Q., Zhang, W., Zhang, Y., Zhou, H., Zhou, J., Wen, X., Curtis, T.P., He, Q., He, Z., Brown, M.R., Zhang, T., He, Z., Keller, J., Nielsen, P.H., Alvarez, P.J.J., Criddle, C.S., Wagner, M., Tiedje, J.M., He, Q., Curtis, T.P., Stahl, D.A., Alvarez-Cohen, L., Rittmann, B.E., Wen, X., Zhou, J. and Global Water Microbiome, C. (2019) Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nature Microbiology 4(7), 1183-1195.

Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., Fuhrman, J.A. and Sun, F. (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. BMC Systems Biology 5(2), S15.

Yao, S., Zhang, C. and Yuan, H. (2022) Emerging investigator series: modeling of wastewater treatment bioprocesses: current development and future opportunities. Environmental Science: Water Research & Technology 8(2), 208-225.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N. and Anokhin, A.P. (2012) Human gut microbiome viewed across age and geography. Nature 486(7402), 222-227.

Yuan, H., Mei, R., Liao, J. and Liu, W.-T. (2019a) Nexus of Stochastic and Deterministic Processes on Microbial Community Assembly in Biological Systems. Frontiers in Microbiology 10(1536), 1536.

Yuan, H., Mei, R., Liao, J. and Liu, W.-T. (2019b) Nexus of Stochastic and Deterministic Processes on Microbial Community Assembly in Biological Systems. 10.

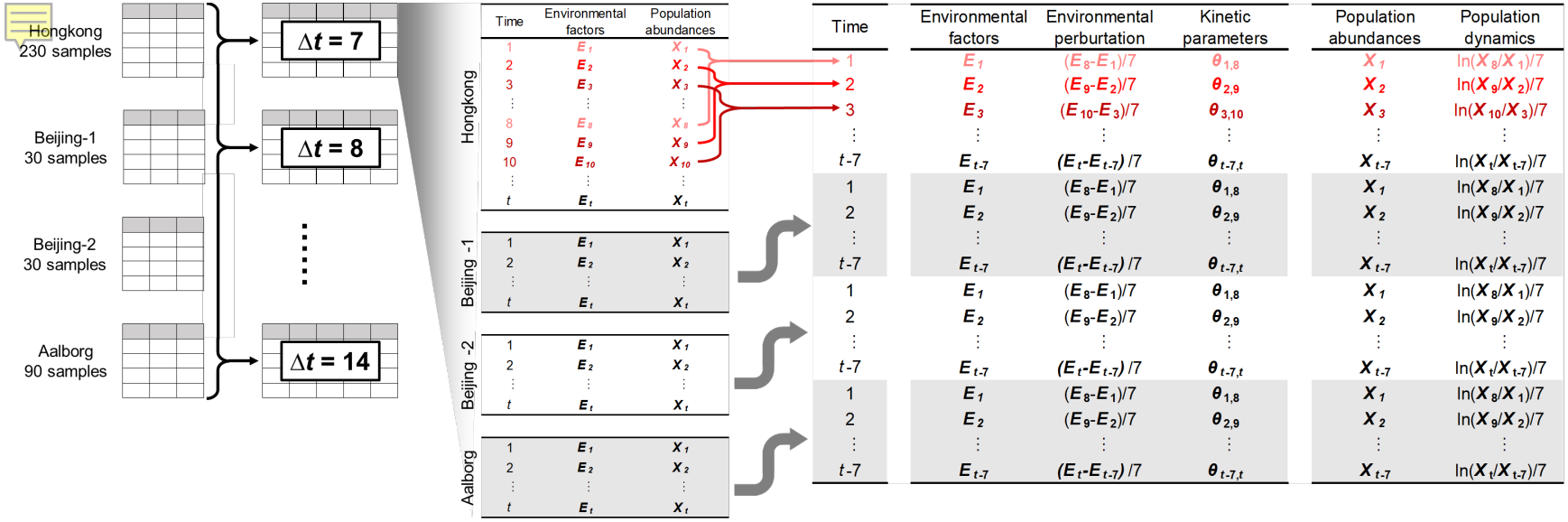
Yuan, H., Sun, S., Abu-Reesh, I.M., Badgley, B.D. and He, Z. (2017) Unravelling and Reconstructing the Nexus of Salinity, Electricity, and Microbial Ecology for Bioelectrochemical Desalination. Environmental Science & Technology 51(21), 12672-12682.

Zhang, Y., Jiang, W.-L., Qin, Y., Wang, G.-X., Xu, R.-X. and Xie, B. (2017) Dynamic changes of bacterial community in activated sludge with pressurized aeration in a sequencing batch reactor. Water Science and Technology 75(11), 2639-2648.

Zhou, J. and Ning, D. (2017) Stochastic Community Assembly: Does It Matter in Microbial Ecology? Microbiology and Molecular Biology Reviews 81(4), 10.1128/mmbr.00002-00017.

Zhou, Y., Oehmen, A., Lim, M., Vadivelu, V. and Ng, W.J. (2011) The role of nitrite and free nitrous acid (FNA) in wastewater treatment plants. Water Research 45(15), 4672-4682.

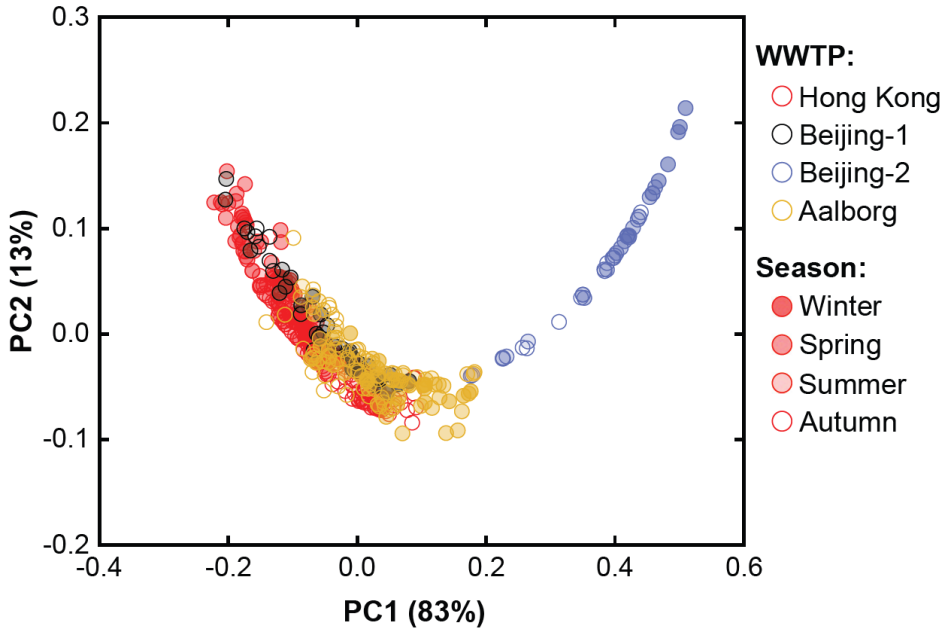
Zomorodian, A. (2012) Topological data analysis. Advances in applied and computational topology 70, 1-39.



776

777 Figure 1. Schematic of applying the data transformation method to the 380 samples in the training set.

778



779

780 Figure 2. Bray-Curis distance-based PCoA of the environmental factors of the four activated sludge systems.

781

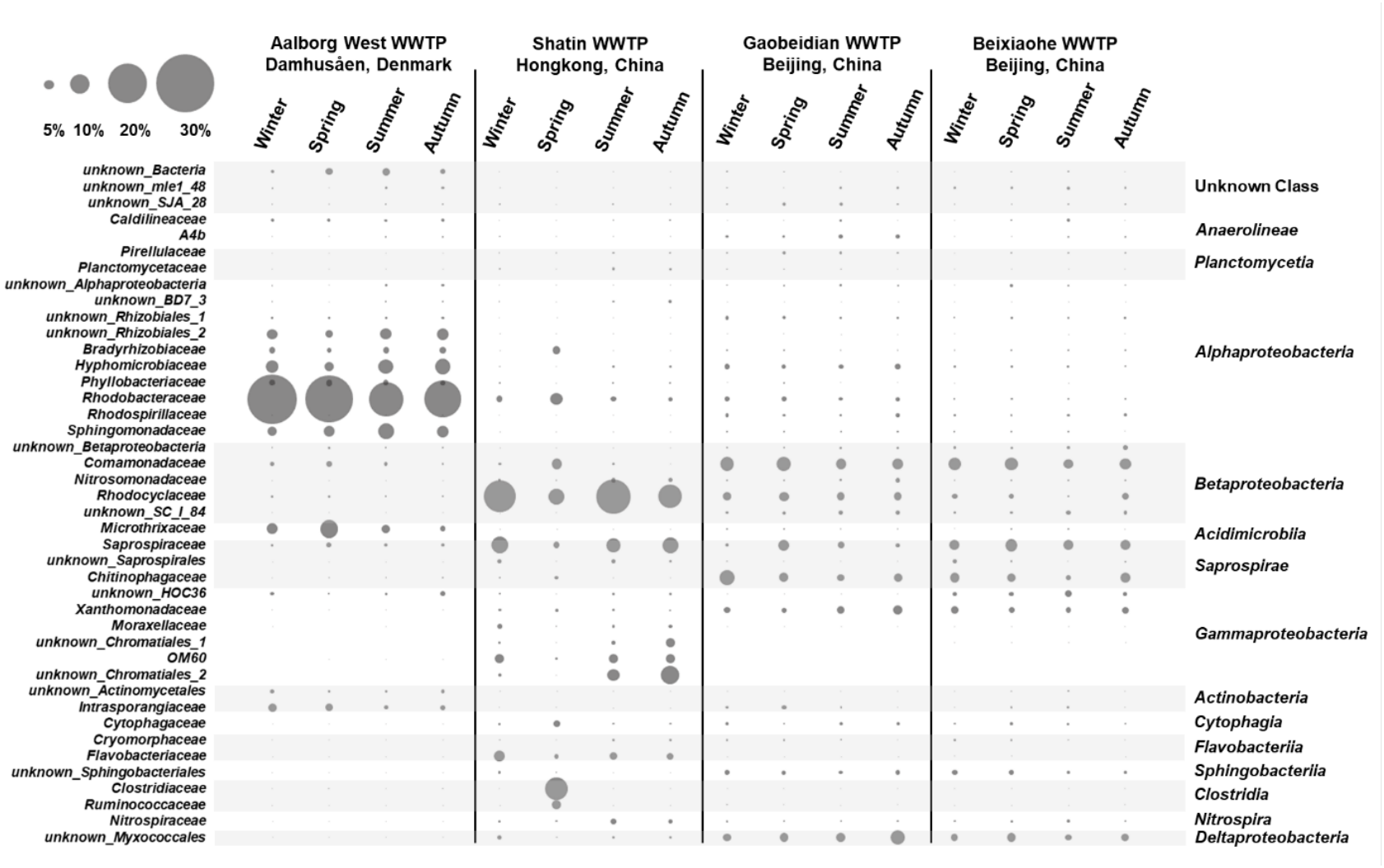
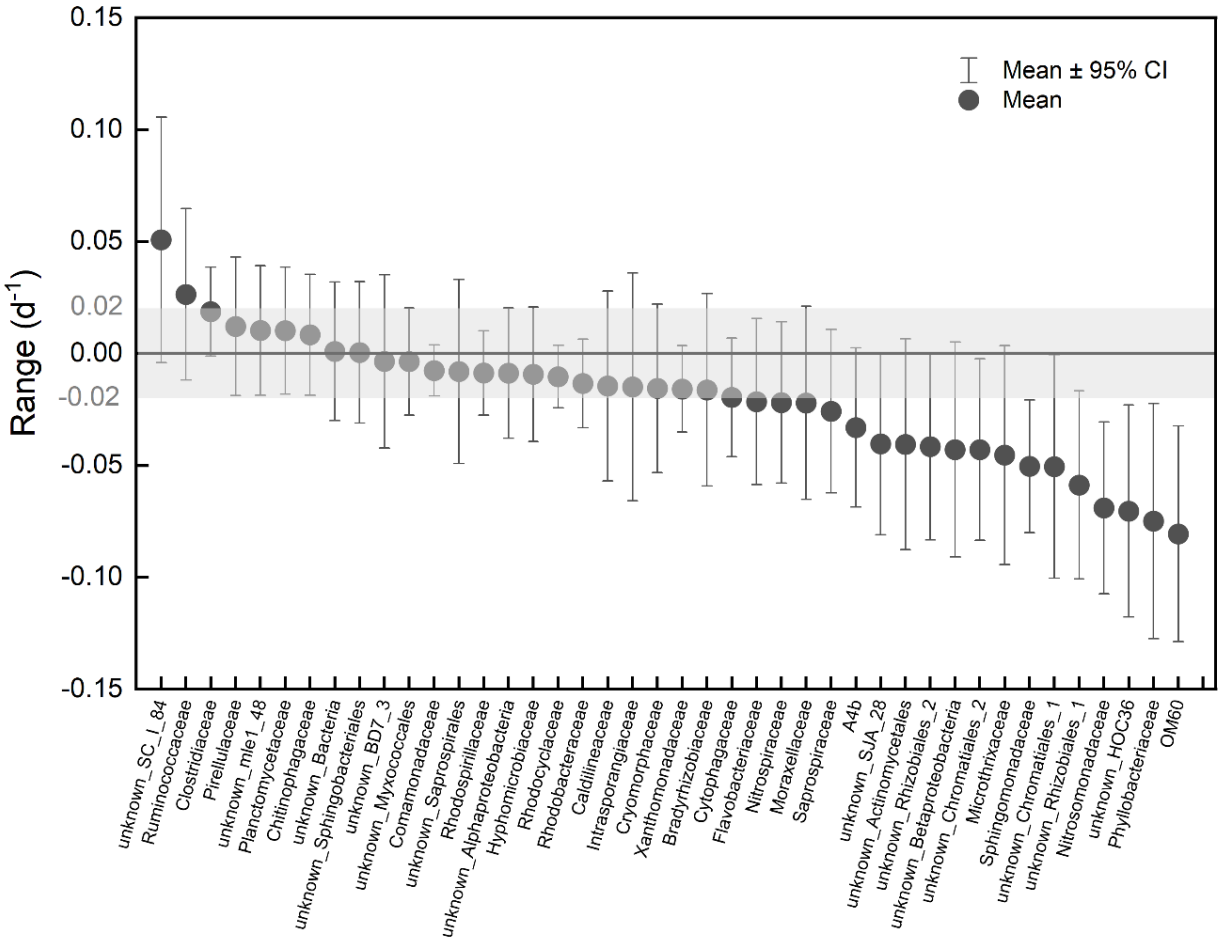
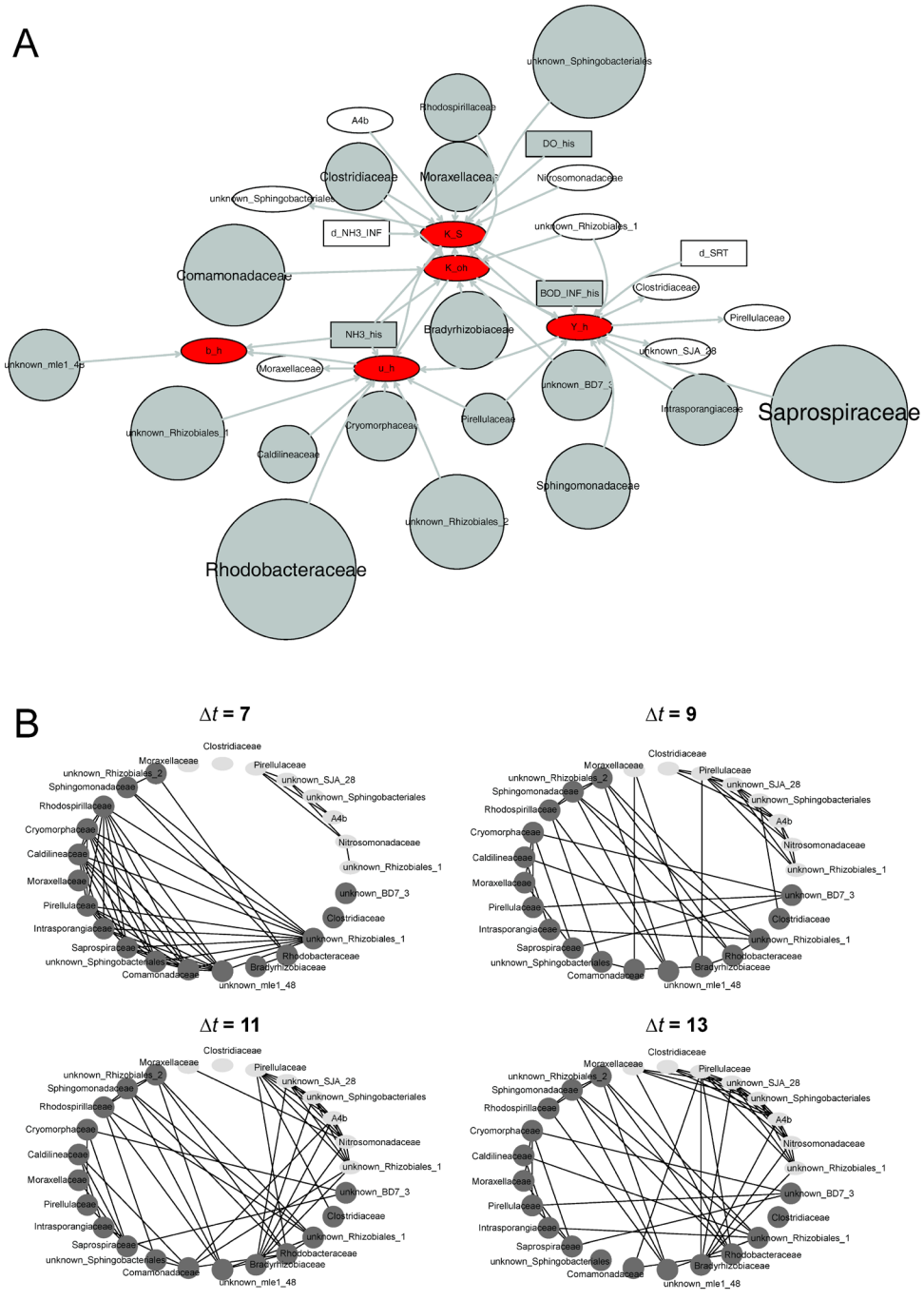


Figure 3. Relative abundance of the 42 core populations selected at the family level in the four activated sludge systems.



785

786 Figure 4. Overall dynamics of the 42 core populations across eight time spans and 936 transformed samples.



787

788 Figure 5. (A) Bayesian network centered on heterotrophic growth kinetics. Red oval indicates
789 microbial kinetic parameters, white oval indicates population dynamics, grey circle indicates
790 historical population abundance, white box indicates environmental perturbation, and grey box
791 indicates historical environmental factors. (B) Topological data analysis at Δt = 7, 9, 11, and 13.

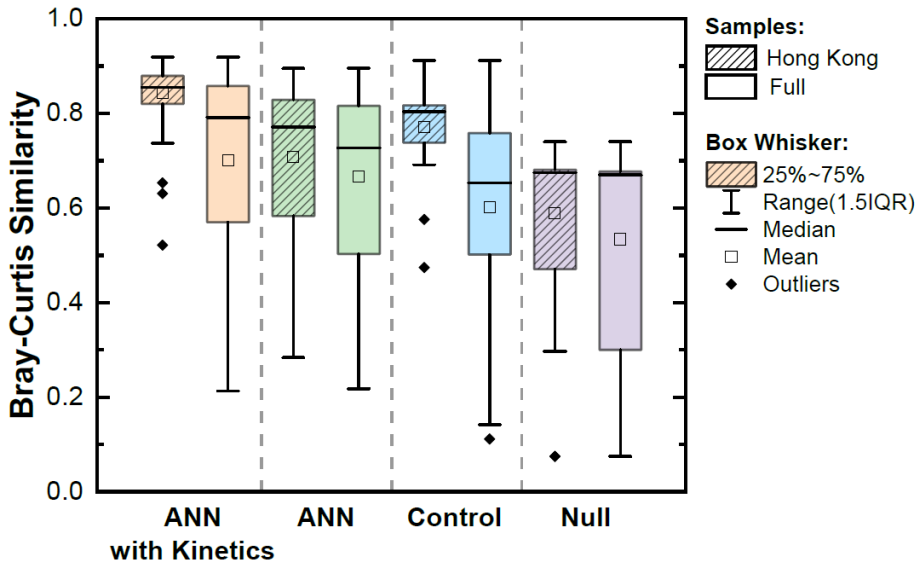


Figure 6. Bray-Curtis similarity between predicted and observed communities.