Mass-immigration determines the assembly of activated sludge microbial communities

Giulia Dottorini[a,1], Thomas Yssing Michaelsen[a,1], Sergey Kucheryavskiy[b], Kasper Skytte Andersen[a], Jannie Munk Kristensen[a], Miriam Peces[a], Dorottya Sarolta Wagner[a], Marta Nierychlo[a], and Per Halkjær Nielsen[a,2]

[a]Department of Chemistry and Bioscience, Section of Biotechnology, Center for Microbial Communities, Aalborg University, 9220 Aalborg, Denmark; and [b]Department of Chemistry and Bioscience, Section of Chemical Engineering, Aalborg University, 6700 Esbjerg, Denmark

Edited by Mary K. Firestone, University of California, Berkeley, CA, and approved April 30, 2021 (received for review October 26, 2020)

The assembly of bacterial communities in wastewater treatment plants (WWTPs) is affected by immigration via wastewater streams, but the impact and extent of bacterial immigrants are still unknown. Here, we quantify the effect of immigration at the species level in 11 Danish full-scale activated sludge (AS) plants. All plants have different source communities but have very similar process design, defining the same overall environmental growth conditions. The AS community composition in each plant was strongly reflected by the corresponding influent wastewater (IWW) microbial composition. Most species in AS across the plants were detected and quantified in the corresponding IWW, allowing us to identify their fate in the AS: growing, disappearing, or surviving. Most of the abundant species in IWW disappeared in AS, so their presence in the AS biomass was only due to continuous mass-immigration. In AS, most of the abundant growing species were present in the IWW at very low abundances. We predicted the AS species abundances from their abundance in IWW by using a partial least square regression model. Some species in AS were predicted by their own abundance in IWW, while others by multiple species abundances. Detailed analyses of functional guilds revealed different prediction patterns for different species. We show, in contrast to the present understanding, that the AS microbial communities were strongly controlled by the IWW source community and could be quantitatively predicted by taking into account immigration. This highlights a need to revise the way we understand, design, and manage the microbial communities in WWTPs.

activated sludge | immigration | community assembly

The interaction of ecological forces underpins community assembly in natural and engineered microbial ecosystems, such as wastewater treatment plants (WWTPs). Since WWTPs are responsible for the protection of human and environmental health and for recovery of energy and nutrients, it is crucial to understand the biological mechanisms that sustain the processes behind the community assembly. After a decade of debate, both stochastic processes, such as dispersal (e.g., immigration), and deterministic processes, such as environmental conditions (e.g., pH, temperature) and biotic interactions (e.g., competition, predation), are generally accepted to determine community assembly in a range of microbial ecosystems, including engineered systems (1–7). However, the difficulty of predicting stochastic processes leaves the community assembly in full-scale WWTPs poorly understood.

WWTPs can be regarded as open ecosystems (8), often consisting of the activated sludge (AS) process and continuously fed by upstream influent wastewater (IWW) (4, 9). Consequently, AS and IWW together define a metacommunity (8) interconnected by the unidirectional and continuous flow (dispersal) of "individuals" from IWW to AS, namely immigration. The immigrating "individuals," which include bacteria [also seen as "invaders" (10)], are considered the main source of biomass that potentially affects diversity, abundance, and assembly of the local AS microbial communities (4, 8, 9, 11–13). This was already proposed and accounted for in the early studies of AS engineering modeling (14–17). The development of sequencing technologies revealed that certain bacteria are shared between IWW and AS (e.g., refs. 9 and 18), so immigration should be tackled as mass flow immigration, here "mass-immigration," corresponding to the mass effect in Leibold's metacommunity paradigm (13, 19). AS plants are characterized by high dispersal rate with low hydraulic retention times [generally 4 to 12 h (20)], so a fraction of AS biomass is composed of IWW biomass (13). Deterministic factors, such as design and operation of AS plants, are believed to contribute to the AS microbial communities. In particular, the solid retention time (SRT) is considered a key operational parameter which directly controls the presence or absence of microorganisms (specific biomass growth) in the system, therefore affecting the community composition and the treatment performance (20). Today, the management of full-scale WWTPs across the world is almost exclusively based on the traditional Baas Becking and Beijerinck's deterministic principle of "Everything is everywhere, but, the environment selects" (21, 22). As a consequence, common WWTP types across the world are designed to establish a specific set of environmental conditions that select for the growth of certain microbial functional guilds carrying out the desired processes (species sorting) (23). These design types include simple oxygenation for carbon removal and nitrification and more advanced systems with complex dynamics of oxic/anoxic conditions for carbon, nitrogen, and

Significance

Wastewater treatment plants are engineering technologies used worldwide to protect the environment and human health. Microbial communities sustain these plants, so it is crucial to know the key factors responsible for the community assembly. We show, in contrast to existing understanding, that microbial immigration largely controls the community structure in these plants and that the fate (growth or death) of immigrating species in the plants is controlled by local factors. The community structure was quantitatively predicted by the immigrating microbial community, highlighting the need to revise the way we today understand, design, and manage microbial communities in wastewater treatment plants.

Author contributions: J.M.K., M.N., and P.H.N. designed research; G.D., K.S.A., J.M.K., M.N., and P.H.N. performed research; G.D., T.Y.M., S.K., K.S.A., M.P., D.S.W., and P.H.N. analyzed data; and G.D., T.Y.M., M.P., and P.H.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]G.D. and T.Y.M. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: phn@bio.aau.dk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2021589118/-/DCSupplemental.

Published June 29, 2021.

phosphorus removal. Microbial functional guilds specifically involved in these processes are, for example, nitrifiers, denitrifiers, and polyphosphate-accumulating organisms (PAOs).

Different approaches have been applied to evaluate the importance of stochastic processes, such as immigration, for community assembly in AS plants. Some approaches are mostly based on the analysis of experimental observations (data driven) (8, 13, 18, 24, 25). The approach proposed by Saunders et al. (9) detects all species in the system (IWW and AS) and establishes mass balances for the calculation of net-specific growth rates of all detected species in IWW and AS. In this way, the approach measures the fate of immigrating bacteria in quantitative terms, identifying microorganisms able or unable to grow in AS, thus distinguishing the bacteria likely important for the function of AS plants (9, 25–27). However, previous studies adopting this approach to investigate immigration did not address whether only the presence or also the abundance of species in the IWW play a role (9, 25). In that case, they also did not address how to predict the AS microbial communities given the community profile in IWW.

Other approaches to evaluate the role of immigration are theoretical, applying neutral and null models (4, 28–30). However, most of the theoretical models are based on simplified assumptions and do not consider the interactions between microbial communities in IWW and AS. As we will demonstrate in our study, data-driven models, such as multiresponse partial least squares regression (PLSR), can be beneficial to predict the abundance of species in AS based on the abundance in IWW. In this way, we can evaluate the impact of immigration without the need to build a specific model with the related simplifications and assumptions. Indeed, PLSR is capable of handling collinear and noisy variables and has the possibility to deconstruct and interpret the model by simple graphical representation (31). PLSR is widely and successfully used in various scientific disciplines dealing with high-dimensional data, such as bioinformatics, genomics, and spectro-metrics and is increasingly applied in the study of microbial ecology of the human microbiome (e.g., ref. 32). However, PLSR has very few implementations in engineered ecosystems such as AS plants (33, 34).

To study in detail the impact of immigration in WWTPs, we have selected 11 Danish full-scale AS plants characterized by very similar process design and thus very similar selective pressure acting on the AS microbial communities, and by different geographical locations representing potentially different dispersal properties. We applied 16S ribosomal ribonucleic acid (rRNA) gene amplicon sequencing and adopted the quantitative approach by Saunders et al. (9) to evaluate the fate of all detectable species in AS immigrating from the corresponding IWW. Afterward, we used PLSR to predict the species abundance in AS from IWW data and find correlations between the abundance of species in IWW and AS. Our results showed that mass-immigration strongly determined the AS microbial community structure (taxa composition and abundance) and that the taxa abundance could be predicted by IWW abundance. The results have a profound impact on the way we understand the assembly of AS plants' communities. Thus, a revision of the way they can be predicted and controlled is needed.

Results

Community Structure in IWW and AS. The 11 Danish AS plants investigated had very similar process design (biological removal of nitrogen and phosphorus), a large AS common core community, but also unique AS microbial signatures stable over years (35). These characteristics defined a similar selective pressure acting on the AS microbial communities. Moreover, the plants were located in different areas across Denmark, providing potential variations in the IWW source communities and dispersal conditions. We applied amplicon sequencing for community analyses of IWW and AS samples (118 samples arranged in 59 IWW/AS pairs according to the sampling week) obtaining 13,756 to 104,449 read per sample (mean: 42,287) and 12,991 to 124,078 (mean: 45,153), respectively.

In general, microbial communities in IWW and AS were different in terms of overall composition and abundance, as evident from the different groupings in the principal component analysis (PCA) plot (Fig. 1*A*) and the differences in species abundance shown by heat maps (*SI Appendix*, Fig. S1). While the 20 most abundant species in IWW made up 45.8% of the total read counts in IWW, they represented only 24.4% of the total read counts in AS (*SI Appendix*, Fig. S1). Moreover, the IWW communities varied more over time compared to AS, as seen by the less-tight clustering of samples in IWW compared to AS (Fig. 1*A*). Despite these differences, we observed pronounced clustering by plant, especially in AS. This indicates that samples from the same plant had a microbial community structure more similar to each other than compared to community structures from samples of other plants (Fig. 1*A*). Interestingly, IWW and AS paired samples (from the same plant and the same sampling week) exhibited a similar community gradient along the PC-2, which was consistent across all the plants. This is confirmed by the high correlation ($r^2 = 0.67$) between PC-2 of IWW and AS (Fig. 1*B*). Detailed analyses of community dissimilarity confirmed that IWW/AS paired samples from the same plant and the same sampling week were overall significantly more similar to each other than compared to any other IWW/AS pair combination (e.g., IWW/AS pair from same plant but different sampling week or IWW/AS pair from different plants) (Fig. 1*C*). This reflects the presence of a plant-specific microbial community between AS and its correspondent IWW source community.

Impact of Immigration on AS Microbial Communities. AS plants with similar process designs are characterized by a core of common abundant bacteria, which are considered to carry out important processes (9, 35). We investigated whether the abundant species in AS of each plant were detected in the corresponding IWW. We found that nearly all of the 50 most abundant species in AS were also detected in the correspondent IWW (Fig. 2*A*). In a few plants (Aalborg East and Haderslev), nearly all of the top 200 species in AS could be detected in IWW, while this was not the case in most of the other plants. This indicated the presence in IWW of some low-abundance species which were difficult to detect. To verify whether this dropout was related to technical detection limitations, we deep sequenced a subset of IWW samples including Aalborg West and Randers AS plants, obtaining 123,550 to 259,110 reads per sample. Deep sequencing greatly improved species-level detection in IWW, leading to the detection of all top 200 species in AS in the corresponding IWW source community (Fig. 2*B*).

Besides a common core community, AS plants with very similar process designs are also often characterized by the presence of unique species not present in other similar plants. Sometimes the presence of transient species can deteriorate the process provoking, for example, bulking (36). First, we identified the presence of unique species in AS samples of each plant; then, we investigated whether these unique species (plant specific) were also detected in the corresponding IWW of the same plant and/or in IWW of different plants (Fig. 2*C*). The number of unique species found in the AS and detected in the corresponding IWW from the same plant were significantly higher than the ones detected in IWW of all the other plants. This shows that the plant-specific species observed in AS were a fingerprint defined by the correspondent IWW.

Fate of Immigrating Bacteria. Microbial species passively transported from IWW into the AS go through one of three possible fates, or growth groups, defined by their behavior in the AS

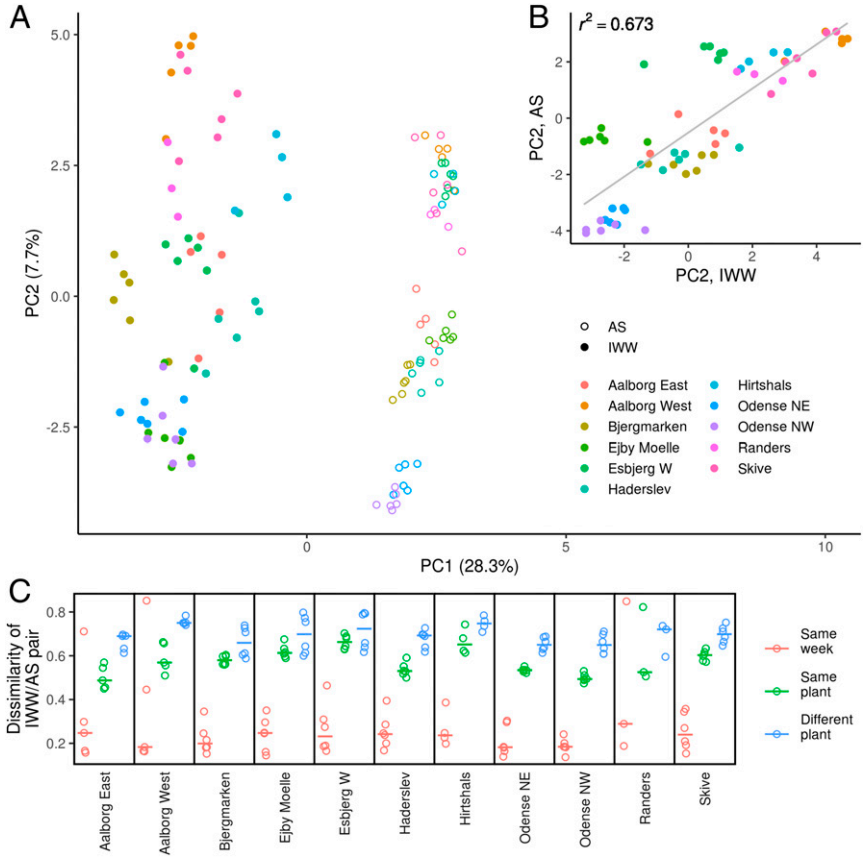

Fig. 1. Community diversity in IWW and AS in each AS plant. (*A*) PCA plot. Every point represents a sample colored by plant and shaped by location, AS (n = 59), or IWW (n = 59). (*B*) Second PC-axis values plotted against each other for IWW/AS paired samples (n = 59) from the same plant. Line shows a linear regression of PC2 values in AS on PC2 values in IWW with corresponding coefficient of determination (r^2). (*C*) Community dissimilarity of IWW/AS paired samples (points) shown for each plant and calculated by the abundance-weighted Bray–Curtis index. Different colors refer to dissimilarity values calculated for the following: IWW/AS paired samples from the same plant and the same sampling week (in red, n = 59); IWW/AS pairs of samples within the same plant (in green, n = 59); and IWW/AS pairs of samples across different plants (in blue, n = 59).

plant: growth, disappearance, or survival (see *Materials and Methods*). The expected "growing species" are immigrating species that can grow in AS and become part of the local AS community; they are assumed to perform essential processes. The expected "disappearing species" are immigrating species that, in absence of mass-immigration, are considered to disappear in AS due to factors such as washout, predation or die-off. Finally, the expected "surviving species" are immigrating species that likely slowly grow or disappear, depending on the operational conditions in the AS plant. The investigated AS plants had an average total SRT of 17.9 d (10.5 to 25.4 d) (*SI Appendix*, Table S1). Consequently, bacteria detected in AS throughout the three-month survey were only present because they could grow in the system (growing species) or because they were continuously added by mass-immigration (surviving and disappearing species). To investigate the fate of immigrating bacteria, we calculated mass balances (see *Materials and Methods*) and partitioned 1,510 species of the entire dataset into the three abovementioned growth groups (Dataset S1). Species with very low abundance in both IWW and AS were considered "ambiguous" and were not assigned to growth groups (see *Materials and Methods*).

The total biomass in terms of volatile suspended solids (VSS) transported by IWW to the plants per day constituted a large fraction (5% ± 1%) of the AS biomass already present in the 11 plants. The greatest fraction in IWW was composed of disappearing species (298 ± 55 species) (Fig. 3*A*), making up on average 72.5% ± 6.6 of the cumulative read abundance in IWW, but only 12.4% ± 5.3 in AS, showing a drastic reduction. They included some species in the IWW that, due to their high abundance in IWW, had a high relative read abundance in AS as well, such as species in the genera *Trichococcus*, *Acidovorax*, *Streptococcus*, and *Arcobacter* (*SI Appendix*, Figs. S1*A* and S2). In contrast, the growing species in AS (61.2% ± 9.3), as well as the most abundant species in AS, were the smallest fraction in IWW (Fig. 3*A*), constituting only 2.1% ± 1.1 on average (361 ± 46 species), most with low relative read abundance (<0.1%, *SI Appendix*, Figs. S1*B* and S2). The growing species in AS included also species assigned to functional guilds, as expected by their role in the AS community. For example, 48 species in the dataset were assigned as PAOs, important for biological phosphorus removal, such as the genera *Tetrasphaera*, *Dechloromonas*, and *Candidatus* Accumulibacter. Of these 48, 32 (67%) were growing, as expected from their potential role in AS plants. Glycogen-accumulating organisms (GAOs), assumed to compete with the PAOs, were also growing and included the genera *Micropruina* and *Candidatus* Competibacter. The growing fraction included also important species for nitrogen removal, such as nitrifiers within the genera *Nitrotoga* and *Nitrospira*. Furthermore, some filamentous genera such as *Candidatus* Microthrix and *Candidatus* Amarolinea were also growing (Dataset S1). Besides these well-known important species for the AS plant, many others with so-far poorly described or unknown functions were also growing. Finally, the surviving species (63 ± 47 species) represented a low cumulative read abundance fraction in both IWW and AS

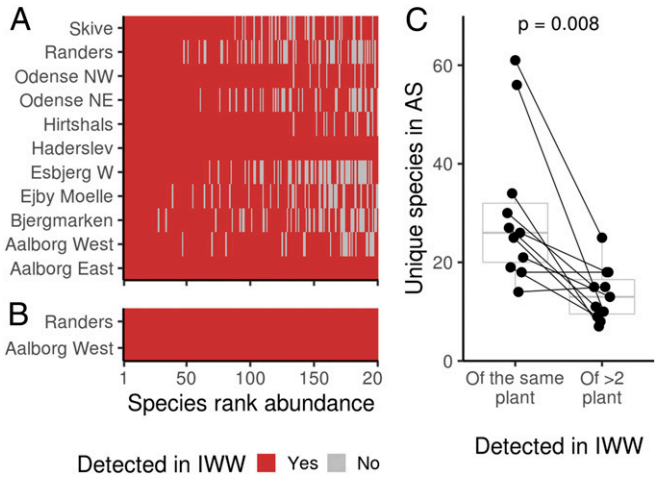

Fig. 2. Community detection in IWW and AS in each plant. (*A*) Top 200 most abundant species in AS ranked by abundance (*x*-axis) for each plant (*y*-axis). Red indicates presence (>1 read) and gray absence of a species in the correspondent IWW. (*B*) Subset of plants (Randers and Aalborg West) in which the same samples were deep sequenced (targeting 200,000 reads) and visualized as in *A*. (*C*) The unique species identified in AS of each plant with median abundance >0.05% were checked for detection (>1 read) in IWW. The *y*-axis shows the number of unique species identified in AS of each plant and detected in IWW. The *x*-axis shows the source of IWW from which the AS unique species were detected, either the IWW sample of the same plant or more than two IWW samples from other plants. Points paired along the gray lines represent the same plant (11 plants, 22 points in total). The *P* value refers to the differences along the *x*-axis across all plants, and it is computed from a paired Wilcoxon rank sum test.

(4.4% ± 5.5 in IWW and 4.4% ± 3.9 in AS) (Fig. 3*A*) and were characterized by both high and low relative read abundances (*SI Appendix*, Fig. S2).

The three different growth groups showed some variation of the cumulative read abundance in both IWW and AS samples from plant to plant, especially in IWW (Fig. 3*A*). However, each species included in the net-specific growth rate calculations had approximately the same fate in AS across different plants (Fig. 3*B*). The assignment to growth groups according to the cutoff (see *Materials and Methods*) was consistent for each species, as 1,465 of 1,510 (97%) species in the entire dataset were assigned to the same growth group. This provided robust conclusions about the fate of immigrating species in AS plants with very similar design.

Prediction of Species Abundance in AS from IWW. Due to the species composition similarities observed between paired IWW/AS samples from the same plant, we aimed to predict the abundance of species in AS based on IWW abundance. First, we performed a global PLSR analysis across plants, regressing the AS abundance table on the IWW abundance table (Fig. 4). The optimal trade-off between model performance and complexity of our data was represented by the use of 10 components (Fig. 4*A*). While some degree of plant-specific community was observed (Fig. 1), no distinct clustering by plant or outliers were detected for the PLSR model (Fig. 4*B*). Before modeling, only species that could be sufficiently quantified after applying filtering criteria were included (see *Materials and Methods*), for IWW and AS separately, across multiple plants. A large fraction of species included in the AS dataset was below the filtering criteria in IWW to be included into the PLSR model (Fig. 4*C*). This is likely an artifact of sequencing depth, as the deep-sequenced samples showed that species in AS were present in IWW, even if in small amounts (Fig. 2*B*). After filtering, a total of 3,533 species were retained in the dataset, of which 844 were shared between IWW and AS table, while 1,225 and 1,464 were only included in the IWW and AS table, respectively (Fig. 4*C*). The PLSR model performed well in predicting the relative abundances of multiple species in the AS based on IWW relative abundances. We considered the abundance of the species sufficiently predictable when more than 40% of the variation of the data can be explained by the model. Indeed, almost half (1,005 of 2,308) of the species included in the analysis from AS had an explained variation (leave-one-out cross validation [LOOCV-R^2]) > 0.4 (Fig. 4*E*). Furthermore, many of the species included from IWW (1535 of 2069) had a variable importance of projection score >1, indicating importance of the variable in the PLSR projection (37) (Fig. 4*D*).

To disentangle the complexity of the predictions from the PLSR model, we used an additional and more simplistic approach, the univariate model, which is a simple linear model. We compared the predictions from both models for the 884 species

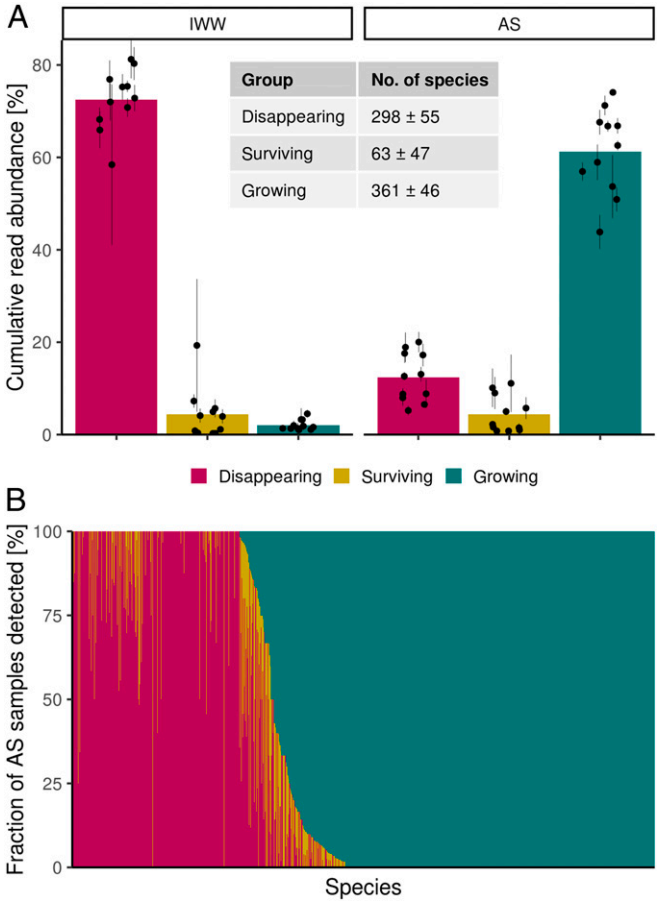

Fig. 3. Fate of immigrating bacteria. (*A*) Cumulative read abundance per growth group (growing, surviving, or disappearing in AS) at species level in IWW and AS samples in 11 AS plants. Each point represents the average across samples for each plant; bars represent SDs. The table in the middle indicates the corresponding plant-wise average number of species for each growth group (mean ± SD). (*B*) Consistency of calculated apparent net-specific growth rate across AS samples. The *x*-axis includes all species in the dataset for which it was possible to calculate an apparent net-specific growth rate in at least three AS samples (n = 1,510). The *y*-axis shows the relative fraction of AS samples for which a species was classified into each growth group (growing, surviving, or disappearing in AS), colored by petrol blue, ochre, and fuchsia, respectively. This relative fraction was calculated based only on AS samples in which a net-specific growth rate could be calculated; the total number of AS samples for calculating this may vary in each plant for each species with a minimum of three samples.

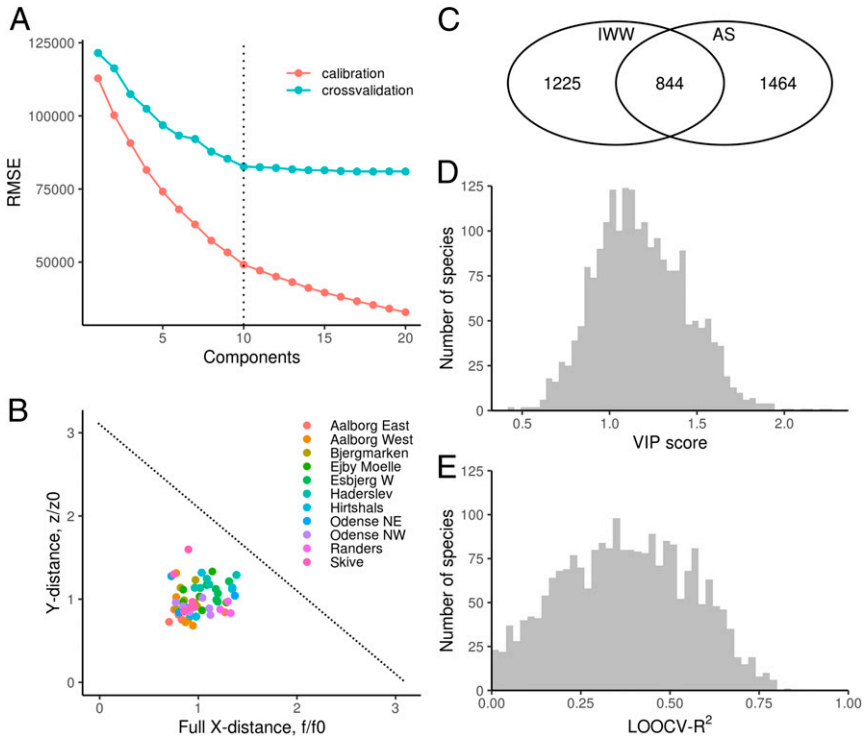

Fig. 4. PLSR model summary. (*A*) Selection of the optimal number of components used for the PLSR model based on cross validation. For the final PLSR model described here, 10 components were chosen, indicated by the vertical dotted line. (*B*) Biplot of the distance to the model of the decomposition matrix of the abundance in IWW (X-matrix, *x*-axis) and in AS (Y-matrix, *y*-axis) for each IWW/AS paired sample (points, $n = 59$). The dotted line shows the critical limit for outliers, defined at 99% confidence level. (*C*) Venn diagram shows the number of unique species from the entire dataset (after filtering criteria) included in the decomposition matrix of abundance in IWW and AS of the PLSR model. (*D*) Variable importance of projection (VIP) score plot shows the importance of individual IWW species in generally predicting the AS abundance. (*E*) The predictive performance of the PLSR model for each species is shown as the LOOCV-R^2.

shared between IWW and AS (Fig. 5*A*). Species abundances well predicted by the univariate model were also well predicted by PLSR, as it can extract the same predictive associations as the univariate model and even more complex relations. While most predictable species (LOOCV-$R^2 > 0.4$) were mainly explained by the PLSR model compared to the univariate model, we could not observe any clear relation between species abundance predictability and their fate (Fig. 5*A*). We interpreted the predictions from the two models in two different ways: "univocal association," when the highest variation in the data were explained by the univariate model, and "interspecies association," when the highest variation was explained by the PLSR model. The "univocal association" predictive pattern means that the abundance of a certain species in AS can be predicted directly by its own abundance in IWW. This implies that a species' abundance in AS is directly reflected and potentially determined by its abundance in IWW. This was exemplified by the species *Streptococcus parasuis* (Fig. 5*B*) in which the univariate model performed remarkably better (LOOCV-$R^2 = 0.76$) than the PLSR model (LOOCV-$R^2 = 0.61$). The "interspecies associations" predictive pattern means that the abundance of a certain species in AS can be predicted by the abundance of itself and/or other species interacting in IWW. This means that the abundance of multiple species in IWW correlates positively and negatively with the abundance of a species in AS. This case was exemplified by one of the most abundant species in the AS plants, *Tetrasphaera* midas_s_5 (Fig. 5*C*). The relative read abundance of this species in AS predicted from IWW by the PLSR model (LOOCV-$R^2 = 0.61$) performed much better than the univariate model (LOOCV-$R^2 = 0.09$). Thus, the abundance in AS of *Tetrasphaera* midas_s_5 was best described by an increase of abundances of other species in IWW.

Linking Functional Potential and Fate of Immigration to Predictability. Species belonging to functional guilds (such as filamentous organisms, PAOs, GAOs, and nitrifiers) have central roles in AS processes, but the effect of immigration on them is unknown. We investigated how the abundance of these species can be better predicted by the two predictive patterns (models), univocal or interspecies association, that we identified. Moreover, we took into account the fate of these species in AS, in light of the mass balance calculations performed above. Most of the species (86%, 305 of 352) belonging to the known functional guilds were able to grow in AS (Fig. 6 and Dataset S1), as expected. For most species, the explained variation (LOOCV-R^2) of the PLSR model was higher than the univariate model (Fig. 6). This was the case for *Tetrasphaera* midas_s_5, *Candidatus.* Amarolinea midas_s_1, and *Nitrospira defluvii*, which are species of particular interest for the function of AS plants due to their high abundance in AS and the role that they are assumed to perform. The "interspecies associations" in IWW were the main responsible for affecting the abundance of these species in AS (Fig. 6). However, a few other species, such as *Dechloromonas* midas_s_1978 among PAOs and *Candidatus* Microthrix midas_s_2 among filamentous bacteria, obtained a similar explained variation in both models. The abundance of some species belonging to the same functional guilds was predicted by different patterns. For example, the abundance in AS of *Microlunatus phosphovorus* and *Tetrasphaera elongata* (both PAOs) was predicted mainly by "univocal association" and "interspecies associations" in IWW, respectively.

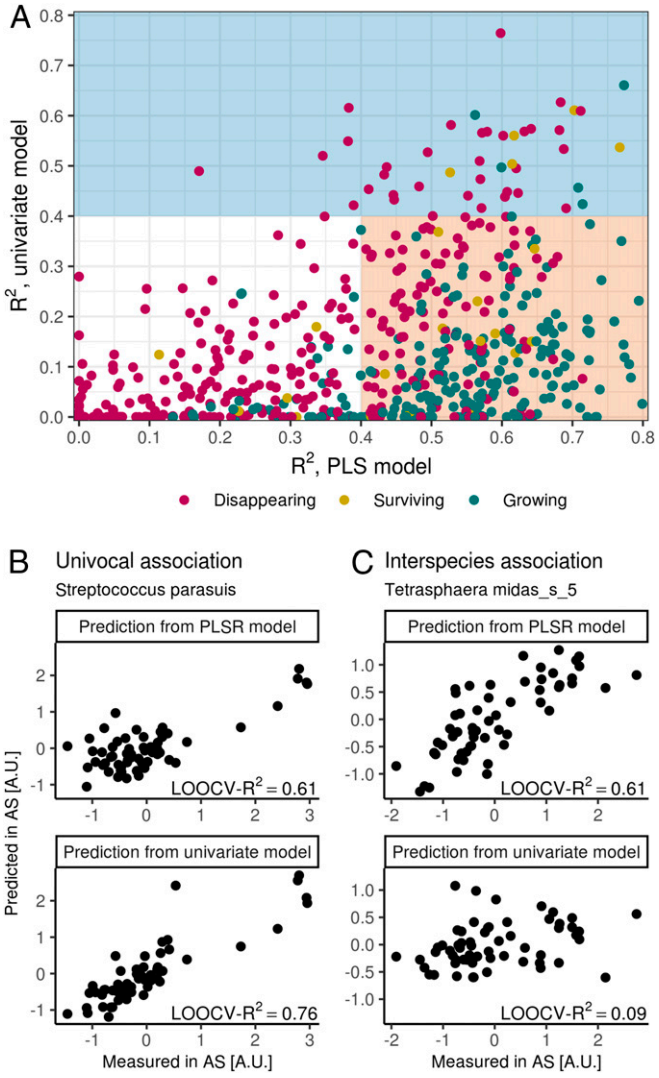

Fig. 5. Prediction by univocal association or interspecies associations of species of interest. (*A*) Codistribution of LOOCV-R^2 from the PLSR model (*x*-axis) and univariate model (*y*-axis) for species abundance in AS, based on IWW abundance. Every point (n = 844) represents a species colored by the growth group (growing, surviving, or disappearing). Ambiguous refers to species that could not be assigned to a growth group. Colored areas indicate good predictability, that is, high explained variance (LOOCV-R^2 > 0.4) for either the PLSR model (light-red area) or the univariate model (light-blue area). (*B* and *C*) Two different species of interest highlighting the two different predictive patterns by which a species' abundance can be predicted in AS (n = 59). On the *x*-axis is shown the standardized relative abundance (A.U.) measured in AS and on the *y*-axis the predicted standardized relative abundance (A.U.) in AS based on IWW standardized abundance either from the PLSR model (upper plots) or the univariate model (lower plots). The associated explained variance (LOOCV-R^2) is provided in the bottom right corner of the plot. *B* shows a univocal association for *Streptococcus parasuis*. Here, the correlation between standardized measured and predicted relative abundance in AS is higher in the univariate model than in the PLSR model (n = 59). *C* shows the case where interspecies interactions in IWW are responsible for the prediction of *Tetrasphaera* midas_s_5 abundance in AS (n = 59). Here, the PLSR model outperforms the univariate model.

Discussion

The open-system character of full-scale WWTPs forces recognition of the active role microbial immigration plays in the assembly of AS microbial communities. The importance has been widely investigated, but past approaches have yielded contradictory results, and a consensus about the role of immigration is still missing. Experimental limitations in previous studies may explain the contradictory results. Early studies of WWTP modeling acknowledged the importance of biomass in IWW (14–17). However, in the absence of powerful microbiological detection methods, IWW biomass was expressed as VSS (e.g., ref. 17), estimated from respirometric assays (16), or was neglected in the models due to its assumed low concentration compared to the biomass produced in AS (38). More recent studies are either based on the inference of immigration patterns without actual analyses of taxa in both IWW and AS communities (39), on the inference of taxa abundance in influent from AS community data (typical of theoretical model; *SI Appendix, Supplementary Discussion*) (28), or on the application of DNA-based methods with low resolution [such as Terminal Restriction Fragment Length Polymorphism (4) or pyrosequencing (8, 24, 40)].

The limit of microbiological methods in detecting bacteria has been a problem in many immigration-related studies, resulting in focusing only on the abundant taxa in the influent. For example, Lee et al. (8) used a sequencing depth of only ∼500 reads per sample (in contrast to ours of >40,000 reads on average per sample). Thus, they found only a small fraction of the operational taxonomic units (4.3 to 9.3%) shared between IWW and AS and that the communities were very different in terms of diversity, composition, and temporal variation. Gonzalez-Martinez et al. (24) reached the same conclusion investigating alpha and beta diversity of IWW and AS adopting a similar pyrosequencing approach with low resolution. The higher resolution of our high-throughput amplicon sequencing revealed that while the AS and IWW microbial communities initially appeared very different from each other in terms of abundance, most species were shared when deep sequencing was performed. In combination with mass balance calculations, multivariate statistics, and high-accuracy species-level resolution using the new MiDAS ecosystem-specific reference database and taxonomy (35), we could reveal the dramatic effect of mass-immigration also for the low-abundant species in IWW, as seen by the abundance prediction by PLSR. Other studies have also suggested that mass-immigration could be more important than previously recognized (13) and revealed the importance of mass-immigration for low-abundant species, even though focusing only on the nitrifiers functional guild (18).

The AS composition of each plant was reflected by the composition of its corresponding source community (both at overall community level and for unique species), suggesting source–sink dynamics are taking place between IWW and AS, as described by Leibold et al. for the mass effect (19). Continuous unidirectional mass flow of immigrating bacteria from IWW into AS plants (small spatial scale) contributes to homogenizing dispersal mechanisms between IWW and AS. This has been observed in natural aquatic systems in which mass effects can be expected when connectivity among habitats is very high (41, 42) but not yet reported for full-scale AS WWTPs. The effect of different source communities (IWW) may explain common observations in full-scale AS plants of temporal and spatial abundance variations of microbial communities (8). For example, IWW likely contributes to the unique microbial community composition and variation over time (13 y) we reported for more than 20 Danish full-scale enhanced biological phosphorus removal (EBPR) AS plants (35). Moreover, mass-immigration might be responsible for the abundance variations of process-critical bacteria (e.g., genus *Ca.* Microthrix) over time and rank abundance variations of different genera within a functional guild (e.g., the PAO genera *Ca.* Accumulibacter, *Tetrasphaera*, and *Dechloromonas*) in different AS plants. Consequently, stochastic processes seem important for the differences observed in community structure in AS plants across the world, as was recently proposed (30).

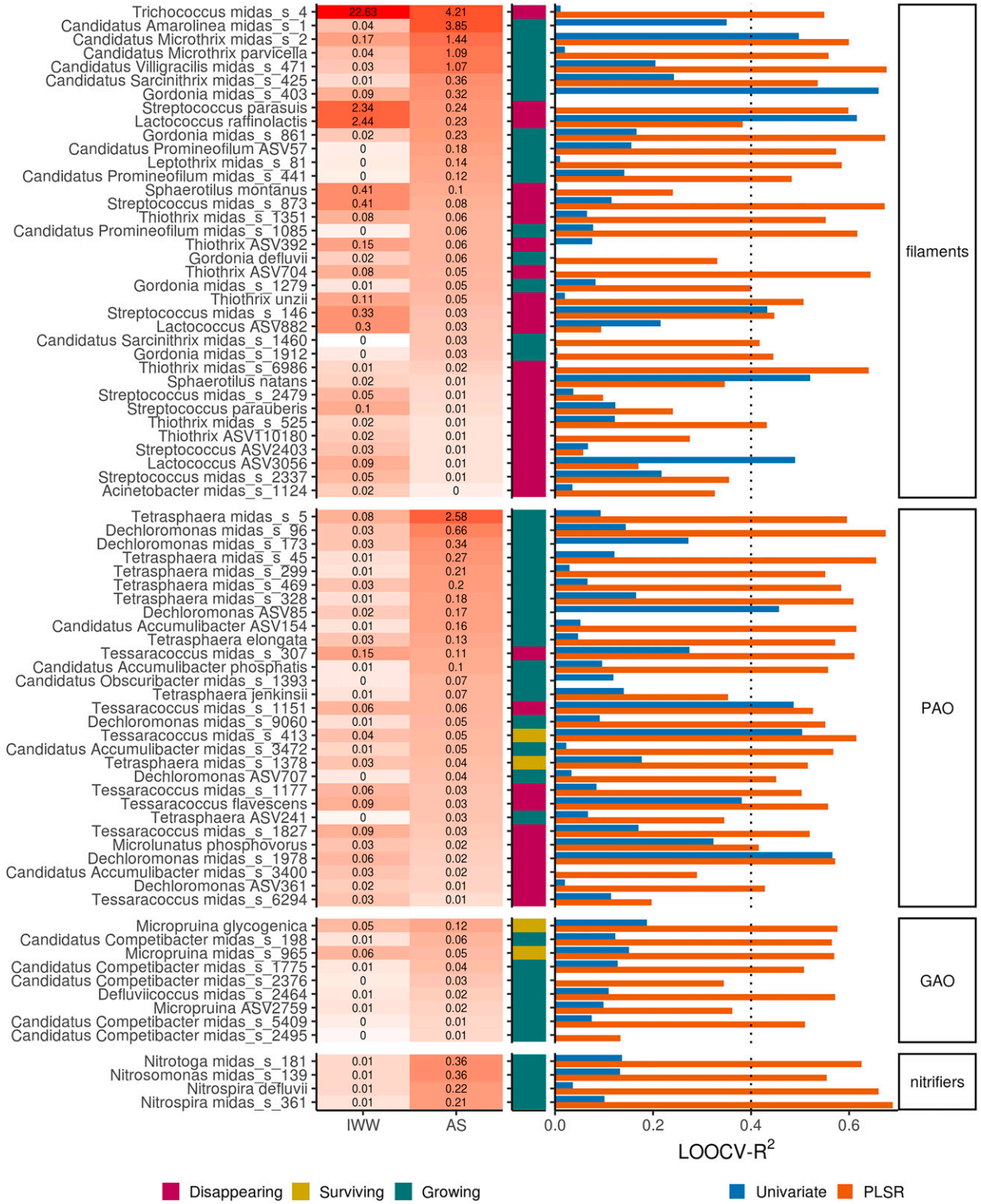

Fig. 6. Abundance prediction of functional guilds. The plot combines the characteristics investigated in this study (relative read abundance, fate, and predictability) for species belonging to known functional guilds (filamentous organisms, PAOs, GAOs, and nitrifiers). On the *y*-axis, to the extreme left, a subset of species belonging to functional guilds is shown. Following on the left, a heat map showing, for each species, the median relative read abundance (%) in IWW and AS (*x*-axis) sorted by relative abundance in AS is shown. Only species with abundance higher than 0.01% in either IWW or AS across all samples are shown, for a total of 78 out of 352 species belonging to functional guilds shown in the plot. In the middle, heat map showing the growth group (or fate) assigned to each species (in fuchsia: disappearing species; in ochre: surviving species; and in petrol blue: growing species). On the right, barplot of the variance (LOOCV-R^2, *x*-axis) explained by the PLSR model (in orange) and the univariate model (in blue) for each species. Dotted vertical line indicates the threshold for predictability (LOOCV-R^2 > 0.4), as explained in *Results*. To the extreme right, grouping of species according to the functional guilds.

Mass balance calculations provided overall metrics indicating that the total amount of biomass immigrating per day represented a large proportion of biomass already present in AS (5%), which is rarely recognized. Furthermore, mass balance calculations were critical to identify the fate of immigrating species in the AS (growing, disappearing, or surviving). A strong rescue-effect caused by continuous mass-immigration was likely responsible for the relatively high cumulative proportion of disappearing species in AS (12.6% of biomass of AS) (9, 19). Disappearing species belonged to the genera *Arcobacter*, *Acidovorax*, *Trichococcus*, *Streptococcus*, and *Blautia*, which are well-known to inhabit sewers (e.g., 43). We believe that these species are considered nongrowing in AS plants because they are "maladapted" (5), likely due to competitive exclusion (44, 45), predation (46, 47), or because of random death or lysis. However, the continuous immigration was responsible for keeping them abundant in the AS plant even if they were not taking part in the process, as indicated by their negative net-specific growth rate. Whether disappearing species may be active in the AS process before they disappear is not known (9). The growing fraction constituted the majority of the biomass in the AS (61.2% cumulative read abundance) and included species belonging to known functional guilds that are expected to grow in AS. The growing species had only low or very low abundances in the IWW and could easily be overlooked without deep sequencing. Their fate indicated that they were randomly transported by the IWW into the AS systems in which they could finally thrive and take part in the wastewater treatment process as a result of species sorting carried out by the design and operation of the plants.

The fate of each species proved to be consistent across the AS plants of this study. This consistency was ensured by investigating AS plants with very similar process design, that is, the EBPR process, and similar operations, that is, relatively long SRTs (10.5 to 25.4 d). Comparing the effect of different source communities (IWW) in plants with different process designs, such as EBPR and simple carbon removal plants, and/or different SRTs, such as long and SRTs of a few days, could lead to biased or inconclusive results as the fate of many immigrating species would vary according to the available local niche. However, it would be very interesting to investigate further the effect of mass-immigration in such plants, which would likely reveal a different fate for many species. We hypothesize the results of such a study would show an even greater effect of mass-immigration than found in our study, as plants with short SRT are regarded more vulnerable to disturbances such as dispersal and colonization of invaders (25, 48, 49).

In our study, the PLSR approach was able to capture the multivariate nature of the data, including collinearity of species, which is otherwise difficult to tackle. PLSR provided good performance and interpretability on our data. Alternative methods widely used in microbial ecology include redundancy analysis and correspondence analysis; however, they have limitations for our type of dataset and scope (*SI Appendix*, *Supplementary Discussion*). PLSR revealed that the abundance of most species in AS could be predicted from the abundance in IWW, either from the same species ("univocal association") or by other species ("interspecies association"). Both cases represent a perspective to interpret microbial immigration in AS plants. Interspecies association was found to be the most common predictive pattern for most species, including functional guilds, in our dataset. This indicated the presence of interspecies abundance correlation in IWW and AS. Even if positive and negative correlations might indicate species co-occurrence and coexclusion, we do not directly imply any metabolic or trophic connection between those species; however, this aspect cannot be excluded. Interspecies associations also suggest the immigration of species from a similar source (coimmigration) along the sewer network. Moreover, they might indicate that microbial processes occurring upstream in the sewer systems induce the increase or decrease of abundance for certain species in IWW. In addition, it is possible that minor variations in the operation of the EBPR plants affected the relative abundances of these species, also influencing the predictions across plants. It is not clear yet why and how these interspecies interactions occur, and further studies are needed.

The AS microbial community structure was determined by both stochastic and deterministic factors, as agreed in the "reconciliation" of niche and neutral theory of community assembly (2, 50). As reviewed by Zhou and Ning (6), plenty of studies have addressed the importance of either niche or neutral theory in the past decades. A common challenge is represented by the difficulty to reasonably reject one or another theory, given for example that different processes (or assumptions) can yield very similar patterns. Moreover, as discussed above, part of this long-standing literature relies on outdated methodological approaches. We propose that in AS plants, deterministic and stochastic forces act in a hierarchical assembly, in similar way as these forces occur at macroscale. First, a dispersal filter (i.e., mass-immigration of the IWW source community) is the key player at the metacommunity level and determines the overall AS species composition (taxa presence/absence) and partially the species abundance (mass effect). Then, the environmental filter (i.e., process design [e.g., EBPR] and operation [e.g., SRT] of AS plants) determines the fate of immigrating species. The combined effect will determine the structure of microbial communities in AS. Our study shows that mass-immigration played a much greater role than previously believed and that it is important to distinguish between the fate of species (growth groups) when understanding and investigating the impact of immigration.

The importance of immigration seems to vary in different engineered ecosystems according to the actual species sorting mechanisms. Immigration was recorded to play a minor role in granular anaerobic sludge bioreactors (25) whereas it had a great impact in anaerobic digesters at WWTPs treating surplus AS and primary sludge (51). The latter is also an example of within-WWTP immigration, and other studies investigating the transfer (immigration) of biomass across full-scale plant's units are available, although they reach contrasting conclusions about immigration (45, 52–54). Therefore, specific conclusions drawn for one specific system (e.g., the AS process) cannot easily be extrapolated to another treatment system because the assembly forces (e.g., source communities, species sorting, and mass effect) will have different strengths.

The current design and operation of full-scale AS WWTP is based on the traditional Baas Becking and Beijerinck's deterministic approach that "Everything is everywhere, but, the environment selects" (21, 22). The importance of microorganisms in the IWW is often acknowledged with the terms "seeding" or "inoculum" for the AS community (14, 43). It is important to review this thinking. Our results showed that mass-immigration is responsible for both the presence and abundance of most species in the AS, especially for those species (the disappearing) that are not fit for it. Secondly, the abundance of species in IWW, no matter at what level, contributes to the abundance (and prediction) of AS microbial communities. Therefore, mass-immigration, and its inherent stochasticity, has proved to be more important for the AS bacterial microbial communities than previously believed. This highlights the need to revise the way we currently understand, design, and manage microbial communities in AS WWTPs. While the optimization of AS communities through operational conditions remains relevant, more focus needs to be on the effect of immigration in WWTPs, considering AS and IWW as a unique entity (open system). The source community is essential for effective process treatment, challenging us to manage AS communities directly in sewer systems.

Materials and Methods

Sample Collection. A total of 11 full-scale AS plants in Denmark were sampled for IWW and AS over 3 mo from October to December 2014. All plants were characterized by phased-isolation oxidation ditch also called Bio-denipho technology (Kruger, Veolia Water Technologies), defining very similar process designs. They mainly treated municipal wastewater with biological nitrogen and/or enhanced biological phosphorus removal configuration (EBPR). The plants' location was spread across Denmark. All the main characteristics of the AS plants and the correspondent number of samples taken from each are summarized in *SI Appendix*, Table S2. Samples were collected every second week, with slight time differences between plants due to variation in their individual sampling programs or dry-weather days. Samples for microbiological analysis were collected as follows: IWW samples were collected as 50-mL subsamples from a 24-h flow proportional sampler, if present, and AS samples consisted of 2-mL subsamples of 50-mL grab samples collected from the aeration tank. IWW and AS samples were taken in duplicates for every sampling week for Aalborg West and Skive plants. All subsamples were stored at −20 °C until further analysis. AS plants design and process information about the plants and IWW were obtained from the plant operators for establishing microbial mass balances, including the following: volume of process tanks, amount of suspended solids in the process tanks, IWW flow, chemical oxygen demand (COD) of IWW, and effluent (*SI Appendix*, Table S1).

Amplicon Sequencing Workflow. Bacterial DNA extraction and 16S rRNA gene amplicon sequencing were performed for both IWW and AS samples. Prior to DNA extraction, samples were thawed, vortexed, and homogenized using 15 mL for each IWW sample (or 30 mL for duplicate samples) and 0.5 mL for AS samples; the homogenization was performed with overhead stirrer (Heidolph RZR 2020) with second gear, speed 9, moving 10 times from top to bottom the sample inside the tissue grinder. Homogenized IWW samples were then vacuum filtered through a 0.2-μm polycarbonate membrane, supported by glass fiber filters, by means of a DHI filtration manifold (Carbon 14 Centralen). Cells from IWW samples immobilized on the membrane's surface, and 0.5 mL homogenized AS samples were later processed according to the same DNA extraction and amplicon sequencing method, as described by ref. 55. DNA extraction was performed using the FastDNA Spin Kit for soil (MP Biomedicals), repeating the bead beating four times instead of once. Genomic DNA was quantified with the Qubit 2.0 fluorometer (Invitrogen) using the Qubit double stranded (ds) DNA BR Assay Kit (Thermo Fischer Scientific). Amplicon sequencing was performed targeting region V1-V3 of 16S rRNA gene using a modified procedure from Caporaso et al. (56). Briefly, 10 ng of extracted DNA was used as a template. The PCR (25 μL) contained dNTPs (400 μM of each), $MgSO_4$ (1.5 mM), Platinum Taq DNA polymerase High Fidelity (2 mU), 1X Platinum High Fidelity buffer (Thermo Fisher Scientific), and a pair of barcoded library adaptors (400 nM). V1-V3 primers used had the following sequences: 27F 5′-AGAGTTTGATCCTGGCTCAG-3′ (57); 534R 5′-ATTACCGCGGCTGCTGG-3′ (58). The PCR amplification was performed by a thermo cycler with the following settings: initial denaturation at 95 °C for 2 min, 30 cycles of 95 °C for 20 s, 56 °C for 30 s, 72 °C for 60 s, and final elongation at 72 °C for 5 min. All PCR reactions were run in duplicate and pooled afterward. The amplicon libraries were purified using the Agencourt AMPure XP bead protocol (Beckmann Coulter) with the following exceptions: the sample/bead solution ratio was 5/4, and the purified DNA was eluted in 23 μL nuclease-free water. The amplicon library concentration was then measured with Qubit dsDNA HS Assay Kit (Thermo Fischer Scientific) on a Qubit 2.0 fluorometer (Invitrogen), and the quality was evaluated with a TapeStation 2200 system using D1000 ScreenTape (Agilent Technologies). Amplicon libraries were then pooled in equimolar concentrations. The library pool was diluted to a final concentration of 4 nM prior paired-end (2 × 300 bp) sequencing on a MiSeq (Illumina). A subset of IWW samples from Aalborg West and Randers AS plants were prepared in the same way but sequenced deeper to evaluate the effect of technical detection limitations. All raw amplicon reads generated by MiSeq were processed using the AmpProc5.0 workflow (https://github.com/eyashiro/AmpProc) which generated amplicon sequence variants (ASVs). The ASVs were mapped to the full-length ASVs (FL-ASVs) from the database generated by Dueholm and colleagues (59) which is specific for wastewater treatment ecosystems and uses MiDAS 3 taxonomy (35). We chose to use the species-level classification to improve the read counts available for each investigated taxa. The classification at species level used in the database is based on 98.7% sequence identity as recommended by Yarza et al. (60) When species-level classification was not available, the taxonomic classification was assigned by combining the ASV classification with the first available taxonomic level (e.g., genus). Sequencing data are available online at European Nucleotide Archive (ENA) as subset of the project number PRJEB28796 (https://www.ebi.ac.uk/ena/browser/view/PRJEB28796) as indicated in Dataset S2. The sequences with higher sequencing depth are available online at National Center for Biotechnology Information (NCBI) with project number PRJNA715652 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA715652/).

Data Analysis. Data were imported into R v3.6.2 (https://www.R-project.org/) using RStudio IDE v1.2.5033 (https://www.rstudio.com/). Prior to analysis, samples with less than 10,000 reads were discarded upfront, and duplicate samples from the same plant and sampling week were combined by taking the median value of relative read abundance for each species. After filtering, only complete sample pairs of IWW and AS were retained. This resulted in a dataset of 118 samples from eleven different AS plants (*SI Appendix*, Table S2). Beta diversity was examined on relative abundance transformed data by the abundance-weighted Bray–Curtis index (61) and PCA. Additionally for PCA, all species not observed in more than 25% (n = 29) of samples were removed to reduce noise, and the remaining data were square-root transformed and standardized to zero mean and unit variance.

To perform PLSR, the mdatools v0.10.1 package (62) was used. Prior to analysis, the data were first Hellinger transformed. The data was then filtered separately for IWW and AS as follows: for each species, it was first determined whether it was detected in a given plant, defined as being present (read count > 1) in >50% of samples from that plant. Secondly, the fraction of plants in which a species was detected should be more than 25% (n > 2). After filtering, species not detected in more than 25% (n > 14) of samples were dichotomized to −1 (absent) and 1 (present). The abundances of the remaining species were standardized to zero mean and unit variance. To evaluate the performance and complexity (i.e., number of components) of the PLSR model, we used the coefficient of determination (R^2) computed from LOOCV-R^2, which represents the variance of the data explained by the model.

Apparent Net-Specific Growth Rate Calculations. To evaluate the fate of immigrating bacteria within the AS communities, we calculated the apparent net-specific growth rate (k) of every species. We used mass balance calculations between IWW and AS samples, revising the approach described in Saunders et al. (9) as follows.

We used the following assumptions: the apparent net-specific growth rate can be described as a first-order process; the biomass concentration of a j-species (X_j) can be described by both the relative read abundance of a j-species (p_j) and the total number of cells (n); and the biomass removed with effluent is typically very little, as seen by the COD concentrations in *SI Appendix*, Table S1, and can be neglected. Under these assumptions, the mass balance equation for every j-species inside the AS process tank of the plant is calculated as follows:

$$V_{PT}\frac{d\left(p_{AS,j}\cdot\ n_{AS}\right)}{dt} = Q_{IWW}\cdot p_{IWW,j}\cdot n_{IWW} - Q_{SP}\cdot p_{SP,j}\cdot n_{SP} + k_j\cdot p_{AS,j}\cdot n_{AS}\cdot V_{PT}, \quad [1]$$

where

- V_{PT} = volume of process tank [m^3]
- $p_{AS,j}$ = relative read abundance of j-species in AS [%]
- n_{AS} = total number of cells in AS [$\frac{cells}{m^3}$]
- Q_{IWW} = flow rate of IWW [$\frac{m3}{d}$]
- $p_{IWW,j}$ = relative read abundance of j-species in IWW [%]
- n_{IWW} = total number of cells in IWW [$\frac{cells}{m^3}$]
- Q_{SP} = flow of surplus sludge [m^3]
- $p_{SP,j}$ = relative read abundance of j-species in surplus sludge [%]
- n_{SP} = total number of cells in surplus sludge [$\frac{cells}{m^3}$] and
- k_j = apparent net-specific growth rate of the j-species [d^{-1}].

Since at steady state there is no net variation over time of the number of cells in AS ($\frac{dn_{AS,j}}{dt}$ = 0), Eq. **1** can be rewritten as follows:

$$Q_{IWW}\cdot p_{IWW,j}\cdot n_{IWW} - Q_{SP}\cdot p_{SP,j}\cdot n_{SP} + k_j\cdot p_{AS,j}\cdot n_{AS}\cdot V_{PT} = 0. \quad [2]$$

Rearranging Eq. **2** by applying the definition of HRT (hydraulic retention time) and SRT (sludge retention time, in which the biomass removed with the effluent is neglected), which are HRT = $\frac{V_{PT}}{Q_{IWW}}$ [d] and SRT = $\frac{V_{PT}\cdot p_{AS,j}\cdot n_{AS}}{Q_{SP}\cdot p_{SP,j}\cdot n_{SP}}$ [d] respectively, the apparent net-specific growth rate (k_j) of every j-species observed in the plant is calculated as follows:

MICROBIOLOGY

$$k_j = \frac{1}{SRT} - \frac{1}{HRT} \cdot \frac{p_{IWW,j} \cdot n_{IWW}}{p_{AS,j} \cdot n_{AS}} [d^{-1}]. \quad [3]$$

The apparent net-specific growth rate of every species in the plant is thus determined by both amplicon data and key design and process information. In particular, amplicon sequencing data are expressed by $p_{AS,j}$ and $p_{IWW,j}$ terms, while the other parameters are calculated or measured as follows:

- SRT, total sludge retention time: $\frac{VSS_{AS} \cdot V_{PT}}{Y_{obs} \cdot Q_{IWW} \cdot (COD_{IWW} - COD_{effl})}$ [d],

where

- VSS_{AS} volatile suspended solids in AS [g/L]
- V_{PT} volume of process tank [m^3]
- Y_{obs} observed AS yield for heterotrophs from municipal wastewaters, $\simeq$ 0.43 $\frac{gVSS}{gCOD}$ (20)
- Q_{IWW} flow rate of IWW [$\frac{m3}{d}$]
- COD_{IWW} chemical oxygen demand of IWW [mg/L]
- COD_{effl} chemical oxygen demand in effluent [mg/L]

- HRT, hydraulic retention time: $\frac{V_{PT}}{Q_{IWW}}$ [d]
- n_{IWW}, total number of cells in IWW: 1.4 $\cdot 10^{14}$ [$\frac{cells}{m^3}$] (20)
- n_{AS}, total number of cells in AS: $VSS_{AS} \cdot N_{AS}$ [$\frac{cells}{m^3}$],

where

- N_{AS} number of cells in AS, 0.5 $\cdot 10^{15}$ [$\frac{cells}{kgVSS}$].

According to Eq. **3**, bacterial species can assume values of apparent net-specific growth rate from minus infinite to 1/SRT, that is, $k = \{-Inf; 1/SRT\}$. In particular, Eq. **3** allows to distinguish the fate of immigrating bacteria into three different growth groups:

- "growing," with a clearly positive growth rate (k > 0), expected to grow in AS;
- "disappearing," with a clearly negative growth rate (k < 0), expected to disappear from AS systems in absence of immigration and;
- "surviving," with a growth rate of ~0 (k $\simeq$ 0), expected to either grow or disappear depending on the stringency of the plant parameters.

These growth groups are all characterized by species with a relative read abundance in IWW and/or AS higher than 0.05%. Species with lower abundance are, instead, defined here as "ambiguous," because we consider that such a low abundance cannot be used to clearly assign these species to a growth group.

The distinction into three different growth groups, instead of simply two based on the theoretical cutoff of k = 0, arises to provide a more realistic situation. It is expected that most immigrating species can be clearly growing or clearly disappearing in AS and that their fate is consistent across AS plants with similar designs. However, the fate of some immigrating species may vary according to the stringency of the plant's design. These species are neither growing nor disappearing in AS but rather surviving with an apparent net-specific growth rate distributed around 0. This implied that in some plants, the surviving species are growing, while in others they are disappearing. Indeed, when theoretically k = 0, Eq. **3** is:

$$p_{AS,j} \cdot n_{AS} = \frac{SRT}{HRT} \cdot p_{IWW,j} \cdot n_{IWW}. \quad [4]$$

For a given sample pair of IWW and AS with known SRT and HRT, Eq. **4** becomes a linear function of the IWW abundance with an intercept through 0. Sample pairs from the same plant are then used to construct an interval of AS abundances for that particular plant, defining the abundance of the surviving species. We used the fifth and 95th percentile, assuming that the variation of AS abundances of surviving species is similar across plants with similar process design, and it is 10%. In this way, species with AS abundances above the interval are considered clearly growing, while species with AS abundances below the interval are considered clearly disappearing. In this dataset, every species was assigned to a certain growth group when for that species, the growth group was observed in >50% of IWW-AS sample pairs (Dataset S1).

Data Availability. Amplicon sequencing data have been deposited in ENA project PRJEB28796 (https://www.ebi.ac.uk/ena/browser/view/PRJEB28796) and National Center for Biotechnology Information PRJNA715652 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA715652/).

ACKNOWLEDGMENTS. We thank all the plant operators for the collection of samples and process parameters. The project has been funded by the Villum Foundation (Grant 16578, Microbial Dark Matter) and Aalborg University.

1. J. S. Clark *et al*., Resolving the biodiversity paradox. *Ecol. Lett.* **10**, 647–659, discussion 659–662 (2007).
2. M. Vellend, Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206 (2010).
3. M. K. Pholchan, Jde. C. Baptista, R. J. Davenport, W. T. Sloan, T. P. Curtis, Microbial community assembly, theory and rare functions. *Front. Microbiol.* **4**, 68 (2013).
4. I. D. Ofiţeru *et al*., Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 15345–15350 (2010).
5. E. B. Graham, J. C. Stegen, Dispersal-based microbial community assembly decreases biogeochemical function. *Processes* **5**, 65 (2017).
6. J. Zhou, D. Ning, Stochastic community assembly: Does it matter in microbial ecology? *Microbiol. Mol. Biol. Rev.* **81**, e00002-17 (2017).
7. D. Ning, Y. Deng, J. M. Tiedje, J. Zhou, A general framework for quantitatively assessing ecological stochasticity. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16892–16898 (2019).
8. S. H. Lee, H. J. Kang, H. D. Park, Influence of influent wastewater communities on temporal variation of activated sludge communities. *Water Res.* **73**, 132–144 (2015).
9. A. M. Saunders, M. Albertsen, J. Vollertsen, P. H. Nielsen, The activated sludge ecosystem contains a core community of abundant organisms. *ISME J.* **10**, 11–20 (2016).
10. M. Kinnunen *et al*., A conceptual framework for invasion in microbial communities. *ISME J.* **10**, 2773–2775 (2016).
11. O. C. Shanks *et al*., Comparison of the microbial community structures of untreated wastewaters from different geographic locales. *Appl. Environ. Microbiol.* **79**, 2906–2913 (2013).
12. Y. Zha, M. Berga, J. Comte, S. Langenheder, Effects of dispersal and initial diversity on the composition and functional performance of bacterial communities. *PLoS One* **11**, e0155239 (2016).
13. D. Frigon, G. Wells, Microbial immigration in wastewater treatment systems: Analytical considerations and process implications. *Curr. Opin. Biotechnol.* **57**, 151–159 (2019).
14. M. Henze, The influence of raw wastewater biomass on activated sludge oxygen respiration rates and denitrification rates. *Water Sci. Technol.* **21**, 603–607 (1989).
15. M. Henze, Characterisation of wastewater for modeling of activated sludge processes. *Water Sci. Technol.* **25**, 1–15 (1992).
16. G. H. Kristensen, P. E. Jorgensen, M. Henze, Characterization of functional microorganism groups and substrate in activated sludge and wastewater by AUR, NUR and OUR. *Water Sci. Technol.* **25**, 43–57 (1992).
17. B. E. Rittmann, P. L. McCarty, *Environmental Biotechnology: Principles and Applications* (McGraw-Hill, ed. 1, 2001).
18. S. Jauffur, S. Isazadeh, D. Frigon, Should activated sludge models consider influent seeding of nitrifiers? Field characterization of nitrifying bacteria. *Water Sci. Technol.* **70**, 1526–1532 (2014).
19. M. A. Leibold *et al*., The metacommunity concept: A framework for multi-scale community ecology. *Ecol. Lett.* **7**, 601–613 (2004).
20. Metcalf & Eddy; G. Tchobanoglous, H. D. Stensel, R. Tsuchihashi, F. L. Burton, *Wastewater Engineering: Treatment and Resource Recovery* (McGraw-Hill Education, ed. 5, 2014).
21. M. A. O'Malley, 'Everything is everywhere: But the environment selects': Ubiquitous distribution and ecological determinism in microbial biogeography. *Stud. Hist. Philos. Biol. Biomed. Sci.* **39**, 314–325 (2008).
22. J. B. H. Martiny *et al*., Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
23. J. l. C. Jansen, E. Arvin, M. Henze, P. Harremoes, *Wastewater Treatment - Biological and Chemical Processes* (Polyteknisk Forlag, ed. 4, 2019).
24. A. Gonzalez-Martinez *et al*., Comparison of bacterial communities of conventional and A-stage activated sludge systems. *Sci. Rep.* **6**, 18786 (2016).
25. M. Ali *et al*., Importance of species sorting and immigration on the bacterial assembly of different-sized aggregates in a full-scale aerobic granular sludge plant. *Environ. Sci. Technol.* **53**, 8291–8301 (2019).
26. A. K. Winegardner, B. K. Jones, I. S. Y. Ng, T. Siqueira, K. Cottenie, The terminology of metacommunity ecology. *Trends Ecol. Evol.* **27**, 253–254 (2012).
27. R. Mei, W.-T. Liu, Quantifying the contribution of microbial immigration in engineered water systems. *Microbiome* **7**, 144 (2019).
28. W. T. Sloan *et al*., Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).
29. J. C. Stegen *et al*., Quantifying community assembly processes and identifying features that impose them. *ISME J.* **7**, 2069–2079 (2013).
30. L. Wu *et al*.; Global Water Microbiome Consortium, Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* **4**, 1183–1195 (2019).
31. M. Stocchero *et al*., PLS2 in metabolomics. *Metabolites* **9**, 51 (2019).
32. O. Libiger, N. J. Schork, Partial least squares regression can aid in detecting differential abundance of multiple features in sets of metagenomic samples. *Front. Genet.* **6**, 350 (2015).
33. P. Aarnio, P. Minkkinen, Application of partial least-squares modelling in the optimization of a waste-water treatment plant. *Anal. Chim. Acta* **191**, 457–460 (1986).

34. A. L. Amaral, E. C. Ferreira, Activated sludge monitoring of a wastewater treatment plant using image analysis and partial least squares regression. *Anal. Chim. Acta* **544**, 246–253 (2005).
35. M. Nierychlo *et al*., MiDAS 3: An ecosystem-specific reference database, taxonomy and knowledge platform for activated sludge and anaerobic digesters reveals species-level microbiome composition of activated sludge. *Water Res.* **182**, 115955 (2020).
36. M. Nierychlo *et al*., Candidatus Amarolinea and Candidatus Microthrix are mainly responsible for filamentous bulking in Danish municipal wastewater treatment plants. *Front. Microbiol.* **11**, 1214 (2020).
37. I. G. Chong, C. H. Jun, Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **78**, 103–112 (2005).
38. M. Henze, W. Gujer, T. Mino, M. C. M. van Loosdrecht, *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3* (IWA Publishing, 2000).
39. J. S. Griffin, G. F. Wells, Regional synchrony in full-scale activated sludge bioreactors due to deterministic microbial community assembly. *ISME J.* **11**, 500–511 (2017).
40. D. C. Vuono *et al*., Disturbance and temporal partitioning of the activated sludge metacommunity. *ISME J.* **9**, 425–435 (2015).
41. J. Heino *et al*., Metacommunity organisation, spatial extent and dispersal in aquatic systems: Patterns, processes and prospects. *Freshw. Biol.* **60**, 845–869 (2015).
42. E. S. Lindström, S. Langenheder, Local and regional factors influencing bacterial community assembly. *Environ. Microbiol. Rep.* **4**, 1–9 (2012).
43. S. L. McLellan, A. Roguet, The unexpected habitat in sewer pipes for the propagation of microbial communities and their imprint on urban waters. *Curr. Opin. Biotechnol.* **57**, 34–41 (2019).
44. S. Louca, M. Doebeli, Transient dynamics of competitive exclusion in microbial communities. *Environ. Microbiol.* **18**, 1863–1874 (2016).
45. S. Günther *et al*., Species-sorting and mass-transfer paradigms control managed natural metacommunities. *Environ. Microbiol.* **18**, 4862–4877 (2016).
46. Y. Hirakata *et al*., Effects of predation by protists on prokaryotic community function, structure, and diversity in anaerobic granular sludge. *Microbes Environ.* **31**, 279–287 (2016).
47. O. H. Shapiro, A. Kushmaro, A. Brenner, Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J.* **4**, 327–336 (2010).
48. D. C. Vuono, J. Munakata-Marr, J. R. Spear, J. E. Drewes, Disturbance opens recruitment sites for bacterial colonization in activated sludge. *Environ. Microbiol.* **18**, 87–99 (2016).
49. F. A. Meerburg *et al*., High-rate activated sludge communities have a distinctly different structure compared to low-rate sludge communities, and are less sensitive towards environmental and operational variables. *Water Res.* **100**, 137–145 (2016).
50. D. R. Nemergut *et al*., Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).
51. R. H. Kirkegaard *et al*., The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Sci. Rep.* **7**, 9343 (2017).
52. G. F. Wells *et al*., Microbial biogeography across a full-scale wastewater treatment plant transect: Evidence for immigration between coupled processes. *Appl. Microbiol. Biotechnol.* **98**, 4723–4736 (2014).
53. R. Mei, J. Kim, F. P. Wilson, B. T. W. Bocher, W.-T. Liu, Coupling growth kinetics modeling with machine learning reveals microbial immigration impacts and identifies key environmental parameters in a biological wastewater treatment process. *Microbiome* **7**, 1–9 (2019).
54. C. Jiang *et al*., Characterizing the growing microorganisms at species level in 46 anaerobic digesters at Danish wastewater treatment plants: A six-year survey on microbial community structure and key drivers. *Water Res.* **193**, 116871 (2021).
55. S. J. McIlroy *et al*., MiDAS: The field guide to the microbes of activated sludge. *Database (Oxford)* **2015**, bav062 (2015).
56. J. G. Caporaso *et al*., QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
57. D. J. Lane, "16S/23S rRNA Sequencing" in *Nucleic Acid Techniques in Bacterial Systematics*, E. Stackebrandt, M. Goodfellow, Eds. (John Wiley and Sons Ltd, 1991), pp. 115–175.
58. G. Muyzer, E. C. de Waal, A. G. Uitterlinden, Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* **59**, 695–700 (1993).
59. M. S. Dueholm *et al*., Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (Autotax). *mBio* **11**, e01557-20 (2020).
60. P. Yarza *et al*., Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
61. J. Oksanen *et al*., The vegan package. *Community Ecol. Packag.* **10**, 719 (2007).
62. S. Kucheryavskiy, mdatools – R package for chemometrics. *Chemom. Intell. Lab. Syst.* **198**, 103937 (2020).