Linking Population Dynamics to Microbial Kinetics for Hybrid Modeling of Engineered Bioprocesses

Zhang Cheng [a], Shiyun Yao [a], Heyang Yuan [a,*].

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

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

* **Corresponding author**

heyang.yuan@temple.edu;

Abstract

Mechanistic and data-driven models have been developed to provide predictive insights into the design and optimization of engineered bioprocesses. These two modeling strategies can be combined to form hybrid models to address the issues of parameter identifiability and prediction interpretability. Herein, we developed a novel and robust hybrid modeling strategy by incorporating microbial population dynamics into model construction. The hybrid model was constructed using bioelectrochemical systems (BES) as a platform system. We collected 77 samples from 13 publications, in which the BES were operated under diverse conditions, and performed holistic processing of the 16S rRNA amplicon sequencing data. Community analysis revealed core populations composed of putative electroactive taxa *Geobacter, Desulfovibrio, Pseudomonas*, and *Acinetobacter*. Primary Bayesian networks were trained with the core populations and environmental parameters, and directed Bayesian networks were trained by defining the operating parameters to improve the prediction interpretability. Both networks were validated with Bray-Curtis similarly, relative root-mean-square error (RMSE), and a null model. The hybrid model was developed by first building a three-population mechanistic component and subsequently feeding the estimated microbial kinetic parameters into network training. The hybrid model generated a simulated community that shared a Bray-Curtis similarity of 72% with the actual microbial community and an average relative RMSE of 7% for individual taxa. When examined with additional samples that were not included in network training, the hybrid model achieved accurate prediction of current production with a relative error-based RMSE of 0.8 and outperformed the data-driven models. The genomics-enabled hybrid modeling strategy represents a significant step toward robust simulation of a variety of engineered bioprocesses.

Keywords: Engineered bioprocess, Microbial population dynmics; Microbial kinetics; Machine learning, Bayesian network; Hybrid modeling.

1. Introduction

Engineered bioprocesses are widely applied to treat waste streams and recover valuable resources (Rittmann and McCarty 2012). To facilitate the design and optimization of full-scale bioprocesses, a number of mechanistic and data-driven models have been developed over the past 60 years (Batstone et al. 2002, Bhat and McAvoy 1990, Henze et al. 2000).

Mechanistic models can provide predictive insights into the fundamental processes in biological systems (Jeppsson 1996). The model structure has been improved with a greater understanding of the microbiomes in biological systems (Donoso-Bravo et al. 2011, Henze et al. 2000, Ng and Kim 2007). For example, the Anaerobic Digestion Model No.1 composed of 19 bioconversion steps and 100 parameters is by far the most comprehensive mechanistic model formulated for engineered bioprocesses and can be readily modified to simulate specific applications (Liu et al. 2017, Rodríguez et al. 2008, Zhao et al. 2019). An inherent problem of this modeling strategy is that the model structure and parameters are largely unidentifiable (Donoso-Bravo et al. 2011, Jeppsson 1996). This stems from the simulation of microbial kinetics, in which the functional populations cannot be fully recapitulated by the Monod expressions, and the associated kinetic parameters cannot be directly measured (Donoso-Bravo et al. 2011, Ng and Kim 2007). Although kinetic parameters can be derived from biochemical measurements, they show considerable variations under different operating conditions (Bernard et al. 2006, Ng and Kim 2007). As a result, a mechanistic model developed for a specific bioprocess needs constant parameter calibration to cope with operational perturbations but still falls short when applied to other biological systems.

Data-driven models are not limited by identifiability issues and can yield more accurate predictions than mechanistic models do when a sufficiently large data pool is provided (Walpole et al. 2017). Artificial neural networks can be constructed with appropriate input variables and network architecture to predict the effluent quality (Mendes et al. 2015, Moral et al. 2008). Recent studies have demonstrated the applicability of several machine learning algorithms, including support vector machine, random forest, extreme gradient boosting, and k-nearest neighbors, to full-scale anaerobic digesters (De Clercq et al. 2019, Wang et al. 2020b). Despite the outstanding learning performance, most of the data-driven models are black boxes that are unable to generate interpretable predictions (Rudin 2019). This is particularly problematic for complex biological systems whose performance is largely determined by the microbial population and activity, and thus no mechanistic insights can be obtained from the simulation. Training data-driven models with microbial population dynamics presents a promising solution to tackle this issue. Previous studies have incorporated genomic data into machine learning models (neural networks and Bayesian networks) to reconstruct microbial communities in natural ecosystems (Kuang et al. 2016, Larsen et al. 2012). Using similar strategies, an array of data-driven models was constructed to simulate the performance and stability of engineered bioprocesses (Lesnik et al. 2020, Lesnik and Liu 2017, Yuan et al. 2017).

Hybrid models can potentially address the limitations of those two modeling strategies (Cote et al. 1995, Karama et al. 2001, Zhao et al. 1997). A common approach is to couple the mechanistic and statistical components in series. The error signals obtained from the mechanistic component are converted into those for the network component, which are subsequently used to update the network weights through back-propagation (Lee et al. 2002). Through such integration, hybrid

71 models yield robust and semi-interpretable predictions for non-linear behaviors such as microbial
72 kinetics (Zendehboudi et al. 2018). By far, all hybrid models are built with physical and
73 biochemical parameters and thus unable to reveal the connections between microbial population
74 and microbial kinetics. We propose to link them and improve the prediction robustness by
75 incorporating genomic data into model construction.
76
77 The objectives of this study are to 1) comprehensively characterize the microbiome in a specific
78 type of bioprocess and 2) formulate a novel hybrid model based on the population dynamics and
79 microbial kinetics. To this end, we used bioelectrochemical systems (BES) as a platform system
80 for model construction. As emerging biotechnology that simultaneously achieves
81 water/wastewater treatment and energy/resource recovery (Logan et al. 2006, Wang and Ren 2013),
82 BES are ideal for this task because they respond quickly to environmental perturbations (Yuan et
83 al. 2016, Yuan et al. 2015), with current production acting as a sensitive indicator of the microbial
84 population and functional dynamics. The microbial communities in BES are highly enriched with
85 relatively low microbial diversity (Yates et al. 2012, Zhu et al. 2014), and hence represent a
86 desirable level of complexity: diverse enough to be relevant to the microbiomes in other
87 bioprocesses yet simple enough to be *in silico* reconstructed. To improve the compatibility of our
88 models, we performed an extensive literature review and collected the 16S rRNA gene amplicon
89 sequencing data from 77 samples in 13 publications, in which the BES were operated under a wide
90 range of conditions. Core populations were selected at different taxonomic levels and used to train
91 Bayesian networks, a machine learning model capable of characterizing the causal relationships
92 among variables (Uusitalo 2007). Meanwhile, microbial kinetics were calculated using a three-
93 population mechanistic component and fed into the training process to improve the prediction.

This hybrid modeling strategy is expected to take advantage of the rapid growth of the genomic database (Kahn 2011), circumvent the time-consuming calibration of the kinetic parameters, and be broadly applicable to a variety of engineered bioprocesses.

2. Materials and Methods

2.1 Data collection and sequence processing

A comprehensive literature review was carried out, and 13 publications containing 77 samples were selected for downstream community analysis and model construction. The detailed information about the selected publications is listed in the Supporting Information (SI) Table S1. The selection criteria are: 1) 16s rRNA gene amplicon sequences were properly deposited in the National Center for Biotechnology Information (NCBI) or DNA Data Bank of Japan (DDBJ) databases for holistic sequence processing; 2) the results reported eight key parameters closely related to BES performance, including substrate composition and concentration, coulombic efficiency (CE), pH, current, anode area, external resistance, hydraulic retention time, and temperature; 3) the results included the variation of chemical oxygen demand (COD) over time or currents/voltage that can be used to calculate the time series of COD. The selected studies show a variety of reactor configurations, operation modes, substrates, and operating conditions (SI Table S1), which is expected to enhance the compatibility of the predictive models developed in the present study.

The sequence data from the selected publications contained both pair-end or single-end reads and were converted into a uniform format before further processing. Briefly, pair-end reads were merged using Vsearch (Rognes et al. 2016), and the chimeric and low-quality sequences were

eliminated using the QIIME2 plug-in DADA2 (Callahan et al. 2016). The Greengenes database (gg_13_8 updated February 2011) was used to conduct sequence alignment and train the taxonomy classifier of the denoised sequences (DeSantis et al. 2006). In addition, due to the various primer-targeted regions (V1-V9) used to amplify 16s rRNA gene in those studies, the primer pair 8F/907R was set as the forward and reverse primers for the classifier to encompass all of the sequences from the samples.

2.2 Community analysis

Alpha (Shannon and Simpson indices) diversity analysis, principal coordinate analysis (PCoA), and redundancy analysis (RDA) were performed using R to unravel intra-sample diversity and inter-sample distance. Core populations were selected at the genus, order, and phylum levels with the following criteria (Yuan et al. 2019): 1) at least one occurrence in the 13 studies with the relative abundance ≥0.05% and 2) average relative abundance ≥2% across all 77 samples. The criteria allow us to retain the major functional populations in the microbial community and compress the genomic data for downstream model construction. For this reason, core population was not selected at the operational taxonomic unit (OTU) level as the 100%-similarity clustering strategy generated over 9,000 OTUs, and the majority of the samples have little overlap on the community composition. To build a phylogenetic tree for the core genera, the sequence of the most abundant OTU within a core genus was selected as a representative. The phylogenetic tree was built using ARB (Ludwig et al. 2004), and the *Silva* database (LTPs132_SSU.arb for 16s rDNA updated June 2018) was used for sequence alignment (Quast et al. 2013).

2.3 Bayesian network analysis

To prepare for network construction, the abundance of the core taxa and the values of the environmental parameters were scaled to 0 – 1 (Bishop 2013):

$$w_{n,i,j} = \frac{w_{i,j} - w_{j,min}}{w_{j,max} - w_{j,min}} \quad \text{(Eq. 1)}$$

where $w_{i,j}$ is the relative abundance of population *j* in sample *i*, $w_{j,max}$ is the maximum relative abundance of population *j*, $w_{j,min}$ is the minimum relative abundance of population *j*, and $w_{n,i,j}$ is the normalized abundance of population *j* in sample *i*. Structure learning and parameter learning were performed using the hill-climbing algorithm and maximum likelihood estimation, respectively (Scutari 2010).

Primary networks (SI Figures S4A, S5A, and S6A) were trained without considering the *a priori* knowledge about the operating conditions, and the node directions were solely inferred by the network algorithm. Directed networks (SI Figures S4B, S5B, and S6B) were trained by defining temperature, anode electrode area, external resistance, and hydraulic retention time as the parent nodes. This was because those parameters remained unchanged throughout the operation and thus were not affected by other parameters. A blacklist function was applied to define the unidirectional relationships between those operating parameters and other variables. The same training procedures were conducted at the genus (38 core taxa), order (32 core taxa), and phylum (13 core taxa) levels. Considering the sample quantity and computational cost, a leave-one-out cross-validation strategy was selected for three validation methods (Bro et al. 2008): Bray-Curtis similarity between the predicted and observed microbial community, relative root-mean-square error (RMSE, Eq. 2), and null model analysis.

$$relative\ RMSE = \frac{\sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}}}{y_{max}} \quad \text{(Eq. 2)}$$

where $\hat{y}_i$ is the predicted value; y_i is the observed value; y_{max} is the maximum observed value; and *n* is the number of samples. Null models were constructed by setting the abundance of all core taxa to the average abundance across all samples (Gotelli 2002).

2.4 Hybrid model construction

A hybrid model was constructed at the genus level to bridge population dynamics and microbial kinetics following the procedures shown in SI Figure S1. Briefly, maximum substrate utilization rates, maximum growth rates, and mediator yield were calculated using the mechanistic component described below and included as the nodes for network training. Although these kinetic parameters are unmeasurable, the normalization step (Eq. 1) converts exact values into a general tendency of microbial activity that can be statistically connected to the actual relative abundance of the core population (Weissman et al. 2021), and thus the estimated kinetic parameters do not need to be validated.

The mechanistic component assumes two-step degradation (SI Figure S2) of the substrates by three populations (Pinto et al. 2011). At the first stage, complex organic matters such as polysaccharides, lipids, and proteins are decomposed by primary degraders to low-molecular intermediate products including acetic acid, propionic acid, ethanol, etc. At the second stage, electroactive and non-electroactive microbes convert the intermediates into electrical energy and methane, respectively. Take electroactive microbes as an example, the mass balance for the growth and activity are described using Eq. 3 and Eq. 4, respective:

$$\frac{dX_{ee}}{dt} = \mu_{ee}\frac{S_1}{K_{ee}+S_1}\frac{M_{ox}}{K_M+M_{ox}}X_{ee} - k_{dee}X_{ee} - D\frac{1+\tanh\left(f_{ee}\left(X_{ee}+X_{ne}-X_{ee,max}\right)\right)}{2}X_{ee} \quad \text{(Eq. 3)}$$

$$\frac{dM_{ox}}{dt} = -Y_M \cdot k_{ee} \frac{S_1}{K_{ee} + S_1} \frac{M_{ox}}{K_M + M_{ox}} + \frac{\gamma \cdot I}{V_a \cdot F \cdot X_{ee} n_e} \quad \text{(Eq. 4)}$$

where X_{ee} and M_{ox} are the concentrations of electroactive microbes and mediator, respectively; K_{ee} and K_M are the substrate half-saturation constant and mediator half-saturation constant, respectively; μ_{ee} and $k_{d,ee}$ are the growth and decay rates, $X_{ee,max}$ is the maximum capacity of electroactive microbes in the anode; Y_M is the mediator yield; γ is the molecular mass of mediator; I is the current production; V_a is the anode volume; F is the Faraday constant (A/d·mol); and n_e is the number of electrons transfer. The mass balance was modified based on the multiplicative Monod expressions from previous studies (Ping et al. 2014, Pinto et al. 2010). The detailed formulation of the mechanistic component and the parameters can be found in the SI Method and Table S2. Because the selected publications provided multiple types of data and operated the reactors under distinct conditions, the mechanistic component was slightly modified for individual reactors (SI Table S3), and Literature #27 (S27) was not considered for mechanistic modeling because a time series of COD was not available.

The maximum substrate utilization rate and maximum growth rate are considered to be the most critical values for simulation of engineered bioprocesses (Rittmann and McCarty 2012), and mediator yield is a unique parameter for BES. Those kinetic parameters were estimated with specific limits according to previous studies while others parameters were retrieved from literature (Kato Marcus et al. 2007, Wilson and Kim 2016). To estimate the parameters, the total substrate concentration over time and the initial values of the kinetic parameters were collected, calculated, or estimated from the selected papers. Because biomass and mediator concentrations were not measurable and unavailable in the literature, some assumptions are made: 1) primary degraders have an initial concentration of 100 mg/L in all reactors, 2) electroactive and non-electroactive

207 microbe grow evenly on the anode surface at an average thickness of 60 μm (Lee et al. 2009,
208 Torres et al. 2008), 3) the maximum biofilm capacity in BES was assumed to be 600 mg/L (Ping
209 et al. 2014), and 4) the electroactive microbe fraction is proportional to the CE. For microbial
210 electrolysis cells whose CE was higher than 100% because of the applied voltage, the
211 concentration of electroactive microbes was assumed to be 500 mg/L. Because the dimension of
212 the anode electrode was not directly provided in some of the studies, the area was estimated based
213 on the specific area and size of the electrode (Logan et al. 2007).
214

215 **2.5 Model prediction**

216 Six additional publications were collected to further demonstrate the robustness of the hybrid
217 model (SI Table S7). In the first step, the operating parameters (i.e., temperature, anode electrode
218 area, external resistance, and hydraulic retention time) and pH from those studies were input into
219 the Bayesian networks to predict current and CE directly. In the second step, the kinetic parameters
220 obtained from the Bayesian network were input into the mechanistic component to calculate the
221 steady-state COD, which was then converted into current production based on the expression of
222 CE (Logan et al. 2006). The predicted and observed values were compared using an RMSE based
223 on relative errors (Walpole et al. 2017):

224 $$relative\ error\ based\ RMSE = \sqrt{\frac{\sum_1^n \left(\frac{\hat{y}_i - y_i}{y_i}\right)^2}{n}} \quad \text{(Eq. 5)}$$

225

226 **3. Results and Discussion**

227 **3.1 BES operated under a variety of conditions**

The 13 selected studies contained 39 samples from microbial fuel cells, 31 samples from microbial electrolysis cells with external voltage input, and 7 samples from microbial desalination cells with saline environments (SI Table S1). In addition to different reactor configurations, the bioreactors were fed with a variety of substrates whose main organic matter could be categorized as non-fermentable (acetate and methanol), fermentable (glucose, ethanol, and propyl alcohol), or complex (brewery wastewater, food waste, and pig slurry). The selected studies also presented multiple operation modes including batch, continuous, and continuous with pulse substrate loading. In terms of the performance, the 77 samples showed significant difference (SI Table S4) in CE which ranged from 0.02% (due to high external resistance, e.g., S22) to 140% (due to applied voltage, e.g., S37). Overall, the selected samples have included the conditions commonly found in BES studies and are expected to yield representative core populations and models.

3.2 Core population in BES

The 2.6 million sequence reads from the 77 samples resulted in approximately 9600 OTUs, based on which the alpha diversity was analyzed. Although the diversity indices (Shannon, Simpson, and Chao1, SI Table S5) did not follow a normal distribution as revealed by the Shapiro-Wilk analysis results ($p < 0.05$), it could well reflect the variation in operating conditions. For example, the Shannon and Simpson indices for the samples fed with complex substrates (i.e., S15, S17, S20, S27, and S42) were 4.25 and 0.96, respectively, significantly higher than the 2.24 and 0.73 of the samples fed with non-fermentable substrates. Similar results were reported in previous studies (Wang et al. 2020a), and a highly diverse microbial community was expected to enhance system stability (Girvan et al. 2005). In addition, the majority of the samples fed with non-fermentable and fermentable substrates showed a Chao1 index of approximately 100, indicating that those

BESs were sufficiently sampled, and the key microbes could be captured when selecting the core population. This is confirmed by the rarefaction curves presented in some of the selected publications.

PCoA based on Bray-Curtis distance showed a critical role of substrate composition in microbial community assembly (Figure 1). Specifically, samples cultivated with starch- and yeast extraction-based synthetic wastewater (S20 and S22) were found in the top right corner of the PCoA graph, while those with complex food waste, brewery wastewater, and pig slurry (S15, S17, S42, and S51) were clustered in the center. S48 used ethanol as the sole carbon source and was isolated from other samples. Additionally, the anode area appeared to be an important factor that drove the microbial community assembly in S48 (SI Figure S3), which was amended with granular activated carbon in the anode. Another key deterministic factor of microbial community assembly is the seed source. Unlike other studies, S49 was inoculated with activated sludge and formed a distinct community structure. Similarly, activated sludge was the seed of S35 and together with temperature (SI Figure S3) led to communities significantly different from other studies. In summary, varied substrate composition, reactor configuration, and operation mode provided a comprehensive pool of microbial communities for model construction.

Core populations were selected at different taxonomic levels based on the occurrence (at least one occurrence in the 13 studies with the abundance ≥0.05%) and abundance (≥2% across all 77 samples) (Ling et al. 2016, Saunders et al. 2016). At the genus level, 38 core taxa were identified, accounting for 55% of the abundance on average. The selection criteria were considered stringent given that the bioreactors were operated under distinct conditions, and the microbial communities

were highly diverse. This was reflected by the loss of several abundant taxa in specific BES. For instance, the core genera made up of less than 20% of the abundance in some of the samples due to the unique flow pattern (plug-flow, S22), substrate (propyl alcohol, S27), and reactor configuration (applied voltage, S37). Nonetheless, the core population included some well-characterized genera such as *Geobacter*, *Desulfovibrio*, *Pseudomonas*, and *Acinetobacter*, which were frequently found abundant in BES and potentially involved in current production.

The presence of *Geobacter* often serves as the indicator to explain the BES performance, in particular current production, because a few members of this genus are highly efficient in extracellular electron transfer (EET) (Logan et al. 2019, Lovley et al. 2011). Indeed, *Geobacter* was identified to be a core taxon (G29, Figure 2) with an average abundance of 15% across all samples and an individual abundance higher than 2% in 39 samples. This genus was dominant in S20 (17%), S35 (71%), S48 (27%), and some of the samples in S15 (9%) and S49 (8%), which all showed a CE over 15%. However, *Geobacter* was not present in other high-CE samples likely because the operating conditions (e.g., high salinity in S12 and S54) did not favor its growth (Miyahara et al. 2015). *Desulfovibrio* spp. from the same class of Deltaproteobacteria are common sulfate-reducing bacteria whose EET ability has also been reported (Aulenta et al. 2012, Gacitúa et al. 2014, Yu et al. 2011). This genus (G28, Figure 2) was abundant in 13 samples (>2%) and dominant in ethanol-fed S48 (27%). *Desulfovibrio* is known to oxidize ethanol with sulfate as the electron acceptor. In the absence of sulfate, *Desulfovibrio* can still grow syntrophically with methanogens and oxidize the ethanol to acetate through interspecies hydrogen transfer (Hensgens et al. 1993, Kremer et al. 1988).

297 *Pseudomonas*, a well-studied genus forming biofilm in many anaerobic environments , is another
298 core taxon that has been reported to carry out EET by using phenazines as an electron shuttle
299 (Rabaey et al. 2004). *Pseudomonas* (G33, Figure 2) was widely present in S12, S15, S22 S35, S49,
300 and S54. This genus potentially plays a critical role in shaping the microbial community structure
301 as the phenazines actively produced for quorum sensing can be scavenged for electron shuttling
302 by other species such as *Acinetobacter*. As shown in Figure 2, *Acinetobacter* (G32) was found in
303 samples where *Pseudomonas* was abundant and was previously speculated to utilize phenazines
304 as electron shuttles for EET (Liu et al. 2013, Yuan et al. 2017). Because EET via electron shuttles
305 is limited by diffusion and the conductivity of the anolyte (Torres et al. 2010), *Pseudomonas* and
306 its metabolic partners are less ubiquitous than *Geobacter* in BES, and *Pseudomonas*-dominated
307 communities are more likely to be found in specific environments such as microbial desalination
308 cells with a high ion concentration (Yuan et al. 2017).
309
310 Several genera from the class Bacteroidia were found to be abundant (Figure 2). Previous studies
311 suggested that this group of bacteria could degrade complex organic compounds including proteins,
312 polysaccharides, and pectins (Dongowski et al. 2000, Grenier et al. 1989). For instance,
313 *Parabacteroides* (G6 & G11, Figure 2) show an abundance higher than 5% in most of the samples
314 fed with complex organic. It has also been reported that some members in the class Bacteroidia
315 degrade biomass and can serve as scavengers of dead cells (Madigan 2014, Reichenbach 1992).
316 Those taxa might act as degraders of soluble organic matter and provide electroactive microbes
317 simple substrates (Tan et al. 2012, Zeppilli et al. 2020). In addition to fermentative bacteria,
318 methanogens were observed with considerable abundance in S48 (Figure 2) and possibly carried

out syntrophic electron transfer with ethanol-consuming in the presence of conductive granular activated carbon (Yuan et al. 2018).

Overall, microbial community analysis demonstrated a core BES population composed of primary fermentative bacteria that convert complex organic matter to simple electron donors, and electroactive microbes and their competitors (e.g., methanogens) growing on the fermentative products. The results thus justify the modeling of BES based on those three guilds (Pinto et al. 2011). However, such a model structure is incapable of differentiating the contribution of different types of electroactive microbes and their EET mechanisms (i.e., direct contact vs. electron shuttling) due to the experimental challenge to measure the associated biochemical parameters (e.g., the concentration of phenazines and other electron shuttles). The same pitfall is also found in mechanistic modeling of activated sludge and anaerobic digestion, in which the core populations consist of functionally redundant taxa occupying the same ecological niches (Ju and Zhang 2015). This explains the constant parameter calibration required by mechanistic models. To address the issue and improve the prediction robustness, a new modeling approach is imperative.

3.3 Reconstruction of microbial community

Two types of Bayesian networks were trained with the same dataset containing environmental parameters and relative abundance of the core populations (SI Figure S4-S6): primary networks whose node directions were not restricted by *a priori* knowledge and directed networks in which the operating parameters were set as the parent nodes. The latter were constructed based on the fact that operating parameters such as external resistance and hydraulic retention time remained

unchanged throughout the operation and hence should be not affected by microbial community dynamics and system performance.

To validate the modeling approach and evaluate the prediction of the community structure, Bray-Curtis similarity between the predicted and observed communities was calculated. At the genus level, the directed network achieved the most accurate prediction, followed by the primary network and a null model (Bray-Curtis similarity: $0.72 > 0.64 > 0.52$, $p < 0.05$, SI Figure S7A). Similar trends were observed at the order and phylum levels, but the prediction accuracy did not show consistent improvement as the taxonomic level increased. The Bray-Curtis similarity from the directed networks dropped slightly to 0.61 at the order level and raised back to 0.79 at the phylum level. The results were not in agreement with the previous findings that prediction accuracy could be continuously improved by training data-driven models at higher taxonomic levels (Kuang et al. 2016, Yuan et al. 2017). It should be noted that those models were built based on highly specific environments and communities (e.g., acid mine drainage and microbial desalination cells), whereas the models in the present study considered a variety of environmental conditions, which might statistically compromise the model robustness (Walpole et al. 2017).

To further validate the modeling approach, relative RMSE was calculated for the 6 environmental parameters and 38 core genera (SI Figure S8-S10). At the genus level, the RMSE values from the directed and primary network were 2% - 17% and 3% - 24%, respectively. The abundances of putative electroactive taxa *Desulfovibrio*, *Pseudomonas*, and *Acinetobacter* were well estimated by both networks with the RMSE ranging from 4% to 10%. On the other hand, the RMSE for *Geobacter* was improved from 24% with the primary network to 16% with the directed network.

364 The poor prediction of *Geobacter* is likely because some members from this genus, despite their
365 dominance in many anaerobic environments (Lee et al. 2016, Lin et al. 2017), are inefficient in or
366 incapable of EET (Lovley et al. 2011, Rotaru et al. 2015). Similar to Bray-Curtis similarity, RMSE
367 was improved at the phylum but not at the order level (SI Figure S9 and S10). The phylum
368 Proteobacteria, which includes the putative electroactive taxa discussed above, is estimated with
369 the highest accuracy (relative RMSE <1%). Overall, the more accurate prediction from the directed
370 networks at all three taxonomic levels suggests that the modeling approach can be enhanced by
371 introducing reasonable structure control.
372
373 After the modeling approach was validated with Bray-Curtis similarity and RMSE, final networks
374 were constructed from the whole dataset to infer microbial interactions (SI Figure S4-S6). In the
375 genus-level networks, putative electroactive taxa *Geobacter* (G29), *Desulfovibrio* (G28), and
376 *Pseudomonas* (G33) did not show any association with CE and current, whilst *Acinetobacter* (G32)
377 was not correlated with the system output. The networks at higher taxonomic levels yielded even
378 less interpretable inference regarding the potential functions of the core taxa. For example,
379 methanogens were predicted to be more related to current production than Proteobacteria (SI
380 Figure S6A). The results collectively indicate that more *a priori* knowledge needs to be included
381 in model training to improve the robustness and interpretability of the inference.
382
383 **3.4 Hybrid modeling of BES performance**
384 To build a hybrid model, the rates for substrate utilization and microbial growth and mediator yield
385 were first estimated using the three-population mechanistic component (SI Table S4). Some of the
386 estimates were constant during calibration (e.g., 15 /d for substrate utilization rate) because they

were the boundary values determined based on previous studies (Bruce and Perry 2001, Kato Marcus et al. 2007, Wilson and Kim 2016). The estimated microbial kinetic parameters were subsequently included in the training dataset to construct a hybrid network at the genus level (Figure 3). It should be noted that the scaled kinetic parameters represent the trend of microbial activity and thus do not require accurate estimation or validation. A whitelist function was further applied to force putative electroactive genera *Geobacter, Desulfovibrio, Pseudomonas*, and *Acinetobacter*to to directly affect current generation and improve the prediction interpretability. The hybrid network yielded a simulated community that shared a Bray-Curtis similarity of 0.72 with the actual genera-level core population, which was comparable to the result from the directed network and significantly better than that of the null model (t-test, $p < 0.05$). The relative RMSE of the hybrid model ranging from 3% to 18% was also similar to those from the directed network.

The hybrid network generated reasonable inference of the relationships between microbial population and kinetics (Figure 3), as evidenced by the strong positive correlation of the EET-related substrate utilization rate with *Desulfovibrio* (coefficient = 0.86), as well as the positive correlation of mediator yield with *Pseudomonas*. Meanwhile, mediator yield was negatively related to glucose, likely because fermentable substrates could lead to significant electron loss (Parameswaran et al. 2010). The EET-related substrate utilization and growth rates were both associated with the genera (G6 & G8) from the class Bacteroidia. As discussed above, those taxa can degrade dead cells and soluble microbial products, thereby creating a favorable environment for electroactive microbes (Ni et al. 2011, Ni et al. 2010). Despite those biologically sound inferences, the hybrid model still contained unexplainable interactions such as the negative

association between the EET-related substrate utilization rate and anode area, underpinning the elimination of data-driven models in prediction interpretability.

The developed models were examined with six new samples that were not included in network training, and the hybrid model (hybrid network + mechanistic component) achieved the most accurate prediction of current production compared with the data-driven models. It can be seen from Figure 4 that the predicted results from the hybrid model agree well with the experimental values with slight deviation at the high current range. The low relative error-based RMSE of 0.8 further indicates outstanding prediction accuracy throughout the examined current range. The hybrid network alone loses the prediction power at high current, resulting in a higher relative error-based RMSE of 6.7, whereas the directed network is incapable of generating satisfactory prediction and shows the highest relative error-based RMSE of 16.3. The significantly improved prediction performance of the hybrid model likely stems from the close connection between population dynamics and microbial kinetics. Under a given condition, each population (either a single species or a functional guild) has specific maximum substrate utilization and growth rates that are largely determined by its unique lifestyle and ecophysiology (Rittmann and McCarty 2012), which can thus be statistically inferred from the genomic data (Weissman et al. 2021). On the other hand, system performance such as current production is affected by not only microbial population and activity, but also many other operating parameters including electrolyte conductivity and external resistance. Data-driven models that infer system performance directly from the microbial population do not consider the contribution of those operating parameters and hence cannot consistently yield accurate predictions.

Despite the robust performance, the hybrid model is not ready for practical implementation as accurate prediction can only be obtained with microbial community information as the input. When the data-driven component is fed solely with operating parameters, and the simulated microbial community serves as the intermediate to estimate the kinetic parameters, the prediction error quickly builds up along the inference, causing considerable uncertainty to the final prediction. Another challenge is that inadequate biochemical and sequencing data from the selected publications compromise the compatibility of both the data-driven and mechanistic components. These issues will be addressed in future studies with proper experimental design and alternative machine learning algorithms such as neural networks and random forest. Ultimately, the hybrid modeling approach is expected to be broadly applicable to various engineered bioprocesses including anaerobic digesters, activated sludge processes, anaerobic ammonium oxidation, etc.

4. Conclusion

We collected 77 samples from 13 studies in which the BES were operated under diverse conditions. Community analysis revealed a core population composed of primary fermentative bacteria, putative electroactive taxa *Geobacter*, *Desulfovibrio*, *Pseudomonas*, and *Acinetobacter*, as well as non-electroactive microbes such as methanogens. Bayesian networks were trained with the core populations and validated with Bray-Curtis similarity, relative RMSE, and a null model, all based on a leave-one-out cross-validation strategy. A hybrid model was built by combining mechanistic modeling and network training and achieved more accurate prediction of current production than data-driven models. This study provides insights into incorporating genomic data into hybrid modeling for robust and interpretable prediction.

Acknowledgement

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

457 **Reference**
458
459 Aulenta, F., Catapano, L., Snip, L., Villano, M. and Majone, M. (2012) Linking Bacterial
460 Metabolism to Graphite Cathodes: Electrochemical Insights into the H2-Producing Capability of
461 Desulfovibrio sp. ChemSusChem 5(6), 1080-1085.

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

465 Bernard, O., Chachuat, B., Hélias, A. and Rodriguez, J. (2006) Can we assess the model
466 complexity for a bioprocess: theory and example of the anaerobic digestion process. Water
467 Science and Technology 53(1), 85-92.

468 Bhat, N. and McAvoy, T.J. (1990) Use of neural nets for dynamic modeling and control of
469 chemical process systems. Computers & Chemical Engineering 14(4-5), 573-582.

470 Bishop, C.M. (2013) Neural networks for pattern recognition, Oxford University Press, Oxford.

471 Bro, R., Kjeldahl, K., Smilde, A.K. and Kiers, H.A. (2008) Cross-validation of component
472 models: a critical look at current methods. Anal Bioanal Chem 390(5), 1241-1251.

473 Bruce, E.R. and Perry, L.M. (2001) Environmental Biotechnology: Principles and Applications,
474 McGraw-Hill Education, New York.

475 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P.
476 (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods
477 13(7), 581-583.

478 Cote, M., Grandjean, B.P., Lessard, P. and Thibault, J. (1995) Dynamic modelling of the
479 activated sludge process: improving prediction using neural networks. Water research 29(4),
480 995-1004.

481 De Clercq, D., Jalota, D., Shang, R., Ni, K., Zhang, Z., Khan, A., Wen, Z., Caicedo, L. and
482 Yuan, K. (2019) Machine learning powered software for accurate prediction of biogas
483 production: A case study on industrial-scale Chinese production data. Journal of Cleaner
484 Production 218, 390-399.

485 DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T.,
486 Dalevi, D., Hu, P. and Andersen, G.L. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene
487 Database and Workbench Compatible with ARB. Applied and Environmental Microbiology
488 72(7), 5069.

489 Dongowski, G., Lorenz, A. and Anger, H. (2000) Degradation of Pectins with Different Degrees
490 of Esterification by Bacteroides thetaiotaomicron Isolated from Human
491 Gut Flora. Applied and Environmental Microbiology 66(4), 1321.

Donoso-Bravo, A., Mailier, J., Martin, C., Rodríguez, J., Aceves-Lara, C.A. and Wouwer, A.V. (2011) Model selection, identification and validation in anaerobic digestion: a review. Water research 45(17), 5347-5364.

Gacitúa, M.A., González, B., Majone, M. and Aulenta, F. (2014) Boosting the electrocatalytic activity of Desulfovibrio paquesii biocathodes with magnetite nanoparticles. International Journal of Hydrogen Energy 39(27), 14540-14545.

Girvan, M.S., Campbell, C.D., Killham, K., Prosser, J.I. and Glover, L.A. (2005) Bacterial diversity promotes community stability and functional resilience after perturbation. Environmental Microbiology 7(3), 301-313.

Gotelli, N.J., & Graves, G. R. (2002) Null models in ecology, Smithsonian Institution Press, Washington.

Grenier, D., Mayrand, D. and McBride, B.C. (1989) Further studies on the degradation of immunoglobulins by black-pigmented Bacteroides. Oral Microbiology and Immunology 4(1), 12-18.

Hensgens, C.M., Vonck, J., Van Beeumen, J., van Bruggen, E.F. and Hansen, T.A. (1993) Purification and characterization of an oxygen-labile, NAD-dependent alcohol dehydrogenase from Desulfovibrio gigas. Journal of Bacteriology 175(10), 2859.

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

Jeppsson, U. (1996) Modelling aspects of wastewater treatment processes.

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

Kahn, S.D. (2011) On the future of genomic data. Science 331(6018), 728-729.

Karama, A., Bernard, O., Gouzé, J., Benhammou, A. and Dochain, D. (2001) Hybrid neural modelling of an anaerobic digester with respect to biological constraints. Water Science and Technology 43(7), 1-8.

Kato Marcus, A., Torres, C.I. and Rittmann, B.E. (2007) Conduction-based modeling of the biofilm anode of a microbial fuel cell. Biotechnology and Bioengineering 98(6), 1171-1182.

Kremer, D.R., Nienhuis-Kuiper, H.E. and Hansen, T.A. (1988) Ethanol dissimilation in Desulfovibrio. Archives of Microbiology 150(6), 552-557.

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

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

527 Lee, D.S., Jeon, C.O., Park, J.M. and Chang, K.S. (2002) Hybrid neural network modeling of a
528 full-scale industrial wastewater treatment process. Biotechnology and Bioengineering 78(6),
529 670-682.

530 Lee, H.-S., Torres, C.I. and Rittmann, B.E. (2009) Effects of Substrate Diffusion and Anode
531 Potential on Kinetic Parameters for Anode-Respiring Bacteria. Environmental Science &
532 Technology 43(19), 7571-7577.

533 Lee, J.-Y., Lee, S.-H. and Park, H.-D. (2016) Enrichment of specific electro-active
534 microorganisms and enhancement of methane production by adding granular activated carbon in
535 anaerobic reactors. Bioresource Technology 205, 205-212.

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

538 Lesnik, K.L. and Liu, H. (2017) Predicting Microbial Fuel Cell Biofilm Communities and
539 Bioreactor Performance using Artificial Neural Networks. Environmental Science & Technology
540 51(18), 10881-10892.

541 Lin, R., Cheng, J., Zhang, J., Zhou, J., Cen, K. and Murphy, J.D. (2017) Boosting biomethane
542 yield and production rate with graphene: The potential of direct interspecies electron transfer in
543 anaerobic digestion. Bioresource Technology 239, 345-352.

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

547 Liu, H., Ishikawa, M., Matsuda, S., Kimoto, Y., Hori, K., Hashimoto, K. and Nakanishi, S.
548 (2013) Extracellular Electron Transfer of a Highly Adhesive and Metabolically Versatile
549 Bacterium. ChemPhysChem 14(11), 2407-2412.

550 Liu, Y., Zhang, Y., Zhao, Z., Ngo, H.H., Guo, W., Zhou, J., Peng, L. and Ni, B.-J. (2017) A
551 modeling approach to direct interspecies electron transfer process in anaerobic transformation of
552 ethanol to methane. Environmental Science and Pollution Research 24(1), 855-863.

553 Logan, B., Cheng, S., Watson, V. and Estadt, G. (2007) Graphite Fiber Brush Anodes for
554 Increased Power Production in Air-Cathode Microbial Fuel Cells. Environmental Science &
555 Technology 41(9), 3341-3346.

556 Logan, B.E., Hamelers, B., Rozendal, R., Schröder, U., Keller, J., Freguia, S., Aelterman, P.,
557 Verstraete, W. and Rabaey, K. (2006) Microbial fuel cells: methodology and technology.
558 Environmental Science & Technology 40(17), 5181-5192.

559 Logan, B.E., Rossi, R., Ragab, A.a. and Saikaly, P.E. (2019) Electroactive microorganisms in
560 bioelectrochemical systems. Nature Reviews Microbiology 17(5), 307-319.

561 Lovley, D.R., Ueki, T., Zhang, T., Malvankar, N.S., Shrestha, P.M., Flanagan, K.A., Aklujkar,
562 M., Butler, J.E., Giloteaux, L., Rotaru, A.-E., Holmes, D.E., Franks, A.E., Orellana, R., Risso, C.
563 and Nevin, K.P. (2011) Advances in Microbial Physiology. Poole, R.K. (ed), pp. 1-100,
564 Academic Press.

565 Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T.,
566 Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S.,
567 Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B.,
568 Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A. and
569 Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Research
570 32(4), 1363-1371.

571 Madigan, M., Bender, K., Buckley, D., Sattley, W. and Stahl, D. (2014) Brock biology of
572 microorganisms Pearson, Boston.

573 Mendes, C., da Silva Magalhes, R., Esquerre, K. and Queiroz, L.M. (2015) Artificial Neural
574 Network Modeling for Predicting Organic Matter in a Full-Scale Up-Flow Anaerobic Sludge
575 Blanket (UASB) Reactor. Environmental Modeling & Assessment 20(6), 625-635.

576 Miyahara, M., Kouzuma, A. and Watanabe, K. (2015) Effects of NaCl concentration on anode
577 microbes in microbial fuel cells. AMB Express 5(1), 34.

578 Moral, H., Aksoy, A. and Gokcay, C.F. (2008) Modeling of the activated sludge process by
579 using artificial neural networks with automated architecture screening. Computers & Chemical
580 Engineering 32(10), 2471-2478.

581 Ng, A.N.L. and Kim, A.S. (2007) A mini-review of modeling studies on membrane bioreactor
582 (MBR) treatment for municipal wastewaters. Desalination 212(1), 261-281.

583 Ni, B.-J., Rittmann, B.E. and Yu, H.-Q. (2011) Soluble microbial products and their implications
584 in mixed culture biotechnology. TRENDS in Biotechnology 29(9), 454-463.

585 Ni, B.-J., Zeng, R.J., Fang, F., Xie, W.-M., Sheng, G.-P. and Yu, H.-Q. (2010) Fractionating
586 soluble microbial products in the activated sludge process. Water research 44(7), 2292-2302.

587 Parameswaran, P., Zhang, H., Torres, C.I., Rittmann, B.E. and Krajmalnik-Brown, R. (2010)
588 Microbial community structure in a biofilm anode fed with a fermentable substrate: The
589 significance of hydrogen scavengers. Biotechnology and Bioengineering 105(1), 69-78.

590 Ping, Q., Zhang, C., Chen, X., Zhang, B., Huang, Z. and He, Z. (2014) Mathematical Model of
591 Dynamic Behavior of Microbial Desalination Cells for Simultaneous Wastewater Treatment and
592 Water Desalination. Environmental Science & Technology 48(21), 13010-13019.

593 Pinto, R.P., Srinivasan, B., Escapa, A. and Tartakovsky, B. (2011) Multi-Population Model of a
594 Microbial Electrolysis Cell. Environmental Science & Technology 45(11), 5039-5046.

595 Pinto, R.P., Srinivasan, B., Manuel, M.F. and Tartakovsky, B. (2010) A two-population bio-
596 electrochemical model of a microbial fuel cell. Bioresource Technology 101(14), 5256-5265.

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.

Rabaey, K., Boon, N., Siciliano, S.D., Verhaege, M. and Verstraete, W. (2004) Biofuel Cells Select for Microbial Consortia That Self-Mediate Electron Transfer. Applied and Environmental Microbiology 70(9), 5373.

Reichenbach, H. (1992) The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications. Balows, A., Trüper, H.G., Dworkin, M., Harder, W. and Schleifer, K.-H. (eds), pp. 3631-3675, Springer New York, New York, NY.

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

Rodríguez, J., Roca, E., Lema, J.M. and Bernard, O. (2008) Determination of the adequate minimum model complexity required in anaerobic bioprocesses using experimental data. Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology 83(12), 1694-1702.

Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584.

Rotaru, A.-E., Woodard, T.L., Nevin, K.P. and Lovley, D.R. (2015) Link between capacity for current production and syntrophic growth in Geobacter species. Frontiers in Microbiology 6(744).

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5), 206-215.

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. (2010) Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software. Journal of Statistical Software 35(3), 1-22.

Tan, H.-Q., Li, T.-T., Zhu, C., Zhang, X.-Q., Wu, M. and Zhu, X.-F. (2012) Parabacteroides chartae sp. nov., an obligately anaerobic species from wastewater of a paper mill. International Journal of Systematic and Evolutionary Microbiology 62(Pt_11), 2613-2617.

Torres, C.I., Marcus, A.K., Lee, H.-S., Parameswaran, P., Krajmalnik-Brown, R. and Rittmann, B.E. (2010) A kinetic perspective on extracellular electron transfer by anode-respiring bacteria. FEMS Microbiology Reviews 34(1), 3-17.

Torres, C.I., Marcus, A.K., Parameswaran, P. and Rittmann, B.E. (2008) Kinetic Experiments for Evaluating the Nernst−Monod Model for Anode-Respiring Bacteria (ARB) in a Biofilm Anode. Environmental Science & Technology 42(17), 6593-6597.

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

634 Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K. (2017) Probability & statistics for engineers
635 & scientists, Pearson Education South Asia Pte, Singapore.

636 Wang, B., Liu, W., Zhang, Y. and Wang, A. (2020a) Bioenergy recovery from wastewater
637 accelerated by solar power: Intermittent electro-driving regulation and capacitive storage in
638 biomass. Water Research 175, 115696.

639 Wang, H. and Ren, Z.J. (2013) A comprehensive review of microbial electrochemical systems as
640 a platform technology. Biotechnology Advances 31(8), 1796-1807.

641 Wang, L., Long, F., Liao, W. and Liu, H. (2020b) Prediction of anaerobic digestion performance
642 and identification of critical operational parameters using machine learning algorithms.
643 Bioresource Technology 298, 122495.

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

647 Wilson, E.L. and Kim, Y. (2016) The yield and decay coefficients of exoelectrogenic bacteria in
648 bioelectrochemical systems. Water research 94, 233-239.

649 Yates, M.D., Kiely, P.D., Call, D.F., Rismani-Yazdi, H., Bibby, K., Peccia, J., Regan, J.M. and
650 Logan, B.E. (2012) Convergent development of anodic bacterial communities in microbial fuel
651 cells. The ISME Journal 6(11), 2002-2013.

652 Yu, L., Duan, J., Zhao, W., Huang, Y. and Hou, B. (2011) Characteristics of hydrogen evolution
653 and oxidation catalyzed by Desulfovibrio caledoniensis biofilm on pyrolytic graphite electrode.
654 Electrochimica Acta 56(25), 9041-9047.

655 Yuan, H.-Y., Ding, L.-J., Zama, E.F., Liu, P.-P., Hozzein, W.N. and Zhu, Y.-G. (2018) Biochar
656 Modulates Methanogenesis through Electron Syntrophy of Microorganisms with Ethanol as a
657 Substrate. Environmental Science & Technology 52(21), 12198-12207.

658 Yuan, H., Abu-Reesh, I.M. and He, Z. (2016) Mathematical Modeling Assisted Investigation of
659 Forward Osmosis as Pretreatment for Microbial Desalination Cells to Achieve Continuous Water
660 Desalination and Wastewater Treatment. Journal of Membrane Science 502, 116–123.

661 Yuan, H., Lu, Y., Abu-Reesh, I. and He, Z. (2015) Bioelectrochemical production of hydrogen
662 in an innovative pressure-retarded osmosis/microbial electrolysis cell system: experiments
663 and modeling. Biotechnology for Biofuels 8(116), 1-12.

664 Yuan, H., Mei, R., Liao, J. and Liu, W.-T. (2019) Nexus of Stochastic and Deterministic
665 Processes on Microbial Community Assembly in Biological Systems. Frontiers in Microbiology
666 10(1536).

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

670 Zendehboudi, S., Rezaei, N. and Lohi, A. (2018) Applications of hybrid models in chemical,
671 petroleum, and energy systems: A systematic review. Applied Energy 228, 2539-2566.

672 Zeppilli, M., Chouchane, H., Scardigno, L., Mahjoubi, M., Gacitua, M., Askri, R., Cherif, A. and
673 Majone, M. (2020) Bioelectrochemical vs hydrogenophilic approach for CO2 reduction into
674 methane and acetate. Chemical Engineering Journal 396, 125243.

675 Zhao, H., Hao, O.J., McAvoy, T.J. and Chang, C.-H. (1997) Modeling nutrient dynamics in
676 sequencing batch reactor. Journal of Environmental Engineering 123(4), 311-319.

677 Zhao, X., Li, L., Wu, D., Xiao, T., Ma, Y. and Peng, X. (2019) Modified Anaerobic Digestion
678 Model No. 1 for modeling methane production from food waste in batch and semi-continuous
679 anaerobic digestions. Bioresource Technology 271, 109-117.

680 Zhu, X., Yates, M.D., Hatzell, M.C., Ananda Rao, H., Saikaly, P.E. and Logan, B.E. (2014)
681 Microbial Community Composition Is Unaffected by Anode Potential. Environmental Science &
682 Technology 48(2), 1352-1358.

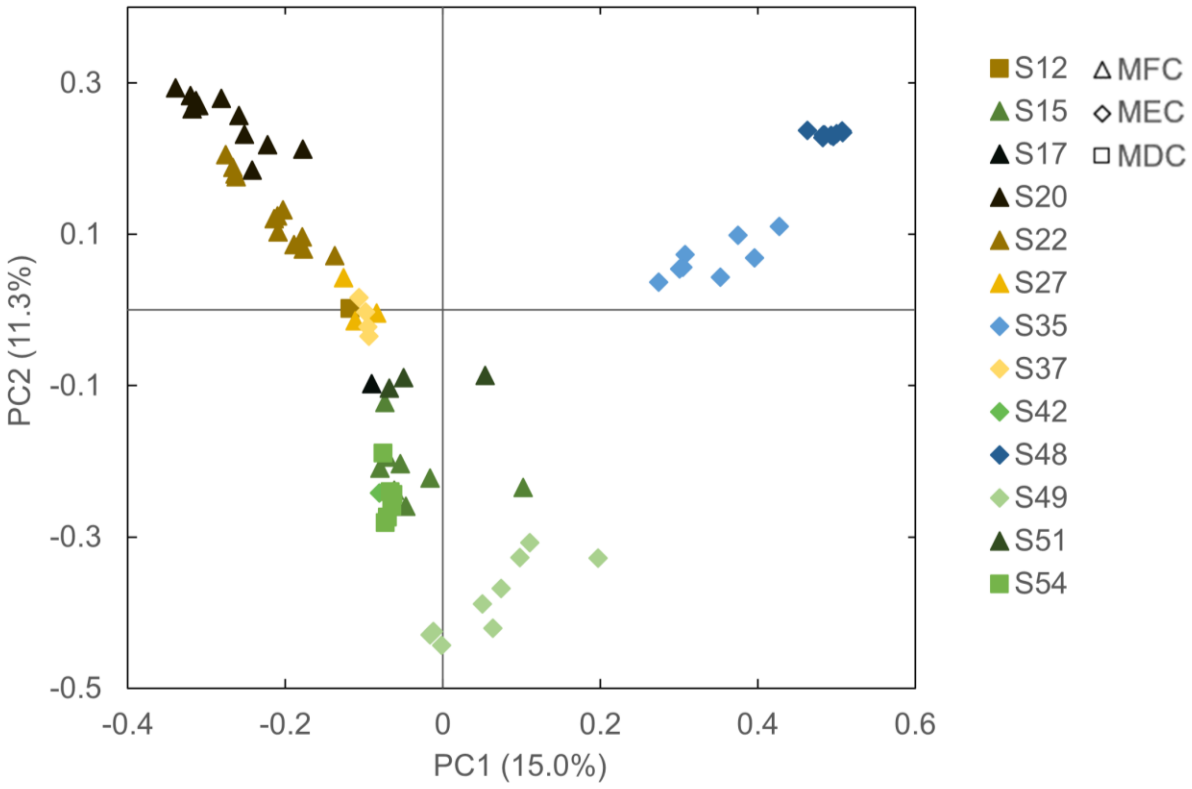


Figure 1. Bray-Curtis distance-based PCoA of the 13 selected studies. MFC: microbial fuel cells, MEC: microbial electrolysis cells, MDC: microbial desalination cells.

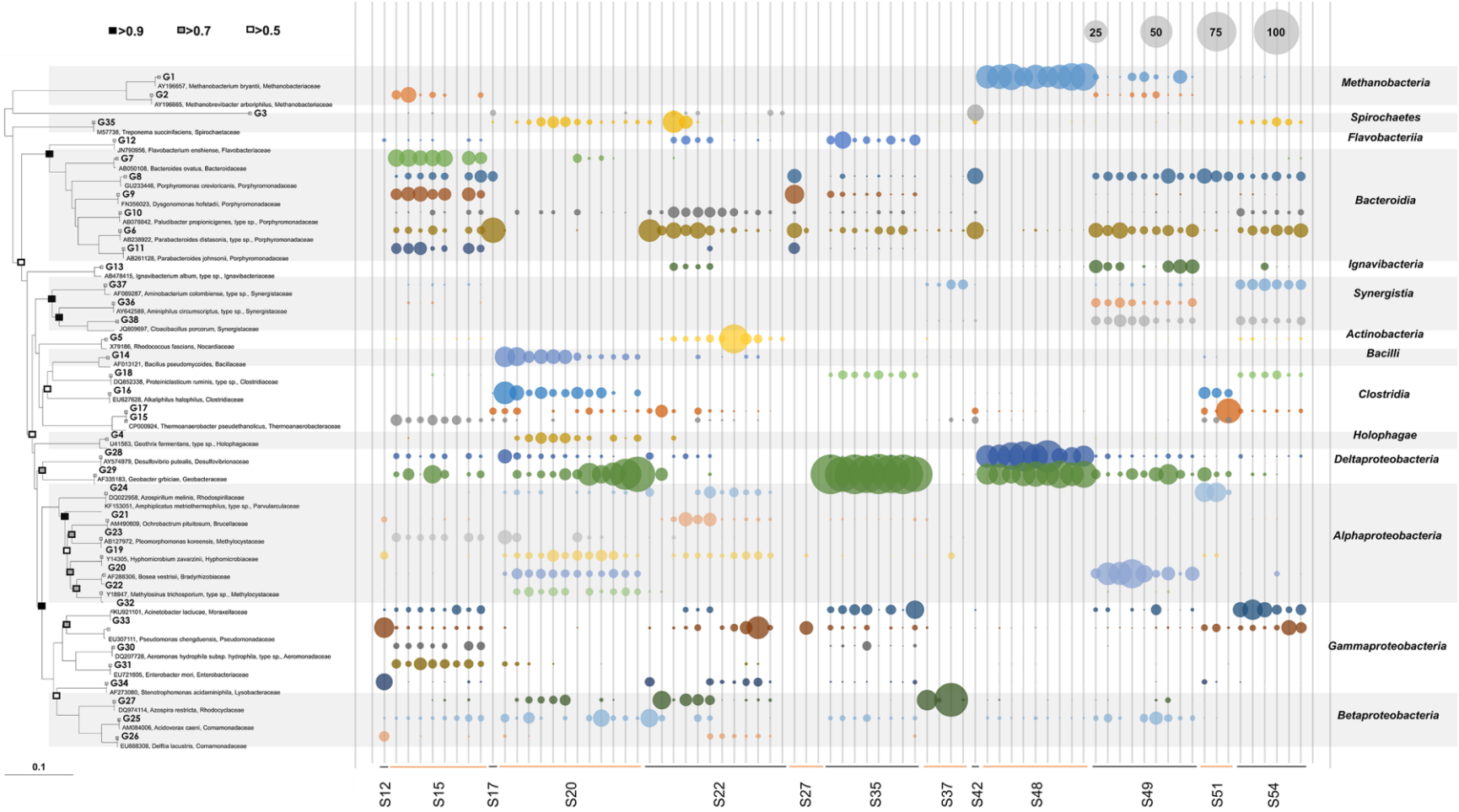


Figure 2. Phylogenetic tree and relative abundance of 38 core genera selected from the 77 samples in 13 publications.

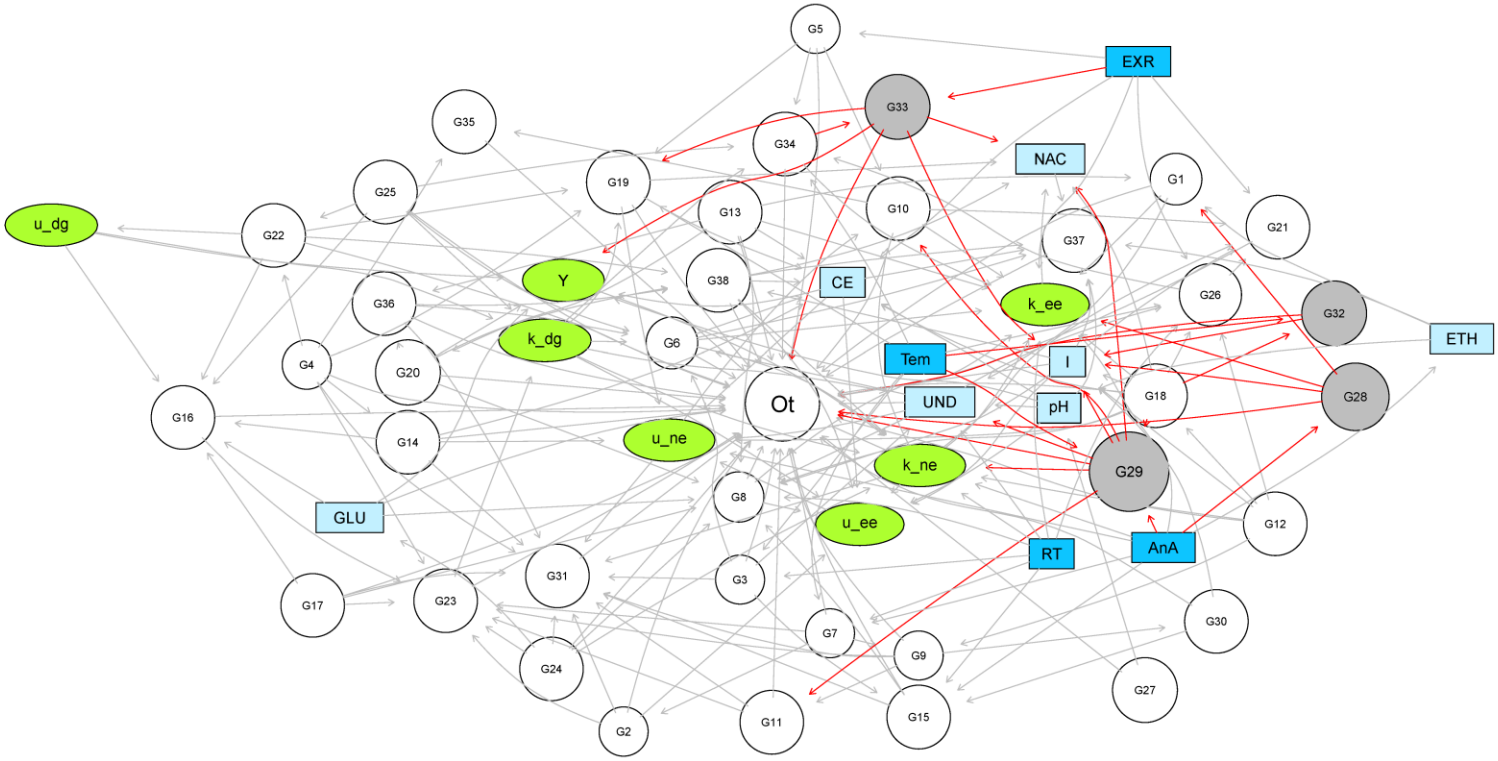
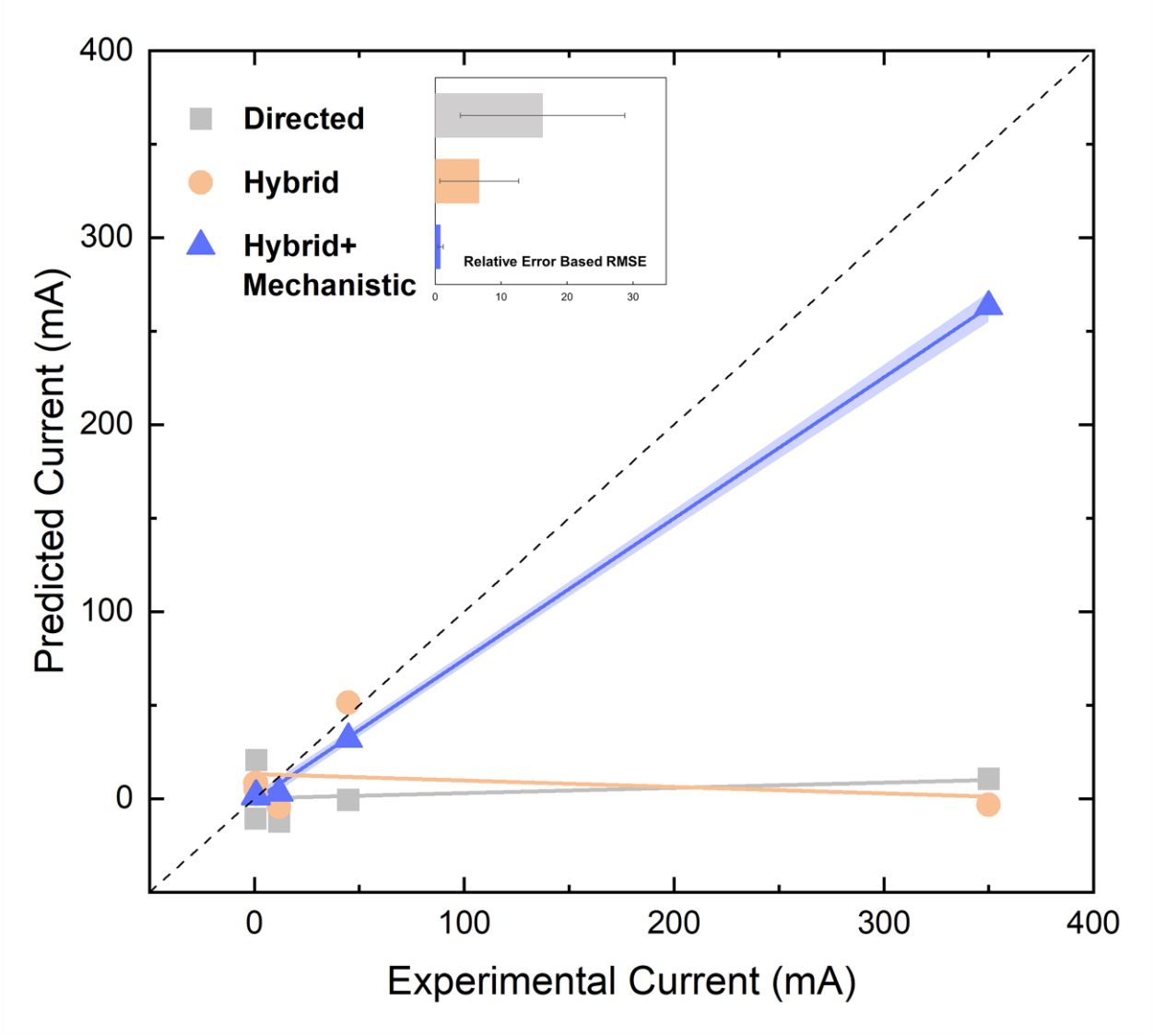


Figure 3. The hybrid Bayesian network at the genus level. *Geobacter* (G29), *Desulfovibrio* (G28), *Pseudomonas* (G33), and *Acinetobacter* (G32) are highlighted in grey and forced to directly affect current. The symbols for the core genera can be found in SI Table S6. Light blue nodes are biochemical parameters. Dark blue nodes are operating parameters unaffected by other nodes and serve only as the parent nodes. Green nodes are kinetic parameters estimated using the mechanistic component, in which u_dg, u_ee, and u_ne are the maximum growth rates for primary degraders, electroactive microbes, and non-electroactive microbes, respectively. k is the maximum substrate utilization rate and Y is the mediator yield. EXR: external resistance, NAC: acetate, CE: coulombic efficiency, I: current, Tem: temperature, ETH: ethanol, GLU: glucose, RT: hydraulic retention time, AnA: anode area, UND: undefined substrate.



683

Figure 4. Comparison of experimental and predicted current from the directed Bayesian network, hybrid Bayesian network, and hybrid model (hybrid network + mechanistic component). Inset in the relative error-based RMSE.