ORT: A workflow linking genome-scale metabolic models with reactive transport codes

Rebecca L. Rubinstein[1,*], Mikayla A. Borton[2], Haiyan Zhou[1], Michael Shaffer[2], David W. Hoyt[3], James Stegen[3], Christopher S. Henry[4], Kelly C. Wrighton[2] and Roelof Versteeg[1,*]

[1]Subsurface Insights, LLC., Hanover, NH, [2]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, [3]Pacific Northwest National Laboratory, Richmond, WA, [4]Argonne National Laboratory, Lemont, IL

*To whom correspondence should be addressed.

Abstract

Motivation: Advanced modeling tools are available for `omics-based metabolic modeling and for reactive transport modeling, but there is a disconnect between these methods, which hinders linking models across scales. Microbial processes strongly impact many natural systems, and so better capture of microbial dynamics could greatly improve simulations of these systems.

Results: Our approach, ORT, applied to environmental metagenomic data from a river system predicted nitrogen cycling patterns with site-specific insight into chemical and biological drivers of nitrification and denitrification processes.

Availability and Implementation: Live interactive models are available at https://pflotranmodeling.paf.subsurfaceinsights.com/pflotran-simple-model/. Microbiological data is available at NCBI via BioProject ID PRJNA576070. The code for ORT (written in Python 3) is available at https://github.com/subsurfaceinsights. The KBase narrative used for the test case is publicly available at https://narrative.kbase.us/narrative/71260.

Contact: rebecca.rubinstein@subsurfaceinsights.com or roelof.versteeg@subsurfaceinsights.com

Supplementary information: Supplementary data are available online.

1 Introduction

Watersheds provide a variety of ecosystem services which are essential for energy, food, and water security. Microbiological processes are a critical component of these services (Anantharaman *et al.*, 2016; Long *et al.*, 2016), driving nutrient cycling and contaminant remediation in both natural and engineered subsurface environments (Rice *et al.*, 1996; Stegen *et al.*, 2018; Furukawa, 2003; Ite and Ibok, 2019; Tchobanoglous *et al.*, 2003). To achieve an actionable understanding of the impact of these microbial processes on macroscopic properties and processes (e.g. soil nutrient availability and water quality), it is necessary to rapidly and cost-effectively obtain, analyze, and interpret related genomic and chemical data. One promising approach for analysis and interpretation is the incorporation of multi-omic data into reactive transport models to better represent microbe-catalyzed biogeochemical reactions. These models can then be used to model site-specific microbiology-informed hydrobiogeochemistry.

Reactive transport models (RTM) are used to simulate coupled chemistry, flow, and transport in biogeochemical systems, allowing us to predict contaminant fate and transport, impact of remediation efforts, or other environmentally significant processes. There are a variety of advanced reactive transport codes (Steefel *et al.*, 2015) that can be used for these types of predictions, including PFLOTRAN (Mills *et al.*, 2009; Hammond and Lichtner, 2010; Gardner *et al.*, 2015), which we used in this work. PFLOTRAN is an open source, massively parallel reactive transport code which supports multi-phase, multi-component, and multi-scale simulation of contaminant transport in porous media, as well as includes a basic implementation of microbial reactions modeled by Monod kinetics. One major benefit of PFLOTRAN compared to other available codes is that it is customizable, allowing users to implement custom reactions or kinetics through the Reaction Sandbox (Hammond, 2017). Using the Reaction Sandbox, we can incorporate microbial dynamics that are not part of the default PFLOTRAN microbial reaction implementation, such as physical inhibition factors or biomass decay. While we have not incorporated these types of dynamics in this initial test case, building our approach around PFLOTRAN provides much-needed flexibility for future development.

The feasibility of using the results of microbiological data analysis to parameterize reactive transport models (RTM) has been shown previously. For instance, Scheibe et al. demonstrated the linking of genome scale models with a reactive transport code (in their case, HYDROGEOCHEM) to improve incorporation of microbiological processes on in situ uranium bioremediation (Scheibe *et al.*, 2009). Specifically, they used a genome scale model of *Geobacter sulfurreducens* to populate a lookup table spanning reasonable expected ranges for all combinations of three key system parameters. This was then used to predict the effects of varying concentrations of three key growth factors

(acetate, Fe(III), and ammonium) on reduction of uranium (VI) at a systems level. More recently, Song et al. developed an enzyme-based approach for simulating microbial reaction kinetics which captured the overall behavior of a consortium rather than rely on individual taxa within the community and coupled it with reactive transport simulations using PFLOTRAN's Reaction Sandbox (Song and Liu, 2015; Song *et al.*, 2017; Hammond *et al.*, 2017). This approach is based on a mechanistic understanding of microbial processes and thus can more accurately predict microbial response to perturbations. However, it requires substantial experimental data, such as enzyme concentrations and kinetics data, as well as advanced microbiological knowledge to implement.

Collectively, while these previous efforts have demonstrated the value of more advanced handling of microbial processes in RTM, they are not well-suited to high-throughput integration. Each requires substantial manual effort to implement for a single use case, so broad application to microbe-driven systems is not reasonable. They also do not provide a clear pipeline for integrating metabolic models based on environmental samples, which typically utilize metagenomic assembled genomes and metabolomics. These data are becoming increasingly available due to enhanced computational workflows which make it feasible to more rapidly process the large volumes of data generated by high throughput instruments. Yet despite these methodological advances, the creation of associated microbiome-informed reactive transport models still remains very much a manual effort.

Metabolic models from environmental samples are becoming increasingly available using tools such as KBase (kbase.us), a US Department of Energy (DOE)-funded high-throughout, web-based, open-source platform designed to enable sharing, curation, and analysis of 'omics data (Arkin *et al.*, 2018). KBase is structured as a narrative interface that users may populate from a library of apps. Each app provides well-defined functionality and outputs. Narratives can also include rich text and raw code cells with code written in Python 3.

Here we present a high throughput workflow that can use environmental (meta)genomes to generate microbiology-informed reactive transport models. This workflow couples free and open source tools KBase and PFLOTRAN to generate reactive transport models informed by environmentally derived genomic, chemical, and physical data. We have exposed two of these models through a user-friendly web interface. They can also be accessed using a software application programming interface (API), allowing the workflow to be leveraged as part of other independent workflows. The results presented here show the feasibility and power of such a pipeline.

2 System & Methods

2.1 Conceptual Model

While previous studies have paved the way for informing RTM with microbial data (Scheibe *et al.*, 2009; Song *et al.*, 2017; Hammond *et al.*, 2017), there is still a need for an automated, high-throughput protocol. Thus, we have designed our pipeline with automation and high throughput in mind. Specifically, this pipeline has well-defined start and end points and inputs and outputs, with each component being fully automatable (Fig. 1). The inputs to this pipeline are annotated genomes, environmental chemistry, physical site data, and a general model type classification. Model classification would mean a description such as "0D batch reactor" or "2D model of unsaturated soil" where "nD" indicates the number of spatial dimensions accounted for in the model grid. Each of these models has specific input requirements which guide the required physical site data.

We use KBase to ingest annotated genomes (Shaffer et al., 2020) and the chemistry (e.g. available carbon sources, electron acceptors, and micronutrients based on metabolome and any other chemical analysis at the site, synthesized into a KBase media recipe), leading to the generation of the overall reactions. Next, these reactions, as well as the site chemistry, physical site data, and model classification are used to build the PFLOTRAN infile, which is automatically constructed using our automated PFLOTRAN input file builder. After the initial model creation, the model is easily updated by repeating the KBase workflow using different inputs and repeating the subsequent steps to generate and retrieve new resulting overall reactions and substitute them into the PFLOTRAN input file. The resulting RTMs can then be executed and exposed either automatically, in our case through our web-based cyberinfrastructure and associated API, or manually. As PFLOTRAN simulates the physical and chemical conditions in the system in space and time, we can feed the resulting chemistry back into KBase as new input data for an iterative modeling approach.

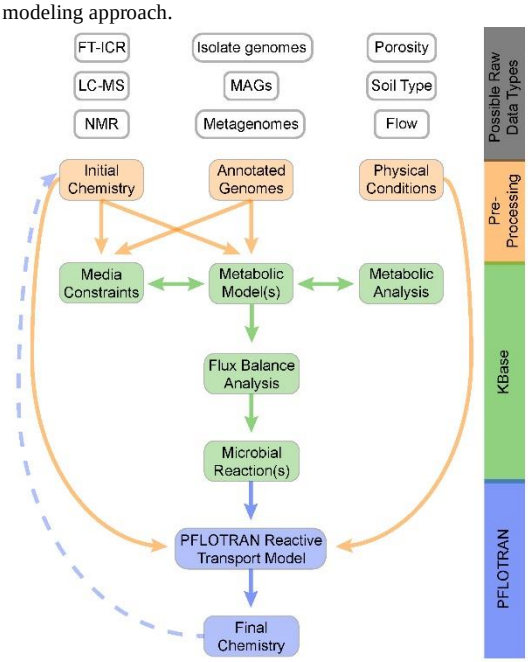

Fig. 1. Flowchart of the proposed workflow where orange boxes are workflow inputs based on site characterization which are pre-processed before use, green boxes are metabolic modeling steps carried out in KBase, and blue show the resulting RTM. The horizontally-aligned boxes and arrows in the KBase workflow represent "optional" curation steps (explained in detail in the text), and the dashed arrow indicates the iteration path (currently manual) wherein the final chemistry of each time step is used to as a new input in KBase.

2.2 Implementation

Our 'Omics to Reactive Transport (ORT) workflow couples chemical (geochemical, metabolite), genomic, and physical environmental inputs, which are highly interdependent but often not in compatible formats, thus it is imperative that this implementation ensures that all components can connect seamlessly. Our workflow automates this connectivity. For example, input data for our test case was manually gathered and pre-processed so that the annotations met the input requirements of KBase (more detail below). Many of these steps required a one-time effort and have now been scripted so that the manual effort is no longer needed. Similarly, chemical data QA/QC and model curation were handled manually, where in the future these need to be (and can be)

automated. While some steps are still manual, these are steps over which we expect users will want to exercise more control, and we have automated the components that link these steps.

Genomes and chemical data are imported into KBase and used as inputs for KBase metabolic modeling apps (process described in detail below). After the completion of the KBase part of the workflow, we use the KBase API to programmatically export the KBase-predicted exchange fluxes from KBase. These fluxes are translated by ORT (available at https://narrative.kbase.us/narrative/71260) into an overall reaction string that describes chemical uptake and secretion from each modeled organism, written in PFLOTRAN-compatible naming conventions. The flux values are used as the stoichiometric coefficients for the corresponding chemicals in the overall reaction used in PFLOTRAN, with positive fluxes indicating reactants and negative fluxes indicating products. The summation of exchange fluxes is not a chemical reaction in the traditional sense, but represents the chemical species removed from and added to the system as a result of the microbial metabolism. Thus, this "pseudo-reaction" provides the information needed by PFLOTRAN to simulate the resulting changes in chemical concentrations.

We then use this reaction string to build our PFLOTRAN model programmatically. Beginning with a model template, we programmatically fill in the microbial reaction string exported from KBase and the initial chemistry (Section). The resulting model can be used to simulate the system behavior. We have a library of roughly eighty different PFLOTRAN model templates (ranging from 0D batch reactors to 3D models of bioremediation) which can be populated programmatically. After filling in the template, we ingest this model into a MySQL database, allowing the model to be exposed through a web interface and an API. Our infrastructure also allows us to expose PFLOTRAN model variables programmatically through the web interface so that users can explore the effect of changing conditions on the system. Users can then select the exposed variables based on what will be most useful or interesting for their research application. The PFLOTRAN and MySQL parts of our workflow run on a completely different server than KBASE, which demonstrates both the value of exposing all software components through APIs and the feasibility of disparate component coupling.

2.3 Test Case

To evaluate the efficacy of our workflow, biological nitrification and denitrification were selected as a classic, well-understood, and extensively studied system, complemented by the availability of relevant real-world environmental samples. We compared a model using traditional (textbook) stoichiometries for nitrification and denitrification to a model derived from metagenome assembled genomes using KBase metabolic modeling tools. Biological nitrogen processing has been the study of extensive research as a core component wastewater treatment (Tchobanoglous *et al.*, 2003; Henze, 1991; Office of Water, 2004), especially since National Pollutant Discharge Elimination System permits began to include effluent nitrogen limits. Excess nitrogen in natural systems from both point and non-point sources has extensive detrimental effects on human health and welfare and the environment (Dodds *et al.*, 2009; Smith *et al.*, 1999; Backer and McGillicuddy, 2006).

For comparison, we selected biological nitrogen cycling in the hyporheic zone, an un-managed system which is similar to that described above. River systems have drawn significant interest as a critical but often underrepresented component of global greenhouse gas models, largely due to lack of experimental data and mechanistic understanding needed to develop strong models. The relationship between hyporheic zone pore water concentrations and N2O emissions is complex, possibly as a result of changes in microbial processes as a result of varying environmental conditions (Villa *et al.*, 2020). In order to maintain the simplicity of the proof of concept and comparison to traditional models, release of gaseous nitrogen species was not considered here, but this is an area of interest to be incorporated in future revisions.

Nitrification is traditionally split into two sub-processes, ammonium oxidation (NH4 → NO2) and nitrite oxidation (NO2 → NO3), while denitrification is often represented as a complete process (NO3 → N2), though in reality it is several sequential reactions. With KBase, we could implement separate models for each step for which genomes are available, but for comparison to the traditional model we used a single model for complete denitrification in this test case. The overall reactions used for the nitrification step were based on experimentally-determined stoichiometries (Liu and Wang, 2012) determined by fitting data collected from bench-scale reactors to traditional half-cell reactions (Rittmann and McCarty, 2012), as given by the following reactions:

$$1.0225\ \mathrm{NH_4^+} + 1.3875\ \mathrm{O_{2\,(aq)}} + 0.09\ \mathrm{CO_{2\,(aq)}} + 0.0225\ \mathrm{HCO_3^-} \rightarrow 2\mathrm{H^+} + \mathrm{NO_3^-} + 0.0225\ \mathrm{Biomass}$$

$$1\ \mathrm{NO_2^-} + 0.0073\ \mathrm{NH_4^+} + 0.4635\ \mathrm{O_{2\,(aq)}} + 0.0292\ \mathrm{CO_{2\,(aq)}} + 0.0073\ \mathrm{HCO_3^-} \rightarrow \mathrm{NO_2^-} + 0.0073\ \mathrm{Biomass}$$

The complete denitrification process stoichiometry was derived from half-cell reactions (Rittmann and McCarty, 2012), scaled to one unit nitrate utilization for comparability with the first two reactions:

$$1\ \mathrm{NO_3^-} + \mathrm{H^+} + 0.869\ \mathrm{CH_3COO^-} \rightarrow 0.458\ \mathrm{N_{2\,(aq)}} + 0.444\ \mathrm{CO_{2(aq)}} + 0.869\ \mathrm{HCO_3^-} + 0.08484\ \mathrm{Biomass}$$

In both cases, the chemical species represented are limited to classical compositions, which in some cases may serve as analogs for a range of compounds. These stoichiometries are not associated with any specific microbes or metabolic pathways, but rather represent the external fluxes observed. While this approach is very effective for process design, it does not offer much insight into the microbiology of a system, and may obscure finer-scale dynamics, particularly in systems with complex carbon sources.

As mentioned previously, reaction rates and other microbial constants are intended to be used as tunable parameters. As a starting point, the rates determined through batch kinetics tests (Liu and Wang, 2012) were used for ammonium oxidation and nitrite oxidation and the denitrification rate was based on rates reported in the literature (Raboni *et al.*, 2014). The same rates (shown in **Table *1***) were used for both the literature-based and genome-based models (described in Section 2.4) in order to directly compare the effects of the different stoichiometries. For future applications, we anticipate reaction rates being used as tunable parameters to fit these models to system-specific experimental data.

Table 1. Initial reaction rates used in web-based nitrogen cycling models

Process	Rate (mol/L·s)
Ammonium Oxidation	1.0×10^{-7}
Nitrite Oxidation	8.51×10^{-8}
Nitrate Reduction	2.34×10^{-8}

2.4 Leveraging Existing Multiomics Data

Sediment was collected and DNA extracted as previously described (Graham et al., 2017). Briefly, 6 sediment cores up to 60 cm in depth were collected at 5-meter intervals in the hyporheic zone of the Columbia River (46°22'15.80″N, 119°16'31.52″W) in March 2015. Each core was sectioned into 10 cm segments from 0-60 centimeter depths and

stored at -80ºC. All biological analyses were carried out at 10-centimeter increments, except for one core that had low yields, which was pooled from 0-30 centimeters to have sufficient input masses (Graham et al., 2017, 2018). DNA was extracted as previously described (Graham et al., 2017, 2018) using MoBio PowerSoil kit (MoBio Laboratories, Inc., Carlsbad, CA).

To identify the metabolites available to microorganisms in these river sediments, we performed 1H Nuclear Magnetic Resonance (NMR) spectroscopy on 17 sediment pore water samples as described previously (Tfaily *et al.*, 2019). Briefly, sediment samples were mixed with water in a 1:1 ratio and then diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d6 as an internal standard. The 1D 1H NMR spectra of all samples were processed, assigned, and analyzed using Chenomx NMR Suite 8.3 with quantification based on spectral intensities relative to the internal standard as described. To obtain a representative bulk summary of the metabolite environment in these sediments, the concentration of 31 of the NMR identified metabolites was averaged across the 17 sediment samples, and this data was used as the chemical data input in our ORT workflow (data available in Supplementary Table SX).

Purified genomic DNA was sent to the Joint Genome Institute (JGI, n=33) under JGI/EMSL proposal 1781 and to the Genomics Shared Resource facility at The Ohio State University (OSU, n=10), producing 43 metagenomes from 34 sediment samples with an average sequencing depth of 3.84 (JGI) 25 Gbp (OSU) per sample. JGI and OSU sequencing was performed as previously described in Graham et al (Graham *et al.*, 2018) and Borton et al (Borton *et al.*, 2018) respectively. Raw reads were processed, assembled, and binned as outlined in previous publications (Shaffer *et al*., 2020) or via the Wrighton Lab GitHub Page (https://github.com/TheWrightonLab). The genomes are available on NCBI via BioProject ID PRJNA576070.

From the sediments we obtained metagenome assembled genomes (MAGs) from which we selected 4 genomes that represented key parts of the nitrogen cycle. To represent nitrification, we chose the most complete genome representatives of the ammonium oxidizing archaea classified by GTDB-Tk (version 1.3.0, as of 1-21-21) as a member of the family Nitrososphaeraceae within the genus TA-21 (previously within the Phylum Thaumarchaeota) and nitrite oxidizing bacterial member of the Nitrospiraceae for nitrification. To represent denitrifcation, we selected two Gammaproteobacterial MAGs, both classified within the family Steroidobacteraceae. Note that neither of these genomes encoded a gene to produce N_2 gas, but the reaction to convert nitrous oxide to nitrogen gas was added to the metabolic models during gapfilling (see Section 2.5). Each nitrogen-cycling genome was annotated using DRAM (Distilled and Refined Annotation of Metabolism (Shaffer et al., 2020)) with default parameters. The raw annotations containing an inventory of all database annotations for every gene from each input genome are reported in Supplementary File SX. These genomes and their annotations were ingested into KBase (Section 2.5) and were the basis for the KBase-derived model (Section 2.6).

2.5 Pre-ingestion Processing

At the beginning of our workflow, user inputs were organized and prepared, which consisted of three broad steps. In this section, we describe these in generic terms, as the same organization would be applied to any system.

(1) Qualitative assessment – to balance model complexity and utility, the system definition phase began with a qualitative description of the system in terms of type (batch, chemostat, continuously stirred tank reactor, etc.), important processes (such as nitrification or sulfur reduction, depending on the system), and parameters of interest (pH, specific chemical species, etc.) that can guide model development. This step includes evaluating if there is any "missing" data, which might render the model inaccurate or impossible, and would need to be estimated in order to produce a viable system (for example, concentrations of biologically necessary compounds that were not measured). These are identified through a combination of subject matter knowledge and comparison with KBase default media recipes. Note that this does not entail delineation of every process and parameter involved in the system, but rather selection of those important to the specific research or application. The goal of this step is to develop a conceptual model of the system of interest, which may be augmented and refined as needed to accommodate new data.

(2) Data Gathering - data describing the site may be drawn from a variety of sources, including direct sampling at the site and public resources such as weather stations or national databases. Biological data could come in the form of annotated genomes or metagenomes collected from the site, or genomes for key microbes as determined using 16S data or literature review could be drawn from public databases. Chemical data could include traditional geochemical analysis as well as metabolomics and metaproteomics to provide a more detailed picture of the chemical profile at the site. Physical data could include temperature, soil porosity, or other parameters of that nature that would be included in the PFLOTRAN input file to produce a more site-specific model.

(3) Translation to KBase and PFLOTRAN - the data produced by the various analyses above are not necessarily in formats that may be directly imported to KBase and/or PFLOTRAN. Therefore, the final step in this phase was to translate these data to forms that can be used by the tools. Aside from managing file formats (see the KBase documentation for details), one major consideration was accounting for any un-measured chemical species identified in the first step of the preparation phase that needed to be added to the KBase media composition to make it biologically viable. Additions were limited to chemical species or compounds known (or reasonably expected) to be present and were added in sufficient concentration that they would not be growth-limiting. This is something that could be evaluated by running the same model on media with a single input concentration varied and comparing the outputs.

These steps are currently carried out manually based on expert knowledge and experience, and could be applied to any system of interest. The existing system could be modified by simply replacing the genomes or input chemistry, and the approach could be expanded to additional systems by implementing a template into which end-users could put their data and conceptual model.

2.6 KBase Metabolic Modeling

Once pre-processing was complete, genomes were uploaded to KBase as paired FASTA and GFF3 text files using the "Import GFF3/FASTA file as Genome from Staging Area" app and then annotated with RASTtk using the "Annotate Microbial Genome" app in KBase. Additional custom annotations from DRAM were uploaded as flat text files using the beta version of "Import Annotations from Staging" app. If using DRAM annotations, preprocessing may be carried out using the provided script at https://github.com/subsurfaceinsights. Notably, both RASTtk and DRAM are available as apps in KBase, allowing users to functionally annotate genomes without high memory computational resources. However, note that the DRAM app in KBase differs from the version used in this example narrative (Shaffer *et al.*, 2020) as the KBase DRAM app annotates using KOfam instead of KEGG genes and does not currently include EC reaction identifiers, so end results may differ from the included narrative. Chemical data was uploaded as flat text files using the "Import Media file (TSV/Excel) from Staging Area". The use of pre-processed flat text files as inputs to the workflow significantly simplifies the process compared to using raw data, especially for genomes, and these can be generated automatically using scripts such as the one developed for the DRAM outputs. This first step brought all of our data in the KBase workspace in an integrated manner.

After this step, we used all this data as inputs to the "Build Metabolic Model" app, and the generated models were used in conjunction with the media objects as inputs to the "Run Flux Balance Analysis" (FBA) app. The output from the FBA app included the reaction and exchange fluxes for each model grown on the corresponding media.

2.7 PFLOTRAN Reactive Transport Modeling

As described, we used our workflow to download the FBA exchange flux values using the KBase API and translate them from KBase objects with ModelSEED (Henry *et al.*, 2010) compound IDs to flat text files with reaction strings written using PFLOTRAN naming conventions. We then used either the KBase-derived reaction strings and biomass yield values or the literature-based stoichiometries introduced in Section 2.3 to fill in the MICROBIAL_REACTION card in our 0D model template. All parameters except the reactions and yield terms were held the same for both the conventional and the site-specific model. The models were evaluated by comparing the relative "system performance" with respect to nitrogen conversion to N_2 or to biomass, as described in greater detail below.

3 General behaviors and trends

Both models exhibited sequential ammonium and nitrite oxidation followed by nitrate reduction, ultimately producing dissolved nitrogen gas. Despite using the same reaction rates, inhibition constants, and initial nutrient concentrations, the overall progress of the system is noticeably different. The genome-based model exhausts the available ammonium within 1.5 hours of the simulation start, while the literature based model does not exhaust ammonium until a little more than 3.5 hours into the simulation. Nitrite concentration peaks earlier and at a lower level for the genome-based model (~23 µM at approximately 1 hr) than the literature based model (~51 µM slightly before 3 hrs). Similarly, nitrate peaks at approximately 20 µM after 1.7 hrs for the genome-based model but peaks at 40 µM at the 6 hr mark for the literature based model. In the 6 hour period shown in Fig. 2, the genome based model has exhausted ammonium, nitrite, and nitrate, while the literature based model is still processing nitrite and nitrate. This variance is expected since we are comparing generic reactions (with generic substrate utilization and biomass production reactions) to site-specific reactions based on the most dominant taxa found at our study site.

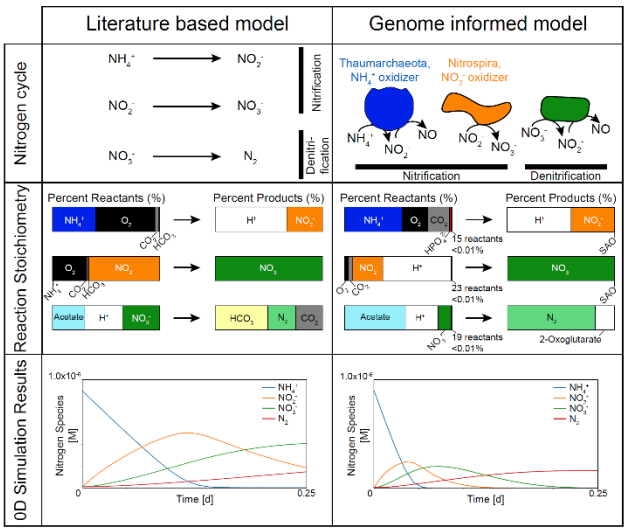

Fig. 2. Using our Omics to Reactive Transport (ORT) workflow allows us to not only tailor a model to a specific environmental site and system, but also provides much finer insight into the changes in chemistry driven by microbial processes. The top frame shows the steps captured by the literature-based and genome-informed models respectively. The literature-based model steps are simplistic and not tied to any specific metabolisms, while the genome-informed model is based on the metabolisms of specific taxa found to be prominent at this site. Given that the expression and activity of nitrite reductase encoded in Thaumarchaeota is poorly understood at this time (Kuypers et al., 2018), we did not incorporate the production of nitric oxide by Thaumarchaeota, and focused only on nitrite outputs from ammonification. The middle frame shows graphical representations of the two sets of reaction stoichiometries. At a high level these are similar, but upon closer inspection there are details that may make a significant difference in a real-world system. In particular, there are a much broader range of carbon compounds utilized and secreted by the genome-informed model. Given appropriate time series data, these generalized stoichiometries could be tuned to a specific site and time period, allowing even more accurate representation of carbon cycling. SAO stands for S-Adenosyl-4-methylthio-2-oxobutanoate, which is a compound predicted by KBase as an output which is not part of standard literature representations. The bottom frame shows the results of using each set of reactions in a 0D PFLOTRAN simulation of nitrogen cycling. The general patterns are similar, but the different stoichiometries result in noticeable differences in the magnitudes even though the same reaction rates were used in both.

One important difference was that the microbiologically-explicit, genome-based stoichiometry provided much greater detail on the chemistry, particularly with respect to carbon catabolism (Fig. 2 and Fig SX). Specifically, the literature-based models relied entirely on either carbon dioxide (nitrification) or acetate (denitrification), however, because we provided additional carbon compounds detected from our bulk sediment metabolome, the site models used 15 to 23 unique additional carbon sources, such as betaine, leucine, and choline (see Supplementary Table X). This greater detail allows us to evaluate more precisely the potential

chemical drivers or limiters of a system which would be entirely overlooked with traditional representations, which presents the opportunity to probe and improve our conceptual and mechanistic understanding of these systems and individual metabolisms.

Instead of generic bacterial enzymatic reactions, we can determine which site specific bacterial – or archaeal – reactions are drivers in the system. Instead of pre-set stoichiometries, our Omics to Reactive Transport workflow uses chemistry determined based on metabolomics. For example, even with the same rate constants, we can see that the genome-informed model utilizes a higher proportion of ammonium in the first step of nitrification, resulting in more rapid depletion of ammonium in the system and earlier generation of nitrite. As a result, subsequent steps begin earlier, resulting in an overall accelerated process. At the same time, both versions exhibit the expected cycling of ammonium to nitrite to nitrate and finally to nitrogen gas. Since PFLOTRAN relies on user-defined chemistry (as opposed to automatically generating reactions), this allowed us to incorporate more realistic, mechanism-driven reactions.

The genome-based model allows for greater chemical breadth. The nitrogen cycling reactions are modulated by a wider range of carbon sources. Additionally, the by-products of this carbon and nitrogen metabolism also resulted in more complex chemical outputs in some cases, such as 2-oxoglutarate or S-Adenosyl-4-methylthio-2-oxobutanoate. These inferred reactions could be further refined by using gene expression data (e.g. metatranscriptomics or metaproteomics data) to calibrate the models (by way of reaction rates, saturation constants, etc.) to a particular set of environmental conditions. Again, this presents an opportunity to test and enhance our understanding of the metabolic processes involved.

These models are publicly available (without sign-in) through Subsurface Insights' web-based PFLOTRAN interface at https://pflotranmodeling.paf.subsurfaceinsights.com/pflotran-simple-model/. For the literature-based model, we have made the input concentrations of ammonium, bicarbonate, and acetate accessible to web users using sliders. For the Hanford 300 Area-specific version of the model, we have made accessible the reaction rate for each of the steps modeled. There is no limit to the number of parameters that may be exposed this way, but for the sake of a user-friendly and un-cluttered demonstration, we limited our selections to three per model. We selected the parameters we did both because the effects of varying them are significant and to highlight the power and flexibility provided by this approach.

4 Discussion

Here we demonstrated an Omics to Reactive Transport (ORT) workflow for creating site specific reactive transport models that include local chemical and biological content. The ORT workflow was applied to a well-understood system, and the results agree generally with literature data. We interpret the differences in magnitude and timing to be due to the difference between generic, simplified reactions and metabolism-informed reactions, as KBase-derived stoichiometries made it possible to capture microbial metabolism in much greater detail than conventional approaches allow.

While the model predictions are borne out by comparison to traditional models, we would need extensive new data which currently is not available to comprehensively validate our modeling results. Rigorous model and method validation will be part of future work. Specifically, the sampling effort underlying our data captured a microbiological and geochemical snapshot of a dynamic system. To expand our model to a 3D site model and validate the results, we will need samples covering a broader range of time and space.

Much of the future work on this workflow will be focused on enhancing and expanding automation and on making it more robust in several ways. In particular, we are developing a user interface that will allow for and provide guidance on QA/QC and the initial setup. Automated metabolic curation would also be highly beneficial to this pipeline. In our effort, curation was carried out manually using two different approaches: metabolism-based and media-based. The former is labor intensive and requires substantial subject-matter expertise to carry out. The latter is more straightforward and relies on a more general system understanding, but still requires manual iteration to obtain reasonable results. Partially or fully automated model curation is needed to allow the workflow to support high-throughput processing of data to produce simulations. Finally, we need to expand on our PFLOTRAN models to include processes such as temperature mediated biological processes and material recycling. While these are non-trivial tasks, we feel that they are feasible and that we have a clear path forward.

While previous researchers have demonstrated the feasibility of coupling genome-scale metabolic models with reactive transport simulations, our work is different in some fundamental ways. First, our approach allows us to use microbial reaction constants as tunable parameters. The reaction rate, half-saturation concentration, and inhibition constants in our study are based on literature review. While we note that these values may not be available for all systems, we anticipate they could still be approximated based on near relatives or determined experimentally. The latter would also allow the models to account for conditions that otherwise might not be captured, such as temperature or soil conditions which would change microbial growth rates. Second, our approach allows for responsive modeling of dynamic systems – by feeding back the chemistry modeled by PFLOTRAN (as well as other temporal changes) we can model system evolution without needing to anticipate changes at the outset. This includes dynamics of microbial communities as they respond to different and varying conditions. This will allow for investigating complex and poorly understood systems in which we can probe microbial behavior under a variety of conditions. Finally, our current results set the stage for the automated generation of microbiology-informed models which can be easily used through either a web interface or an API by end users who are not specialists in reactive transport modeling or even microbiology. We expect this to be of interest to a very broad community and support research efforts in many fields.

Acknowledgements

Tasya Rodzianko, Doug Johnson, and Erek Alper at Subsurface Insights work on the cyberinfrastructure and web interface that was used in this work. Garret Smith, Pengfei Liu, and Lindsey Solden provided additional microbiological expertise and processing. Field sample data was collected by Evan Arntzen, Alex Crump, Brad Fritz, Dave Kennedy, Sarah Fansler, Nate Phillips, Sadie Montgomery, Kyle Parker, and Rob Macklet at Pacific Northwest National Laboratory. Processing of fine sediments was also performed by Ray Clayton and Chris Strickland and cultural support was provided by Doug McFarland and Joy Ferry.

Funding

This work has been supported by the SBIR Award DE-SC0019619, Integrated Management and Analysis Platform for Multi Domain Site Data (program manager Paul Bayer) from the DOE Biological and Environmental Research program. A portion of the metagenomic sequencing for this research was

performed by the Department of Energy's Joint Genome Institute (JGI) via sequencing award no. 1781. Metabolite support was provided by Environmental Molecular Sciences Laboratory (EMSL) via award no. 50334. Both JGI and EMSL facilities are sponsored by the Office of Biological and Environmental Research and operated under contract nos. DE-AC02- 05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). Field sample collection and processing was part of the Scientific Focus Area (SFA) project at PNNL, sponsored by the U.S. Department of Energy, Office of Science, Environmental System Science (ESS) Program. This contribution originates from the ESS Scientific Focus Area (SFA) at the Pacific Northwest National Laboratory (PNNL).
Conflict of Interest: none declared.

References

Anantharaman,K. *et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*, **7**, 13219.

Arkin,A.P. *et al.* (2018) KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol*, **36**, 566–569.

Backer,L.C. and McGillicuddy,D.J.Jr. (2006) Harmful Algal Blooms: at the Interface Between Coastal Oceanography and Human Health. *Oceanography*, **19**, 94–106.

Borton,M.A. *et al.* (2018) Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc Natl Acad Sci USA*, **115**, E6585–E6594.

Dodds,W.K. *et al*. (2009) Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages. *Environmental Science & Technology*, **43**, 12–19.

Furukawa,K. (2003) 'Super bugs' for bioremediation. *Trends in Biotechnology*, **21**, 187–190.

Gardner,W.P. *et al.* (2015) High Performance Simulation of Environmental Tracers in Heterogeneous Domains. *Groundwater*, **53**, 71–80.

Graham,E.B. *et al.* (2018) Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. *Science of The Total Environment*, **642**, 742–753.

Hammond,G.E. *et al.* (2017) Application of a hybrid multiscale approach to simulate hydrologic and biogeochemical processes in the river-groundwater interaction zone. Sandia National Lab. (SNL-NM), Albuquerque, NM (United States).

Hammond,G.E. (2017) PFLOTRAN Reaction Sandbox: A Flexible Extensible Framework for Vetting Biogeochemical Reactions within an Open Source Subsurface Simulator.

Hammond,G.E. and Lichtner,P.C. (2010) Field-scale model for the natural attenuation of uranium at the Hanford 300 Area using high-performance computing: MODEL FOR NATURAL ATTENUATION OF URANIUM. *Water Resour. Res.*, **46**.

Henry,C.S. *et al.* (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, **28**, 977–982.

Henze,M. (1991) Capabilities of Biological Nitrogen Removal Processes from Wastewater. *Water Science and Technology*, **23**, 669–679.

Ite,A.E. and Ibok,U.J. (2019) Role of Plants and Microbes in Bioremediation of Petroleum Hydrocarbons Contaminated Soils. *International Journal of Environmental Bioremediation & Biodegradation*, **7**, 1–19.

Kuypers,M.M.M. *et al.* (2018) The microbial nitrogen-cycling network. *Nat Rev Microbiol*, **16**, 263–276.

Liu,G. and Wang,J. (2012) Probing the stoichiometry of the nitrification process using the respirometric approach. *Water Research*, **46**, 5954–5962.

Long,P.E. *et al.* (2016) Microbial Metagenomics Reveals Climate-Relevant Subsurface Biogeochemical Processes. *Trends in Microbiology*, **24**, 600–610.

Mills,R.T. *et al.* (2009) Modeling subsurface reactive flows using leadership-class computing. *J. Phys.: Conf. Ser.*, **180**, 012062.

Office of Water (2004) Primer for Municipal Wastewater Treatment Systems United States Environmental Protection Agency, Washington, D.C.

Raboni,M. *et al*. (2014) Calculating specific denitrification rates in pre-denitrification by assessing the influence of dissolved oxygen, sludge loading and mixed-liquor recycle. *Environmental Technology*, **35**, 2582–2588.

Rice,C.W. *et al*. (1996) Role of Microbial Biomass Carbon and Nitrogen in Soil Quality. In, Doran,J.W. and Jones,A.J. (eds), *Methods for Assessing Soil Quality*, Special Publication. Soil Science Society of America, Madison, WI, pp. 203–215.

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

Scheibe,T.D. *et al.* (2009) Coupling a genome-scale metabolic model with a reactive transport model to describe in situ uranium bioremediation. *Microbial Biotechnology*, **2**, 274–286.

Shaffer,M. *et al.* (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*, **48**, 8883–8900.

Smith,V.H. *et al*. (1999) Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environmental Pollution*, **100**, 179–196.

Song,H.-S. *et al.* (2017) Regulation-Structured Dynamic Metabolic Model Provides a Potential Mechanism for Delayed Enzyme Response in Denitrification Process. *Frontiers in Microbiology*, **8**.

Song,H.-S. and Liu,C. (2015) Dynamic Metabolic Modeling of Denitrifying Bacterial Growth: The Cybernetic Approach. *Industrial & Engineering Chemistry Research*, **54**, 10221–10227.

Steefel,C.I. *et al.* (2015) Reactive transport codes for subsurface environmental simulation. *Computational Geosciences*, **19**, 445–478.

Stegen,J.C. *et al.* (2018) Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat Commun*, **9**, 585.

Tchobanoglous,G. *et al.* (2003) Wastewater engineering treatment and reuse 4th ed. McGraw-Hill Higher Education, Boston, US.

Tfaily,M.M. *et al*. (2019) Single-throughput Complementary High-resolution Analytical Techniques for Characterizing Complex Natural Organic Matter Mixtures. *JoVE*, 59035.

Villa,J.A. *et al*. (2020) Methane and nitrous oxide porewater concentrations and surface fluxes of a regulated river. *Science of The Total Environment*, **715**, 136920.