2022

FROM THERMAL SPRINGS TO SUBWAY BENCHES: EXPLORING THE DIVERSITY OF CARBON MONOXIDE DEHYDROGENASES THROUGH METAGENOMES, PHYLOGENETICS, AND MACHINE LEARNING

Isaac Bigcraft
Michigan Technological University, isbigcra@mtu.edu

FROM THERMAL SPRINGS TO SUBWAY BENCHES: EXPLORING THE DIVERSITY OF CARBON MONOXIDE DEHYDROGENASES THROUGH METAGENOMES, PHYLOGENETICS, AND MACHINE LEARNING

By

Isaac Bigcraft

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Biological Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY

2022

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Biological Sciences.

Department of Biological Sciences

Thesis Advisor:	*Stephen Techtmann*
Committee Member:	*Trista Vick-Majors*
Committee Member:	*Carsten Kuelheim*
Department Chair:	*Chandrashekhar Joshi*

Table of Contents

List of Figures

Abstract

Carbon monoxide is well known as a toxic gas but can also be an important input and intermediary for microbial metabolisms. Carbon monoxide dehydrogenases (CODHs) serve as key enzyme complexes for a variety of microbial carbon monoxide (CO) utilization pathways. Such pathways include the Wood-Ljungdahl pathway, which is important in methanogenesis and acetogenesis, metal and sulfate reduction pathways, hydrogen production, and others. The CODH enzymes allow microbes to turn the traditionally toxic waste gas of CO into a useful carbon and energy source. Despite the flexibility of CODH enzymes, the use of carbon monoxide is still believed to be a fringe metabolism. Here we seek to expand the known diversity, distribution, and phylogeny of CODH catalytic subunit proteins by searching an expansive dataset of over 50,000 metagenome assembled genomes. Our work has shown that this dataset contains 5,426 putative CODH protein sequences found within 4,001 metagenome assembled genomes. Despite the considerable expansion of the known set of CODH sequences, our phylogenetic analysis has validated the protein's previously established phylogeny while showing a wider environmental and taxonomic distribution of CODHs. Often considered to be found primarily in areas with high levels of CO, CODHs are typically associated with thermal and extremophiles. In addition to the expected high temperature environments, CODHs were found in metagenomes from diverse environments from soils to subway benches, and in phyla ranging from archaeal Euryarchaeota to bacterial Actinobacterota. We also have constructed a machine learning model to extract functional predictions and information using a sequence-only method to predict gene ontologies (GO-terms) associated with CODH function. While our model can achieve accurate prediction of GO-terms, our work has shown some of the current limitations in the approach. This study reveals CODHs to be a more diverse and ubiquitous enzyme than previously anticipated. Despite tripling the number of sequences in the phylogeny, we provide strong support for the previously established clades and report no new clades. This work has also identified some key areas for experimental follow up regarding the importance of carbon monoxide and CODH genes in many environments.

1 Introduction

Carbon monoxide (CO) is a household name due to its danger as a colorless, odorless, toxic gas. Usually found as an atmospheric gas in only harmless trace concentrations, CO is also found in more substantial concentrations in a variety of natural and anthropogenic locales. Hydrothermal vents, industrial gas flues, and volcanic environments all experience CO levels greater than 1 ppm (Robb & Techtmann, 2018). Despite carbon monoxide's toxic reputation, some microorganisms can thrive in CO-rich environments by utilizing the gas as a carbon and energy source for downstream reactions in a process known as carboxydotrophy. While carboxydotrophs have been found to be common in extreme environments, the low concentrations of carbon monoxide in most environments suggests that carboxydotrophy is likely a fringe metabolism performed by a limited range of microorganisms in restricted environments. However, some microorganisms release CO as a metabolic byproduct, creating local pockets of high CO concentrations in many more, much less extreme environments (Voordouw, 2002). It has been proposed that these biogenic pockets of CO may support the growth of carboxydotrophs in a wider range of common environments (Techtmann et al, 2009).

Metabolism of carbon monoxide is largely facilitated by versatile carbon monoxide dehydrogenase (CODH) enzymes. This family of enzymes consists of protein complexes involved in the oxidation or reduction of CO. CODHs are divided into two distinct protein families: aerobic and anaerobic CODHs. CODH complexes can be made up of a variety of different subunits to modulate the complex's function, but across CODHs the active site containing catalytic subunit remains foundational. These foundational subunits catalyze the reversible reaction $CO + H_2O <=> CO_2 + 2H^+ + 2e^-$; with aerobic and anaerobic CODHs requiring differing metals in their active sites. Aerobic CODHs, also known as coxSML CODHs, rely on a Mo-Cu active site to transfer electrons from CO to oxygen or nitrate (King et al., 2007). Anaerobic CODHs instead utilize a Ni-Fe based active site for either the oxidation or reduction of CO/CO_2 for a much wider variety of pathways (Techtmann et al., 2012). More specifically, anaerobic CODHs have either a CooS or Cdh catalytic subunit, both of which have very similar structures with only minor differences. CooS proteins are found predominantly in bacterial genomes with some exceptions, while Cdh proteins are found almost exclusively in archaeal genomes (Inoue et al., 2019). This study focuses specifically on the versatility of anaerobic CODHs, which will be abbreviated to simply CODHs. Additionally, due to having a very similar structure and function, both CooS and Cdh proteins will be hereafter referred to collectively as 'CODH proteins', as they are the defining and identifying characteristic of all anaerobic CODH complexes.

Interestingly, the reaction catalyzed by anaerobic CODH proteins can serve in a variety of different downstream pathways; a characteristic unusual for a single enzyme family and one not shared with aerobic CODHs. CODH facilitated oxidation of CO can be used as the initial step of the Wood-Ljungdahl pathway for methanogens and acetogens, or the reaction can provide a source of electrons to be utilized for metal and sulfate reducing pathways. These electrons from CO can also fuel hydrogen production in hydrogenogens through the energy converting hydrogenase or can be used to produce reducing equivalents to be used for other cellular functions. These diverse functions are further illustrated in Figure 1 by Robb and Techtmann (2018). The metabolic fate of CO catalyzed by a specific CODH has been proposed to be determined by the presence of specific additional subunits in the CODH complex (Robb and Techtmann, 2018). Many CODH possessing organisms have been found to have multiple different CODH operons encoded in their genomes (Techtmann et al., 2011). For example, the prototype thermophilic carboxydotroph – *Carboxydothermus hydrogenoformans*, encodes five CODHs, shown in Figure 2 originally from Wu et al. (2005), which have been proposed to have diverse metabolic fates including hydrogen production, generation of reducing equivalents, coping with oxidative stress, as well as the Wood-Ljungdahl pathway for autotrophic growth on CO (Wu et al., 2005; Svetlitchnyi et al., 2001). Similarly, the thermophilic metal reducing bacteria *Thermincola potens* contains four distinct CODH gene clusters (Byrne-Bailey et al., 2010). The phenomenon of a single organism encoding multiple CODH gene clusters has been shown to be common in CODH possessing bacteria and archaea (Techtmann et al., 2012). The presence of multiple CODHs in a single genome suggests that these organisms may be capable of utilizing CO as an input or electron source to multiple different pathways simultaneously (Techtmann et al., 2011).

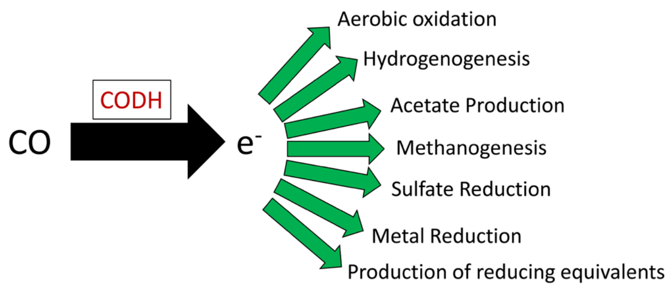

Figure 1. Originally published as Figure 1A in Robb and Techtmann (2018) with the caption “Potential fates of CO in microbial physiologies”

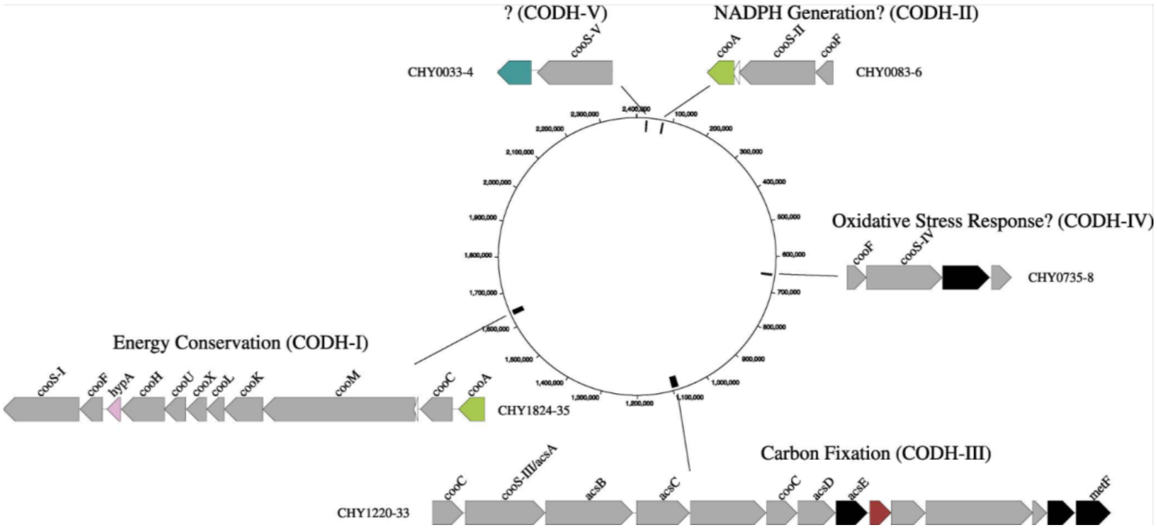

Figure 2. Originally published as Figure 3 in Wu et al. (2005), *Life in Hot Carbon Monoxide: The Complete Genome Sequence of Carboxydothermus hydrogenoformans Z-2901*, with the caption "The genome locations [in *C. hydrogenoformans*] of the genes encoding the five CooS homologs (labelled CooS I-V) are shown. Also shown are neighboring genes that are predicted to encode the five distinct CODH complexes (CODH I-V) with each CooS homolog. Possible cellular roles for four of the five CODH complexes are indicated."

Elucidating the functional roles of CODHs can have important broader implications in understanding the ecological roles and evolution of CO and CO-based metabolisms, as well as informing the biotechnological applications of these metabolisms. Carbon fixation using CO, facilitated by CODHs, has been proposed to be an ancient metabolic pathway present in the Last Universal Common Ancestor (LUCA) (Adam et al., 2018). Other studies have indicated that CODHs as part of the Wood-Ljungdahl pathway represent one of the 355 proteins that trace back to the LUCA. Weiss et al. (2016) posit that the origin of autotrophic life involved the Wood-Ljungdahl pathway for CO_2/CO fixation. Therefore, elucidating the mechanisms and evolution of CO based carbon fixation via that CODH-containing Wood-Ljungdahl pathway could further clarify the evolution of life on earth, the evolution of later forms of carbon fixation, and potential pathways for carbon fixation in astrobiological contexts (Russell et al., 2014; King 2015).

In modern environments where CODHs are found, the organisms encoding these genes may be utilizing them for important and diverse functional purposes. A given CODH could be a step in the Wood-Ljungdahl pathway, serving as an important factor in anaerobic carbon fixation especially in high CO environments (Robb and Techtmann, 2018). Alternatively, that given CODH could serve as a source of metabolic intermediates to a wide range of other pathways with further ecological significance.

There is growing evidence that carbon monoxide-based metabolisms may be important and ubiquitous in diverse environments, potentially serving as a critical intermediate step in broader community-driven processes, rather than a niche form of carbon fixation only reserved for thermophilic carboxydotrophs (Cordero et al 2019; Bay et al 2021).

Carboxydotrophs and CODHs are additionally important in industrial processes including biofuel production or as CO detecting biosensors (Tissera et al., 2019, Reginald et al., 2021). Typically produced as an industrial waste gas or as synthesis gas (CO/H_2), CO can be captured and utilized by harnessing CODH-possessing acetogenic bacteria to produce useful products such as ethanol or other biofuels. Such industrial applications also serve the critically important dual purpose of offsetting carbon emissions by repurposing harmful CO and CO_2 into valuable chemicals.

Phylogenetic analyses of catalytic CODH subunit proteins have previously been used to elucidate the evolution of the protein family. Among some of the previous phylogenetic work done with CODH proteins, Techtmann et al. (2012) established a six-clade phylogenetic structure for the protein. Inoue et al. (2019) verified the established six clades and introduced a seventh clade (Clade G) with a single representative member. Inoue et al. (2019) also described more specific CooS subclades based on structural motifs. Phylogenetic analysis has also been used to better understand the ancient origin of this protein family. Adam et al. (2018) used CODH phylogeny to predict CO utilization capabilities of the LUCA based on observed evolutionary trends throughout the protein family. Complicating the evolutionary analysis of the CODH protein family, however, is that both Techtmann et al. (2012) and Adam et al. (2018) showed indications of horizontal gene transfer occurring with CODHs. These studies found that the evolution of the catalytic CODH subunit does not exclusively follow taxonomic or functional patterns; while the CODHs of clade A were found overwhelmingly in archaeal genomes, other clades were composed of more diverse collections of taxonomic groups, including both bacteria and archaea. This lack of clear taxonomic signal in CODH phylogeny was demonstrated to likely be caused by genetic transfer events (Techtmann et al 2012).

Most studies on the phylogeny and distribution of CODHs have focused on genomes of isolated bacterial groups. This has thus far biased our findings based on what can be cultured in the lab. Metagenomic sequencing has shown a vast array of uncultivated organisms that are present and important members of many environmental microbial communities (Parks et al., 2017; Nayfach et al., 2021), and many of these uncultivated phyla have been shown to contain CODHs and some posited to perform carboxydotrophy (Farag et al., 2020). Drawing upon the rapidly increasing wealth of data provided by metagenomics can be an effective way to query a broad range of otherwise inaccessible organisms, environments, and samples to expand phylogenetic analyses. However, the flexibility in metabolic outcomes of CODHs make it difficult to confidently determine the functional implications of the presence of these proteins using current metagenomic methods. Typically, metagenomic methods of protein annotation rely solely on the predicted amino acid sequence to determine a protein's putative function; genetic context data is generally not available or not considered. Without genetic context, the availability of the accessory subunits needed to indicate the fate of CO in a given CODH are unknown and the protein complex's overall function unclear. Common methods will discern - based on the presence of a catalytic CODH subunit - that the organism could perform some form of CO metabolism, but the pathways and functions that CODH facilitates will remain uncertain. As a result, specialized analysis is necessary to completely and accurately predict the full functional capabilities of CODH proteins and their metabolic implications.

Machine learning (ML) methods have previously shown promising ability to accurately annotate protein function and pathway involvement (Gligorijević et al., 2021). Machine learning is foundationally adept at pattern recognition, even without specific knowledge of recognized patterns from the model's designers. This aspect of machine learning can be leveraged to make accurate predictions based on patterns and mechanisms even when those patterns are not completely understood by researchers. It is currently not known if catalytic CODH subunit proteins have telltale sequence characteristics that determine their subunit binding and functional involvement. These proteins could potentially have specific binding sites for accessory proteins, subunit attachment sites, function-accommodating alterations in structure, or other information that is encoded in the sequence. While this information can be determined through biochemical assays, it's possible that these sequences could be detectable by a ML approach, despite lacking direct knowledge of the structure. With such an approach a complete prediction of a CODH's function could potentially be attained solely from the sequence of the catalytic subunit.

To expand the breadth of knowledge and understanding of CODH proteins and their corresponding complexes, we have identified catalytic CODH proteins in one the largest collections of metagenome assembled genomes from publicly available metagenomic datasets. Using a Hidden Markov Model, we have searched over 50,000 MAGs to identify novel putative CODH protein sequences. The distribution of originating environment and genome taxonomy of these CODHs have been analyzed to elucidate evolutionary patterns. With the goal of expanding consideration to the distribution of downstream function of these CODHs, a machine learning model was trained to predict and assign functional labels to putative catalytic CODH protein sequences. These analyses have shown a more ubiquitous nature of CODH proteins, present everywhere from thermal springs to subway benches, suggesting broader ecological implications from their involvement in microbial functional pathways.

2 Methods

2.1 Dataset

This study utilized two publicly available metagenomic datasets, the "Genomes from Earth's Microbiomes" catalog (GEMs) and the "Uncultivated Bacteria and Archaea" dataset (UBA), as representatives of a broad range of environmental samples for analysis (Figure 3). The UBA dataset (Parks et al., 2017) includes 7,903 metagenomically assembled genomes (MAGs) from 1,550 environmental samples, each with generally high reported genome quality and completeness. MAGs are species specific draft genomes constructed from unspecific community level metagenomic reads. Though they often represent incomplete genomes, MAGs are useful for targeted genomic analysis of organisms otherwise impractical or impossible to isolate. The detailed methods for the generation of the metagenome assembled genomes from the UBA are described in Parks et al. (2017). Briefly, 1,550 metagenomes from primarily environmental and non-human gut microbiome samples were individually assembled using the CLC de novo assembler (CLCBio) and binned into MAGs using MetaBAT (Kang et al., 2015). The MAGs were subsequently refined based on anomalous genomic or taxonomic signatures using CheckM (Parks et al., 2015), CompareM and RefineM (Parks et al., 2017), and only high-quality MAGs were published. The GEMs dataset (Nayfach et al., 2021) includes 52,515 MAGs from 10,450 different environmental samples. The MAGs from the GEMs dataset were generated similarly, with metagenomes from 10,331 different samples in the IMG/M database were individually assembled using metaSPAdes (Nurk et al., 2017), binned using MetaBAT (Kang et al., 2015), and quality controlled using CheckM (Parks et al., 2015) and RefineM (Parks et al., 2017). Only MAGs meeting the MIMAG standard as high-quality MAGs were kept, requiring them to have ≥90% completeness and ≤5% contamination (Bowers et al., 2017). The MAGs were clustered into OTUs based on GTDB taxonomy (Chaumeil et al., 2020). More detailed methods for the generation of the GEMs database are described in Nayfach et al. (2021).

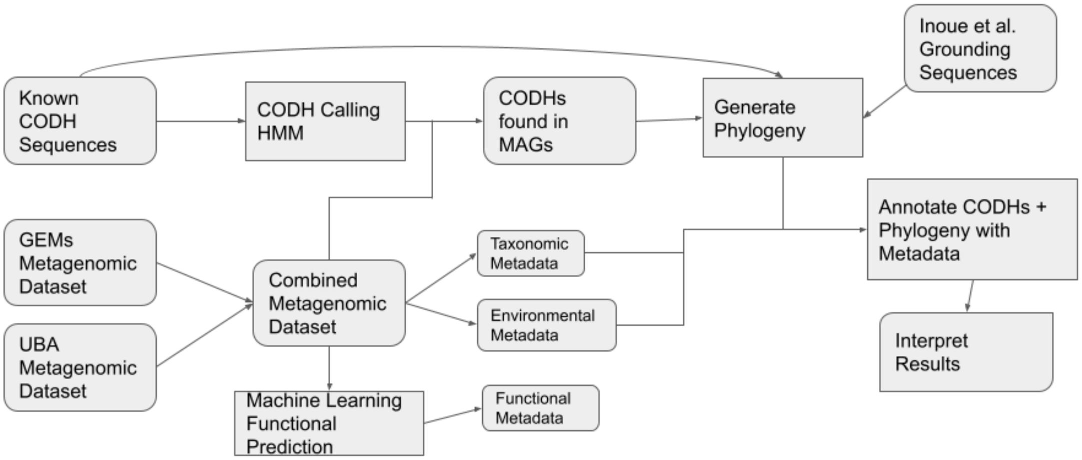

Figure 3. Schematic of the analysis pipeline used for this study. Throughout the flowchart rounded boxes represent datasets and pointed boxes indicate analysis steps.

Environmental metadata from these samples was compiled to categorize each MAG into one of 10 environmental categories slightly modified from the GOLD ecosystem classification system. The GOLD ecosystem classification is a hierarchical classification scheme used in the Genomes Online Database (GOLD) to group ecosystems into similar categories; providing a standardized framework for classifying and referring to ecosystems (Ivanova et al., 2010). In this study we have modified the GOLD ecosystem categories as follows: Ecosystem categories that were subgroups of 'Host-Associated', including 'Animal', 'Human', or 'Plant', were condensed into one broad 'Host-Associated' label. The 'Aquatic' category was subdivided into 'Marine', 'Freshwater', 'Thermal Springs', and 'Other Aquatic', to better suit the distribution of CODHs. The 'Other' category broadly includes 'engineered' GOLD ecosystem categories, such as lab enrichments or bioremediation samples. Additional metadata provided with the GEMs and UBA datasets regarding assigned MAG taxonomy was also used. The distribution of samples and their assigned environmental categories is visualized in Figure 4.

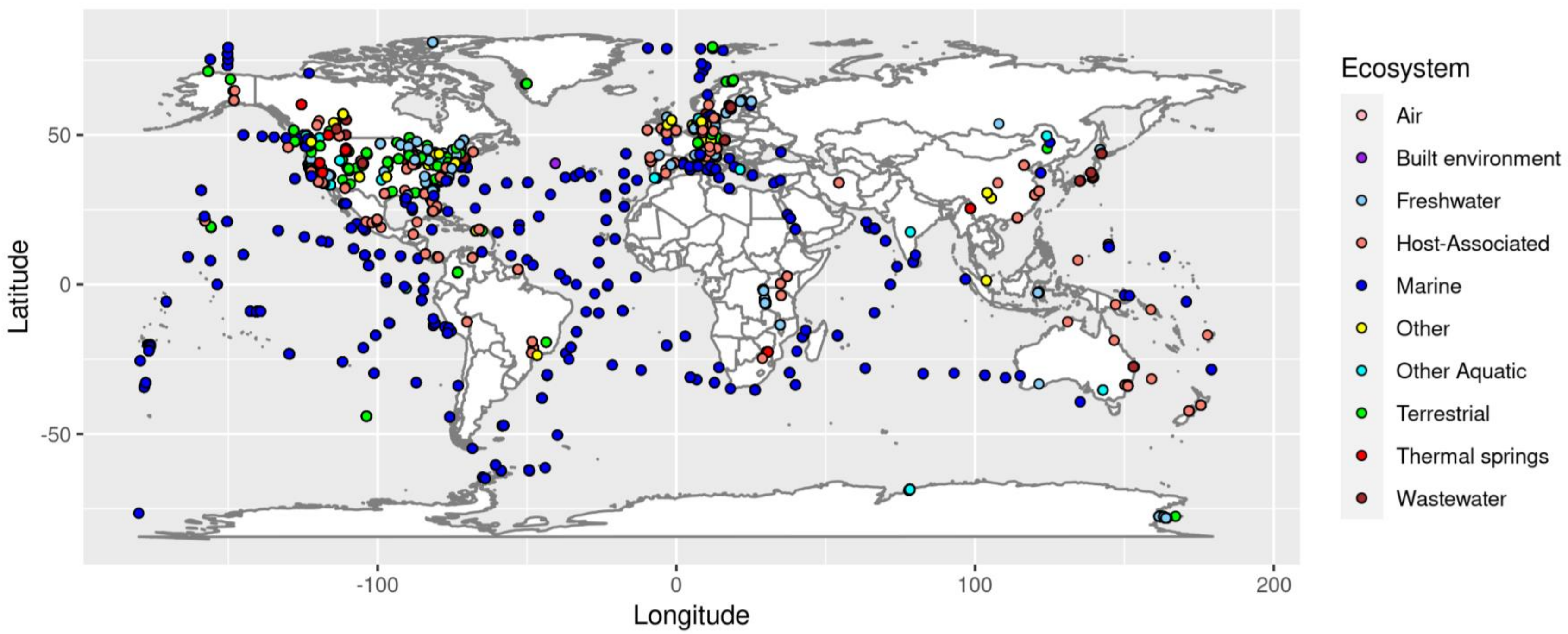

Figure 4. Map of geographic locations of all MAGs in metagenomic dataset. Where applicable, MAGs are represented by a point placed at the latitude and longitudinal coordinates supplied in the dataset's metadata. Points are colored by the metadata reported, slightly modified GOLD ecosystem categories of the MAGs. Some MAGs do not have associated geographic coordinate data

2.2 Hidden Markov Model

313 CODH amino acid sequences, representative of the known range of the protein family's variation, were curated for use as a training set for a Hidden Markov Model (HMM). Using HMMER with default parameters, a HMM was built to detect putative CODHs from input protein sequences (Eddy, 2009). The built HMM was applied with HMMER's default parameters to each of the called protein sequences available in the GEMs and UBA metagenomic datasets with an e-value cutoff of $1e^{-50}$ to identify sequences as CODHs.

2.3 Construction of Phylogenetic Tree

Putative CODHs found using the built HMM were combined with the original training sequences and CODH sequences with known phylogenic clades from Inoue et al. (2019). Inoue et al. (2019) previously constructed the most comprehensive phylogeny of the CooS gene to date. The inclusion of known sequences from the previously curated training set and from Inoue et al. (2019) ensures that the resulting phylogeny includes representatives of all known clades of CODHs, allowing our phylogeny to be grounded in and compared with previously established phylogenetic work. The combined putative and known CODH sequences were aligned using MAFFT with automatically selected parameters (Katoh et al., 2002). The resulting multiple sequence alignment was used with FastTree to generate a distance based phylogenetic tree (Price et al., 2010). FastTree was run with default parameters, meaning a JTT+CAT modelled approximated maximum likelihood tree was constructed. ggtree was used for tree visualization and metadata overlay (Yu, 2020).

2.4 Predictive ML Model

To attempt to elucidate functional pathway involvement from putative CODH sequences, a ML model was trained to identify input sequences as being involved in 'Methanogenesis', 'Acetogenesis', both, or neither. 780,014 protein sequences were used from the UniProtKB by using search parameters finding all sequences labeled with the Gene Ontology (GO) terms 'methanogenesis [15948]' or 'acetyl-CoA metabolic process [6084]'. This set of sequences included both reviewed and non-reviewed sequences. The non-reviewed sequences were assigned GO terms in the UniProtKB in an automated fashion, while the reviewed sequences were curated and manually annotated to ensure proper annotation. Additionally, we included sequences labelled as reviewed and having the GO terms 'carbon-monoxide dehydrogenase (acceptor) activity [18492]' or '4 iron, 4 sulfur cluster binding [51539]'. The GO terms 'methanogenesis' and 'acetyl-CoA metabolic process' were used during model training as indicators of methanogenesis or acetogenesis pathway involvement, respectively. 15% of the training data, chosen randomly, was held out of training for model validation.

Gligorijević recently developed a deep learning model based on Graph Convolutional Networks for automatically predicting protein function from protein sequences, termed Deep Functional Residue Identification (DeepFRI). We constructed a model using a similar structure to DeepFRI (Gligorijević et al., 2021), with a schematic of our model shown in Figure 5. The primary difference between our approach and that of DeepFRI is in the vectorization step. DeepFRI uses simple one-hot vectorization, characterizing each amino acid by a vector where a one is present in a specific location to identify the amino acid the rest of the vector is filled with zeroes. Sequences input to DeepFRI can also be supplemented with contact map information. Our model instead utilizes a novel vectorization step inspired by Asgari et al. (2015) to better capture the biological context of the protein sequence when transforming amino acid sequences into the vectorized values needed for machine learning. This vectorization step serves as the first step in our model process and involves splitting input sequences into 3 frameshifts of 3 amino acid trimers which are then vectorized using the FastText text vectorization algorithm treating each trimer as a 'word' (Bojanowski et al., 2017). In this method, each amino acid 'word' is assigned a vector based on the context the 'word' was found in throughout the training data. Based on the findings of Asgari et al. (2015), borrowing text vectorization strategies for protein vectorization can lead to proteins being vectorized in a manner more closely based on their biochemical properties, as 'words' with similar chemistry should be found in similar sequence contexts. T-distributed stochastic neighbor embedding (t-SNE) plots of vector distance colored by biochemical properties were generated and used to verify that our protein vectorization step captured additional context rooted in real amino acid biochemistry (Figure 6 and Figure 7). T-SNE plots are a form of dimensionality reduction that enables graphical representation of highly dimensional data in 2-D space, with observations of similar properties being found close together on the plot. Molecular weight and hydrophobicity were used as representative biochemical properties, with both molecular weight and hydrophobicity approximated for each trimer by summing the individual values for each constituent amino acid of the trimer. If the vectorization preserves biological information than we expect that amino acid trimers with similar biochemical properties would group together on the t-SNE plots. We chose molecular weight and hydrophobicity as representative properties as they are values that can be directly computed from amino acid trimer sequences and have biochemical significance for protein function.

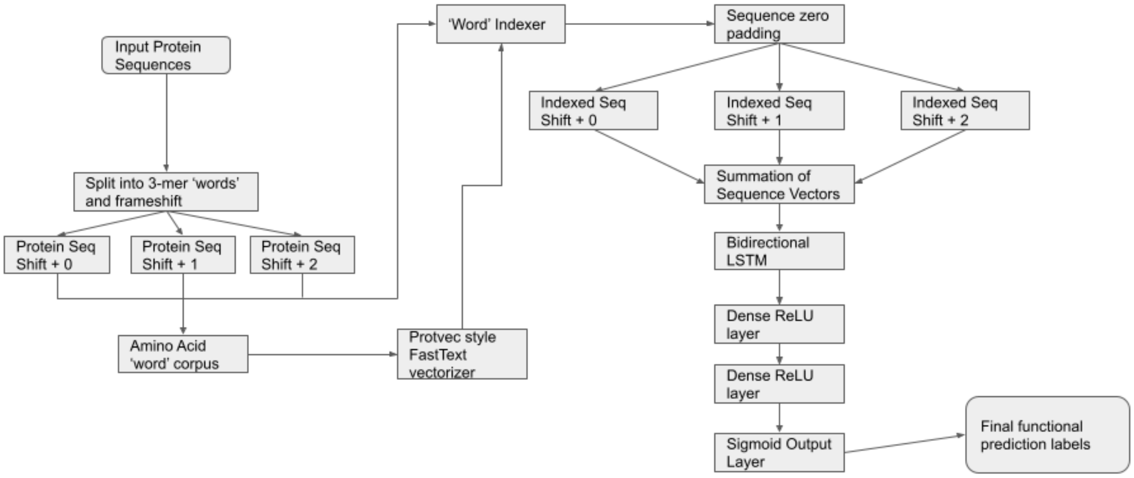

Figure 5. Schematic of the predictive model for CODH functions, starting with input amino acid protein sequences and resulting in predicted labels for the function of those sequences.

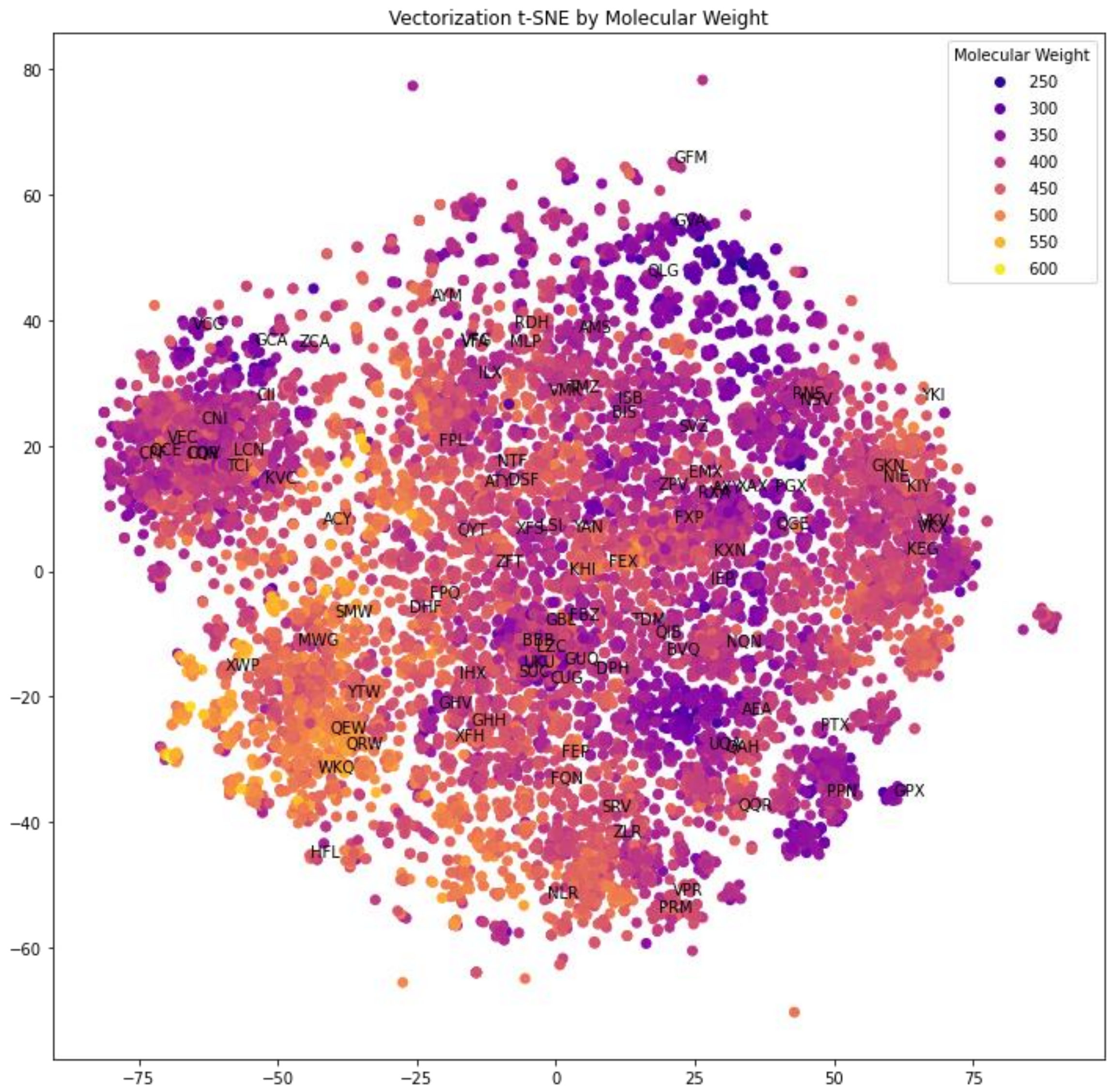

Figure 6. t-SNE plot depicting similarity of vectorized amino acid trimer ‘words’ based on biochemical property molecular weight. Similar clustering indicates that nearby ‘words’ have similar context and are interpreted to have similar ‘meaning’.

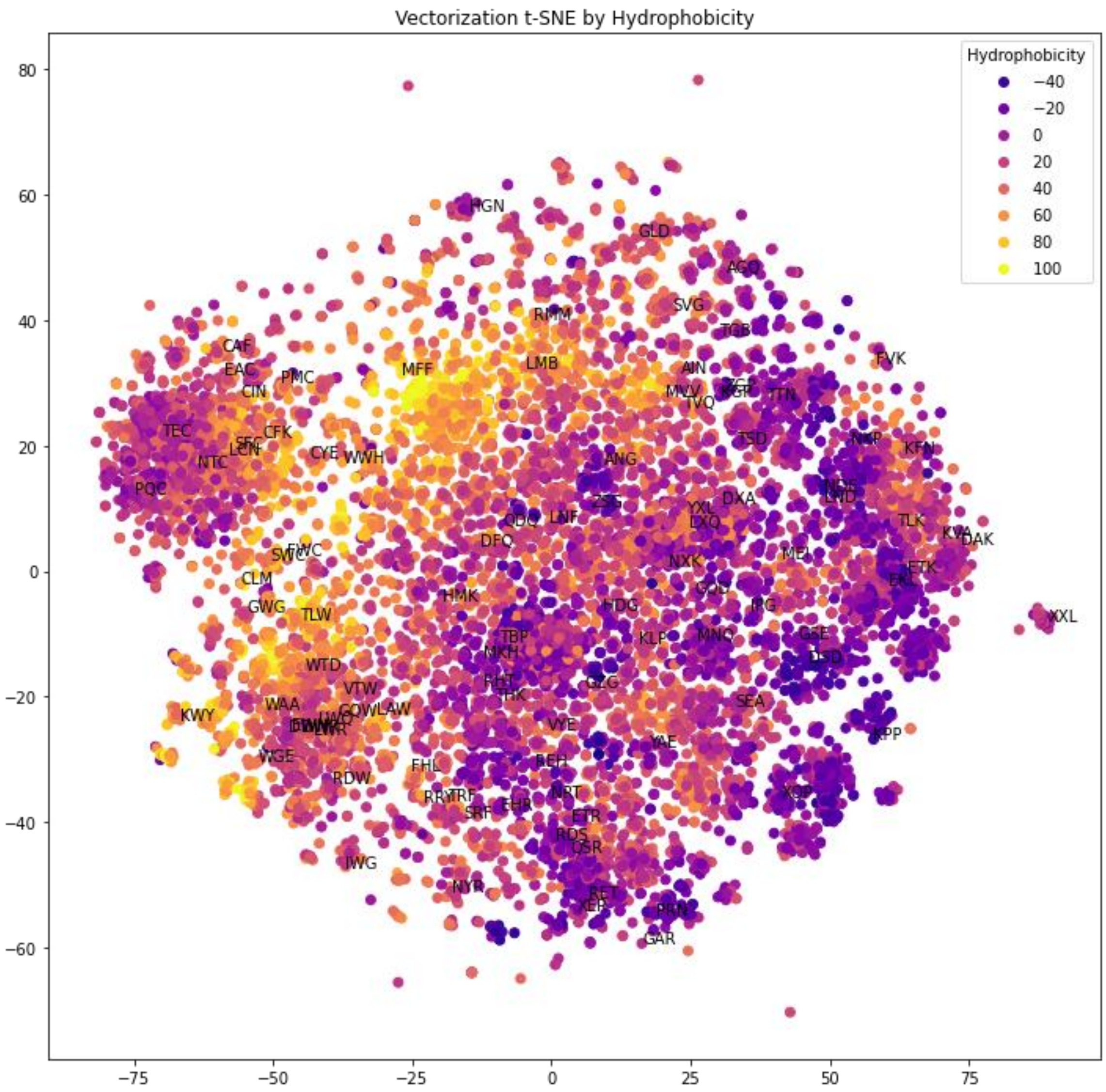

Figure 7. t-SNE plot depicting similarity of vectorized amino acid trimer ‘words’ based on biochemical property of hydrophobicity approximated by summing the individual hydrophobicity levels of each amino acid that constitute a trimer ‘word’. Similar clustering indicates that nearby ‘words’ have similar context and are interpreted to have similar ‘meaning’.

After vectorization, input sequences are zero-padded to a uniform length of 500 amino acid trimers. Sequences greater than 500 amino acid trimers, or 1500 total amino acids, were not present. Vectors calculated from each of the three possible frameshifts of the input sequence are summed together, resulting in the processed input vector that is input into a recurrent neural network (Figure 5). The recurrent neural network consists of four sequential layers: first is a bidirectional Long Short-Term Memory (LSTM) layer, followed by two Rectified Linear Unit (ReLU) activated dense layers, and finalized with a sigmoid output layer. The bidirectional LSTM layer interprets information from both the forward and reverse directions of an input sequence, as well as contextual information from distant portions of the sequence. Additionally, the LSTM layer eliminates any artifacts caused by zero-padding input sequences. The results of the LSTM layer are passed on to two dense layers, which can extract additional predictive ability from the LSTM layer. Finally, the sigmoid output layer condenses the model's prediction into a value between 0 and 1 for both ‘Methanogeneis’ and ‘Acetogenesis’ labels, where sequences assigned values greater than 0.5 are predicted to be members of that label.

The prediction model was trained in batches of 100 samples for 5 epochs (epoch = covering entire dataset once), using a binary cross entropy loss function and the Adam optimization algorithm (Kingma et al., 2014). The model was implemented in Python using the Gensim implementation of FastText and Keras and Tensorflow libraries to implement the neural network (Rehurek et al., 2010; Abadi et al., 2016). Model accuracy was validated using SciKit-Learn (Pedregosa et al., 2011).

3 Results

Searching the GEMs and UBA metagenome datasets using the CODH targeting HMM resulted in 5,426 hits. 4,960 of those hits were from the GEMs dataset, and 466 were from the UBA dataset. Of these hits, 25.3% of them were found in the same MAG as at least one other putative CODH (Table S2). Combined with the 313 protein sequences used to build the HMM and the 1,942 grounding CODH sequences from the previously constructed phylogeny described in Inoue et al. (2019), a total of 7,665 sequences were used to generate an expanded phylogeny of the CODH protein family. The distribution of CODHs found are visualized in Tables 1 and S3 and Figure 8; with Table 1 reporting the proportion of putative CODHs found in each ecosystem category, Table S3 reporting the proportion of putative CODHs grouped by their phylum-level taxonomic assignment, and Figure 8 visualizing the geographic distribution of samples containing CODHs.

Table 1. Proportion of GEMs dataset MAGs containing putative CODHs grouped by their metadata-reported, slightly-modified GOLD ecosystem labels.

Ecosystem Type	**MAG Count**	**CODH Count**	**Proportion Containing CODHs**
Air	21	0	0
Built environment	2640	70	0.026
Freshwater	7263	546	0.075
Host-Associated	20775	1120	0.053
Marine	8581	389	0.045
Other	2980	459	0.154
Other Aquatic	1860	228	0.122
Terrestrial	3356	300	0.089
Thermal springs	1545	322	0.208
Wastewater	2613	220	0.084

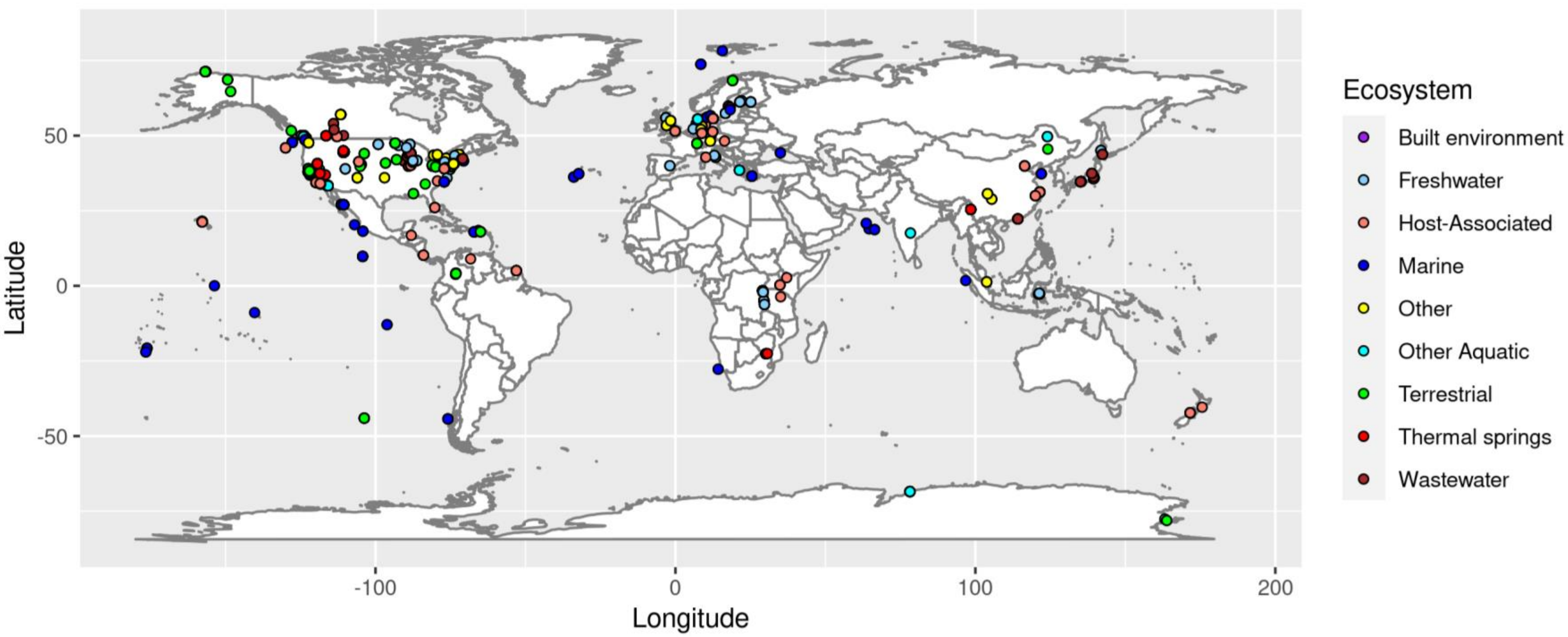

Figure 8. Map of geographic locations of all MAGs in the metagenomic dataset in which putative CODHs were found. Where applicable, MAGs are represented by a point placed at the latitude and longitudinal coordinates supplied in the dataset's metadata. Points are colored by the metadata reported, slightly modified GOLD ecosystem categories of the MAGs. Some MAGs do not have associated geographic coordinate data.

The newly generated phylogeny maintains the same clade structure as reported in Inoue et al. (2018) and initially described in Techtmann et al. (2012). Each clade includes the expected reference sequences, along with expansion from new putative CODH sequences from the MAGs. Of particular note is that no new clades were found despite the inclusion of over 5,000 new CODHs sequences. Based on the matching clades, the naming scheme will be hereafter inherited from Techtmann et al. (2012) and Inoue et al. (2018) as shown in Figure 9. Notably, clade G (graphically located just above clade A), originally represented by a single CODH in Inoue et al. (2018), is here expanded with 30 closely matching sequences found from the metagenome datasets.

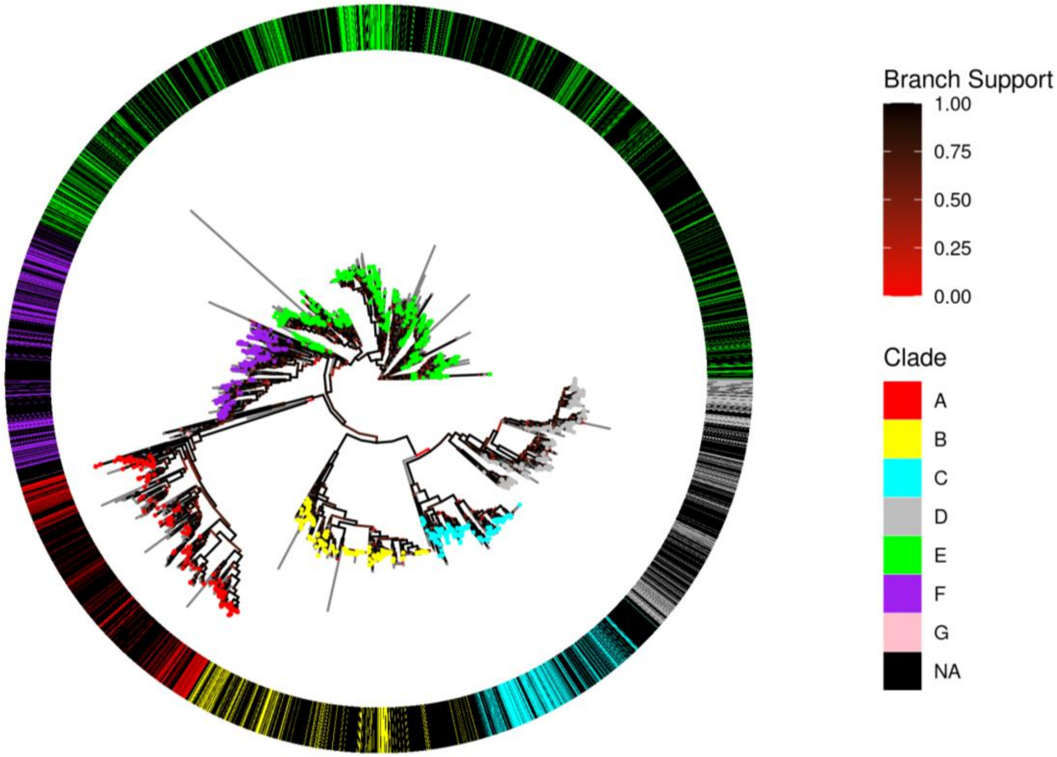

Figure 9. Phylogenetic tree of CooS/Cdh protein sequences. Tree tips and outer ring are colored by the clade assigned to the sequence by Inoue et al. (2019) where applicable. As a result, only grounding sequences are colored. Branches are colored by branch support reported by FastTree. NA represents sequences that were not previously used in the Inoue et al. (2019) tree and thus were not previously assigned to a clade. Branch supports are shown by a gradient of branch colors, with black representing 100% branch support and red representing 0% branch support. A larger version of this figure is available at https://github.com/isbigcra/ThermalSpringsAndSubwayBenches.

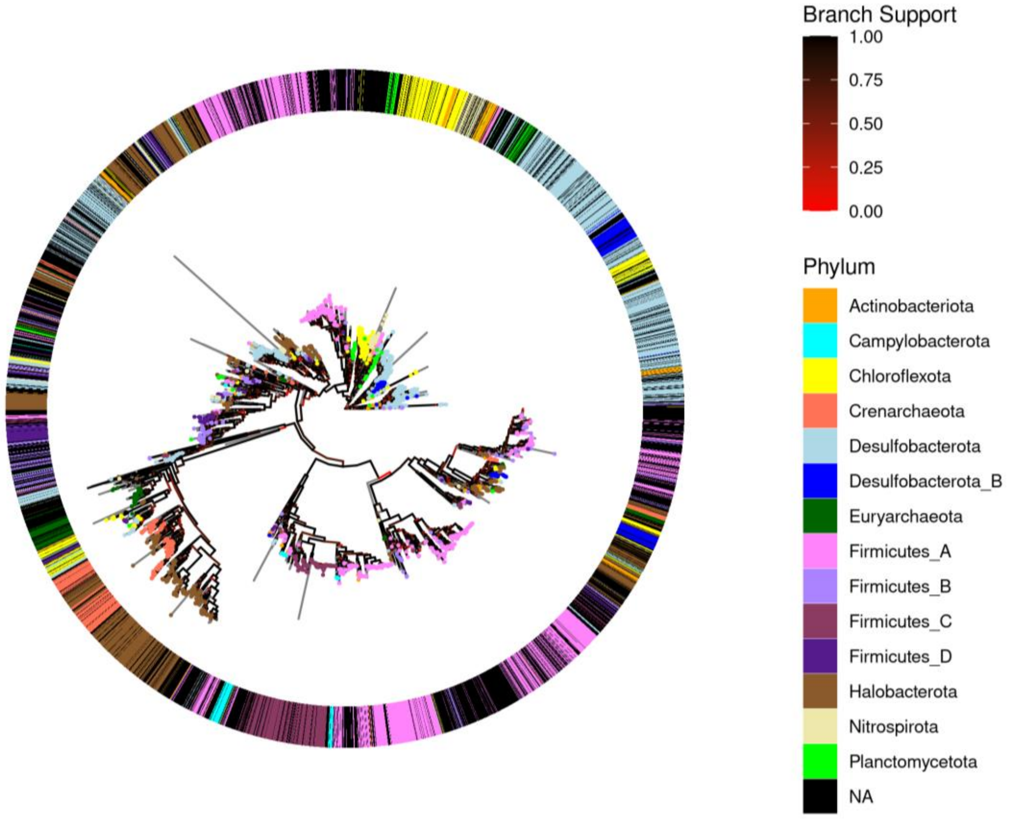

Figure 10. Phylogenetic tree of CooS/Cdh protein sequences. Tree tips and outer ring are colored by the taxonomy assigned to sequence's source MAG at the phylum level. Taxonomy of MAGs was assigned using the GTDB taxonomic classification scheme. Branch supports are shown by a gradient of branch colors, with black representing 100% branch support and red representing 0% branch support. A larger version of this figure is available at https://github.com/isbigcra/ThermalSpringsAndSubwayBenches.

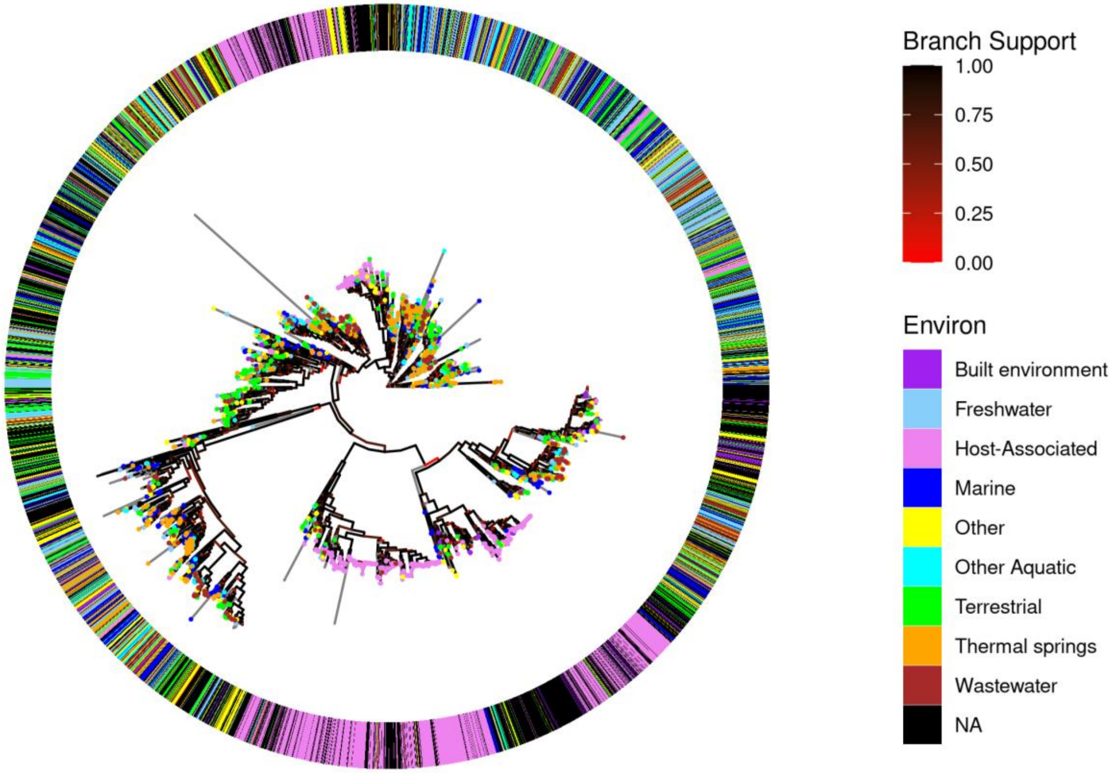

Figure 11. Phylogenetic tree of CooS/Cdh protein sequences. Tree tips and outer ring are colored by modified GOLD ecosystem category label (Environ) assigned to the sequence's source MAG. Branch supports are shown by a gradient of branch colors, with black representing 100% branch support and red representing 0% branch support. A larger version of this figure is available at https://github.com/isbigcra/ThermalSpringsAndSubwayBenches.

3.1 Taxonomic Distribution Within Clades

Matching with previously reported phylogenies, and shown in Figure 10, most protein sequences found from archaeal sources are concentrated in clade A (Techtmann et al., 2012). Archaeal phyla including Crenarchaeota, Euryarchaeota, and Halobacterota dominate the clade. Also present are sequences from some Chloroflexota, Firmicutes D, Desulfobacterota, a single Nitrospirota, and a single Planctomycetota.

Clade B is predominated by Campylobacterota, Firmicutes A, and Firmicutes C. Some Actinobacteriota sequences are included, along with a branch with similar composition to clade A. Notably, clade B includes many clusters of very similar sequences with very short branch lengths, all found in likely very similar Clostridia species from similar ‘Host-Associated’ environments. The frequency of these sequences may be more indicative of a bias towards human microbiome samples in the GEMs dataset and does not necessarily reflect the abundance of the clade.

Clade C consists mostly of protein sequences found in MAGs from the phylum Firmicutes A. A few clade C sequences are from Crenarcheaota, Euryarcheaota, Firmicutes B, Firmicutes C, and Halobacterota MAGs. Clade C also has short branch length clusters similar to clade B.

Clade D contains sequences from a more diverse set of phyla including Actinobacteriota, Chloroflexota, Crenarcheaota, Desulfobacterota and Desulfobacterota B, Euryarchaeota, Firmicutes A through D, and Halobacterota. Additionally, a single protein sequence from Planctomycetota was included.

Clade E is the largest of the clades but is less diverse than its size would suggest. The clade is mostly made up of Desulfobacterota, but also includes branches of sequences from Actinobacteriota, Chloroflexota, Desulfobacterota B, Firmicutes A B and D, Halobacterota, Nitrospirota, and Plactomycetota. Notably, clade E contains the majority of branches belonging to Nitrospirota and Planctomycetota MAGs, with both phyla typically not previously associated with carboxydotrophy. Interestingly, there is one clade E branch closely related to clade F that includes some sequences from the archaeal phyla Crenarchaeota and Euryarchaeota. This is of note due to the previously reported association of archaea with Clade A. The presence of archaeal CODHs in clades typically associated with bacteria may be an indication of horizontal gene transfer.

Clade F is mostly composed of sequences from Firmicutes B and D, with sequences from Desulfobacterota and Halobacterota are also being minor constituents of this clade. Some sequences from Chloroflexota, Firmicutes A and C, Nitrospirota, and Planctomycetota are also present. Clade F additionally includes a single sequence found in a Campylobacterota MAG.

Clade G was originally reported with a single CooS sequence from a Deltaproteobacterium MAG (Inoue et al., 2019). Our results have further populated this clade. Clade G is here mostly composed of other Desulfobacterota (previously classified as Deltaproteobacteria), along with some sequences from Firmicutes B and C, and a single sequence from Chloroflexota. This taxonomic distribution is most similar to the distribution of phyla found populating clades E and F.

3.2 Environmental Distribution Among Clades

CODH hits from the HMM were present in metagenomes from a wide variety of environments. When grouped by broad GOLD ecosystem categories, as shown in Figure 11, the different clades do not appear to have any clear ecosystem-related patterns besides a grouping of sequences from 'Host-associated' metagenomes. 'Host-associated' sequences make up a substantial majority of clades B and C, and a branch of clade E. Other clades do not have consistent patterns in reported source environment, suggesting that generalized environmental conditions do not clearly the impact evolution of CODHs. Many 'Host-associated' branches appear to be composed of very similar sequences, indicating that their clustering may be caused more by their being from MAGs of closely related species rather than being from similar 'Host-associated' environments.

3.3 Distribution of Taxa Among Environments

Among putative CODH containing MAGs from 'Host-associated' environments, Firmicutes A and C dominated (Figure 12). These MAGs were almost entirely Clostridia from human digestive system metagenomes. Firmicutes A and C also made up the majority of sequences from ‘built environment’ MAGs, which were largely metagenomes from New York City subway benches collected by Afshinnekoo et al. (2015), likely human associated, and also belonging to the class Clostridia. CODH protein sequence hits from Chloroflexota, Desulfobacterota, and Halobacterota were found to be abundant in all environmental categories excluding ‘Host-associated’ (Figure 13). Crenarchaeota sourced CODH sequences were primarily found in 'Thermal Springs' environments and Actinobacteriota sequences were primarily in 'Terrestrial' environments. Euryarchaeota CODHs were markedly present in wastewater and bioreactor metagenomes.

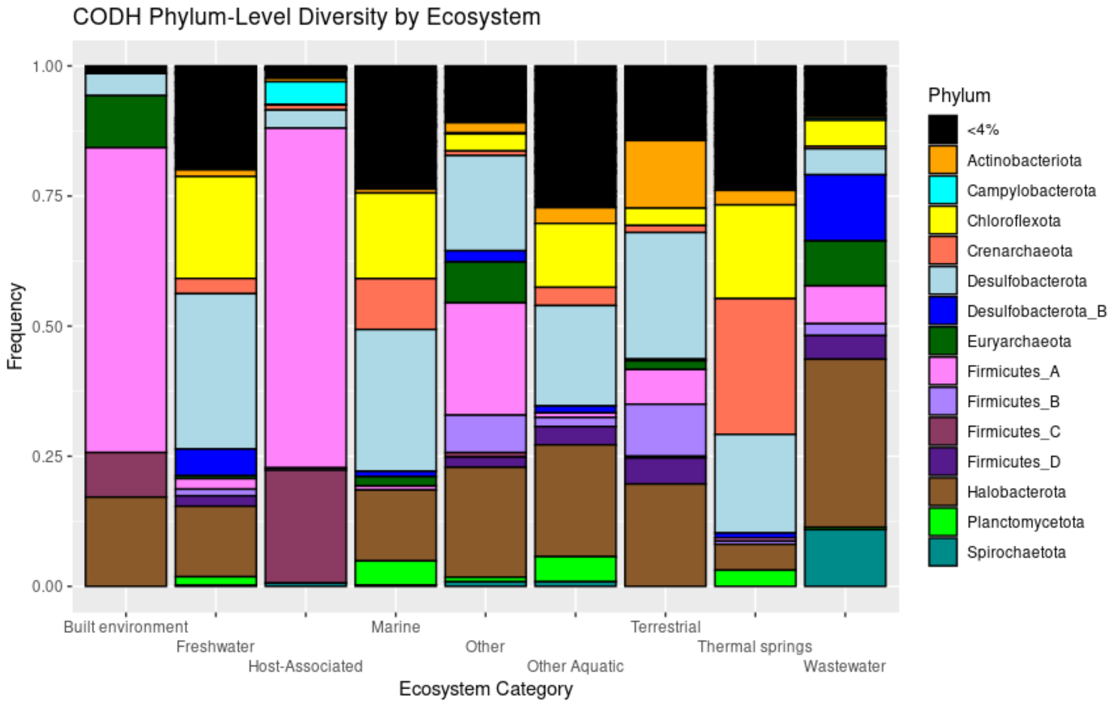

Figure 12. Bar chart showing taxonomic diversity of CODHs among different GOLD ecosystem categories at the phylum level. Frequency indicates the count of CODH containing MAGs for given conditions.

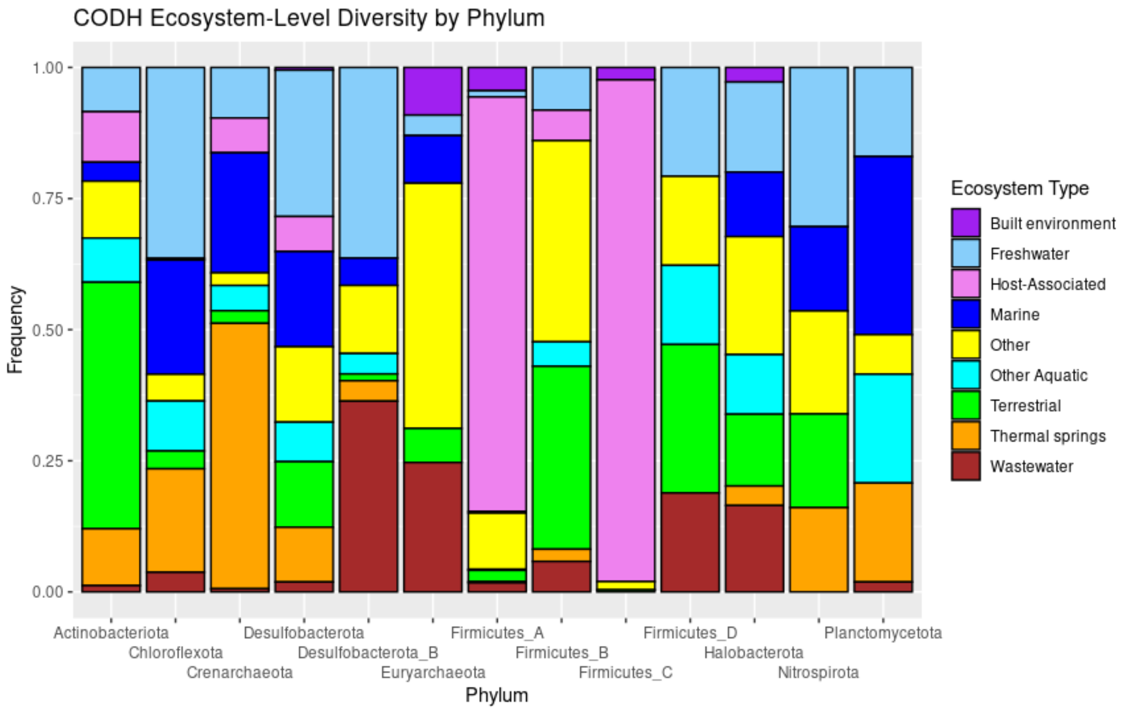

Figure 13. Bar chart showing diversity of source ecosystems among different CODHs at the phylum level. Frequency indicates the count of CODH containing MAGs for given conditions.

3.4 Performance of Recurrent Neural Network Model

After the initial training epoch, the GO label predictive machine learning model achieved a binary accuracy of 98.98% on validation data. After the fifth and final training epoch, the model reported a binary accuracy of 99.4%. Verifying the model on the held out 15% of the training data yielded an overall accuracy of 98.8%, with generally high precision, recall, and f1 score metrics for both 'Methanogenesis' and 'Acetyl-CoA metabolic process' labels. Accuracy, precision, recall, and f1 are all standard metrics for assessing performance of machine learning models. These metrics suggest the model to be high performing and capable of accurately predicting assignment of the two GO labels for input protein sequences.

Applying the machine learning predictions to the putative CODH protein sequences yielded unexpected results, however. Despite methanogenesis pathways being known as exclusive to Archaea, the model assigned 'Methanogenesis' labels to sequences from both Bacteria and Archaea across the phylogenetic tree. Approximately 11% of all input CooS sequences were predicted with the 'Methanogenesis' label, with no clear patterns in taxonomy, phylogeny, or source environment among those predictions. The 'Acetyl-CoA metabolic process' label, despite not being known to be limited to Archaea, appeared to be noticeably biased towards being assigned to archaeal sequences by the model. Almost all the sequences assigned this label were concentrated in clade A or in a clade F branch of predominantly Halobacterota sequences. 47.6% of Archaeal sequences were assigned the 'Acetyl-CoA metabolic process' label by the model, contrasted with only 2.5% of bacterial sequences being assigned the label. 25% of sequences from 'Thermal Springs' were assigned the label, with other ecosystem categories having around 18% of their sequences labelled. Some of the sequences included in the phylogeny from non-metagenomic sources were also included in the ML model training data. These sequences appeared to be labelled with appropriate GO terms by the model, as expected.

4 Discussion

In this study putative CODH protein sequences were found in MAGs from a wide range of environments and phyla, extending their distribution far beyond the extreme environments and extremophilic organisms CODHs are often associated with. The presence of CODHs in diverse environments such as marine and freshwater, soils and deep subsurface, digestive systems, thermal springs, wastewater, and subway benches suggests a broader ecological significance beyond an ancient yet fringe carbon fixation pathway. Due to the versatile nature of the enzyme complex, however, it is currently unclear which pathways are being utilized in each of these environments. As many of the environments CODH encoding MAGs were present in do not have obvious sources of abundant CO, we speculate that these microbes may instead be using biogenically sourced CO produced by other organisms. Previous work has shown that both methanogens and sulfate reducers can produce CO as part of their metabolism, providing plausible sources of such biogenic CO (Voordouw et al., 2002).

Many MAGs additionally encode multiple different putative CODH proteins (Table S2), a phenomenon also observed in previous studies, further suggesting that these CODHs may be contributing to a variety of different pathways within a given environment (Wu et al., 2005; Techtmann et al., 2012). Each of the CODH copies, in most cases, appeared in different clades despite being from the same MAG; providing further evidence that CODHs could be involved in multiple different pathways, even within a single organism. The presence of multiple CODHs within a MAG, each found in different clades, additionally provides support for the assertion that CODHs have undergone horizontal gene transfer events. If these sequences were the result of gene duplication, the sequences would be phylogenetically very similar and therefore would be expected to be found within the same clade. Additionally, such horizontal gene transfer events have been previously reported for CODHs (Techtmann et al., 2012; Adam et al., 2018). It is also possible, however, that cases where a MAG appears to encode an extra CODH with no close same-phylum relatives could instead be the product of metagenomic binning artifacts or chimera sequences. Addressing this, the datasets used only included medium and high quality MAGs, which according to the MIMAG standard must have less than 10% contamination (Bowers et al., 2017). The overall mean contamination of MAGs in the GEMs dataset was 1.3%, indicating that binning artifacts and chimeric MAGs are unlikely in the dataset (Nayfach et al., 2021). Cases where a multitude of similar MAGs exhibit this phenomenon are therefore much more likely to be the result of gene transfer events.

The observed trend of large clustering of CODH sequences from host-associated metagenomes is likely due to a bias in the GEMs dataset towards human host-associated metagenomes. Many of the host-associated clusters are not deeply branching, suggesting that the member sequences of the cluster are very similar with very small branch lengths between tips in those clades. Most of these host associated CODHs were found in Clostridia MAGs. Considering these effects of sampling bias, there do not appear to be any other clear trends between clade membership and environmental source. This suggests that the evolution of CODHs is generally independent of environment, and that although CODHs from different clades could perform different functions, those functions do not appear to be restricted or driven by the ecosystem category in which they are found. It is still possible that more specific environmental parameters, such as temperature, light, or oxygen levels, which could vary within an ecosystem category, could direct CODH evolution. As these kinds of environmental parameters are known to influence microbial activity, they are a clear next step in investigating CODH evolution.

Beyond generalized trends, the specific phyla to which putative CODH containing MAGs belong are also noteworthy. Many CODHs were found in MAGs from known and expected carboxydotrophic phyla, such as Halobacterota, Desulfobacterota, and Chloroflexota. CODHs were also found in MAGs from phyla that contain members only recently considered to perform CO metabolism, such as Actinobacterota and Planctomycetota (Jiao et al., 2021). Unlike the aforementioned carboxydotrophic phyla including Halobacterota or Desulfobacterota, the mechanisms and function of CODH activity is still not completely understood in Actinobacterota and Planctomycetota.

CODHs were additionally found in MAGs from multiple poorly characterized phyla, including Armatimonadetes, Omnitrophota, Bipolaricaulota, and others. While the presence of putative CODH proteins does not directly reveal the functional capabilities of members of these phyla, the ability to encode CODHs lends strong support to hypotheses that these organisms have the potential to use CO as a metabolic intermediate in their metabolism. It’s possible that these CODHs could be used in performing hydrogenogenesis, methanogenesis, or acetogenesis through CO metabolism and the Wood-Ljungdahl pathway. More targeted research will be necessary to more completely elucidate the metabolic abilities of these organisms and how they are using CO.

To explore solely using sequences to identify metabolic potential, we constructed a predictive model to predict GO-terms from amino acid sequence. The GO-term predictive model, from available metrics, is highly effective at accurately labelling protein sequences with 'Methanogenesis' and 'Acetyl-CoA metabolic process' GO pathway labels. Additionally, the t-SNE plots suggest that the initial vectorization step is able to capture meaningful biological context from input amino acid trimer 'words'. This biological context is passed as input for the recurrent neural network portion of the model, which can continue maintenance of contextual information, as well as further considering context from position in the input sequence. As such, the model's predictions likely have some biological foundation beyond simple sequence similarity. While other similar machine learning studies have generally been successful at prediction of protein characteristics, including GO labels, and other biological predictive tasks (Gligorijević et al., 2021), applying text vectorization strategies to amino acid sequences is a novel approach to provide additional biological context to protein identification machine learning models.

Although the model appears to be quite effective at assigning sequences into the two GO terms of methanogensis and acetogenesis, when its predictions were applied to putative CODH sequences from MAGs, the results did not match expectations. It was expected that CODHs involved in methanogenesis would be restricted to the archaea, as methanogenesis via CODHs is exclusively found in the archaea. Conversely, we expected that acetogenesis would be distributed amongst both the bacteria and the archaea. Our results, however, show that the methanogenesis label was given to sequences from both archaea and bacteria. While it is possible that the predictions are still highly accurate and highly unexpected, it is much more likely that the task assigned to the model and the task intended for the model did not match. The data used to train the model to predict GO labels may not have been high enough quality, and the model may have learned patterns and biases outside of the labels' intended meanings. Both 'Methanogenesis' and 'Acetyl-CoA metabolic process' labels are GO pathway labels, which incorporate proteins from the entirety of the pathway, potentially resulting in a too-diverse collection of sequences. Additionally, a minority of Uniprot sequences are manually and confidently assigned GO labels and a majority of sequences are assigned labels with an automated method based on similarity. Although automated assignments are generally reliable, it is possible for automated annotation to mislabel some proteins. We used both manually and automatically annotated sequences due to the limited number of sequences within these two GO term categories that had been manually curated. The quantity and diversity of database sequences is limited to the number of experiments done to find them; proteins less frequently studied - CODHs among them - will have poorer database representations. For these described reasons, the training data used for the model may have led it to ultimately make erroneous 'Methanogenesis' label predictions. The 'Acetyl-CoA metabolic process' label appears to still be effective at predicting Wood-Ljungdahl pathway functionality at some level, although that is distinct from the intended scope of the label.

DeepFRI, a similar neural network model for predicting GO labels for input protein sequences, gave similarly erroneous results to the predictive model described here when applied to putative CODH sequences. Scrutinizing the training data used by DeepFRI revealed that CODH proteins, and proteins with related functions, were completely absent. Without similar sequences in the dataset, and therefore having complete training data appropriate for the assigned task, it is unreasonable to expect meaningful predictions from the model. As an additional but likely less impactful confounding factor, both GO-term labelling predictive models were trained using complete sequences from Uniprot but were applied to proteins from MAGs that may not be complete, further reducing model performance.

With CO metabolism proposed as the ancient carbon fixation mechanism for the LUCA, understanding the mechanisms and evolution of this pathway is key to understanding early evolution on Earth or in astrobiological contexts. More fully understanding these mechanisms will require further research experimentally characterizing and confirming the broad diversity present in CODHs shown in this study. For example, CODH encoders, or their CODH proteins, could be individually isolated and analyzed to confirm which pathways, if any, these complexes are involved in throughout non-thermophilic communities. Additionally, as CODHs are able to reversibly catalyze the reaction of CO → CO_2, understanding if these diverse CODH enzymes are primarily using CO or CO_2 as their input is important to deduce the kinds of metabolic pathways these enzymes facilitate (Hadj-Saïd et al., 2015). Current sequence based approaches, as previously described, are lacking in their ability to answer these functional questions for CODHs; based solely on sequence data, neither the directionality or pathway involvement of a CODH can be readily discerned. This highlights the need for further experimental data to elucidate the mechanisms of diverse CODHs and to inform the development of more advanced sequence based approaches to their prediction.

Machine learning has been shown to be a promising approach to predicting protein functions, but high-quality training data is severely lacking for CODHs and their range of functions. Experimentally collecting a large dataset of ideal training data based on biochemical data would be an impossibly monumental task due to the difficulty of isolating these organisms, challenges associated with high throughput protein expression and purification, among other challenges. So a different approach to model training may be necessary. Future models may instead rely on additional information from genomic context, active sites, or predicted protein folding from programs such as Alphafold (Jumper et al., 2021) to make predictions; with genomic context likely being the most feasible information source to incorporate. Future models predicting CODH function may also rely more on data clustering rather than defined label categorization, limiting the depth of training data necessary. With a known set of possible CODH functions, putative CODH sequences could hypothetically be clustered based on sequence features or recruitment to limited representative sequences; each cluster being indicative of one of the possible functions.

The expansion of CODH phylogeny with over 5000 putative CODHs performed in this study has provided substantial support for the previously established clades of the CooS protein. The broad diversity in taxa and environments of MAGs found to encode these CODHs provides a basis for future research to investigate the functions of CODH enzymes beyond the Wood-Ljungdahl pathway in extremophiles. This diversity also further illustrates the need for understanding the ecological role of CODHs; an enzyme found everywhere from thermal springs to subway benches is sure to be biologically significant.

5 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Adam, P. S., Borrel, G., & Gribaldo, S. (2018). Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proceedings of the National Academy of Sciences*, *115*(6), E1166-E1173.

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., ... & Mason, C. E. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell systems*, *1*(1), 72-87.

Asgari, E., & Mofrad, M. R. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, *10*(11), e0141287.

Bay, S. K., Dong, X., Bradley, J. A., Leung, P. M., Grinter, R., Jirapanjawat, T., ... & Greening, C. (2021). Trace gas oxidizers are widespread and active members of soil microbial communities. *Nature Microbiology*, *6*(2), 246-256.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology*, *35*(8), 725-731.

Byrne-Bailey, K. G., Wrighton, K. C., Melnyk, R. A., Agbo, P., Hazen, T. C., & Coates, J. D. (2010). Complete genome sequence of the electricity-producing "Thermincola potens" strain JR. *Journal of bacteriology*, *192*(15), 4078-4079.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925.

Cordero, P. R., Bayly, K., Man Leung, P., Huang, C., Islam, Z. F., Schittenhelm, R. B., ... & Greening, C. (2019). Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *The ISME journal*, *13*(11), 2868-2881.

Eddy S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics*, *23*(1), 205–211.

Farag, I. F., Biddle, J. F., Zhao, R., Martino, A. J., House, C. H., & León-Zayas, R. I. (2020). Metabolic potentials of archaeal lineages resolved from metagenomes of deep Costa Rica sediments. *The ISME journal*, *14*(6), 1345-1358.

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., ... & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature communications, 12*(1), 1-14.

Hadj-Saïd, J., Pandelia, M. E., Léger, C., Fourmond, V., & Dementin, S. (2015). The carbon monoxide dehydrogenase from Desulfovibrio vulgaris. *Biochimica Et Biophysica Acta (BBA)-Bioenergetics*, *1847*(12), 1574-1583.

Inoue, M., Nakamoto, I., Omae, K., Oguro, T., Ogata, H., Yoshida, T., & Sako, Y. (2019). Structural and phylogenetic diversity of anaerobic carbon-monoxide dehydrogenases. *Frontiers in microbiology*, *9,* 3353.

Ivanova, N., Tringe, S. G., Liolios, K., Liu, W. T., Morrison, N., Hugenholtz, P., & Kyrpides, N. C. (2010). A call for standardized classification of metagenome projects. *Environmental Microbiology*, *12*(7), 1803-1805.

Jiao, J. Y., Fu, L., Hua, Z. S., Liu, L., Salam, N., Liu, P. F., ... & Li, W. J. (2021). Insight into the function and evolution of the Wood–Ljungdahl pathway in Actinobacteria. *The ISME Journal*, *15*(10), 3005-3018.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *3*, e1165.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), 3059-3066.

King, G. M., & Weber, C. F. (2007). Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nature Reviews Microbiology*, *5*(2), 107-118.

King, G. M. (2015). Carbon monoxide as a metabolic energy source for extremely halophilic microbes: implications for microbial activity in Mars regolith. *Proceedings of the National Academy of Sciences*, *112*(14), 4465-4470.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., ... & Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature biotechnology*, *39*(4), 499-509.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, *27*(5), 824-834.*

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, *25*(7), 1043-1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., ... & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, *2*(11), 1533-1542.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.

Reginald, S. S., Etzerodt, M., Fapyane, D., & Chang, I. S. (2021). Functional Expression of a Mo–Cu-Dependent Carbon Monoxide Dehydrogenase (CODH) and Its Use as a Dissolved CO Bio-microsensor. *ACS sensors*, *6*(7), 2772-2782.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.

Robb, F. T., & Techtmann, S. M. (2018). Life on the fringe: microbial adaptation to growth on carbon monoxide. *F1000Research*, *7*.

Russell, M. J., Barge, L. M., Bhartia, R., Bocanegra, D., Bracher, P. J., Branscomb, E., ... & Kanik, I. (2014). The drive to life on wet and icy worlds. *Astrobiology*, *14*(4), 308-343.

Svetlitchnyi, V., Peschel, C., Acker, G., & Meyer, O. (2001). Two membrane-associated NiFeS-carbon monoxide dehydrogenases from the anaerobic carbon-monoxide-utilizing eubacterium Carboxydothermus hydrogenoformans. *Journal of Bacteriology*, *183*(17), 5134-5144.

Techtmann, S. M., Colman, A. S., & Robb, F. T. (2009). 'That which does not kill us only makes us stronger': the role of carbon monoxide in thermophilic microbial consortia. *Environmental microbiology*, *11*(5), 1027-1037.

Techtmann, S. M., Colman, A. S., Murphy, M., Schackwitz, W., Goodwin, L., & Robb, F. T. (2011). Regulation of multiple carbon monoxide consumption pathways in anaerobic bacteria. *Frontiers in microbiology*, *2*, 147.

Techtmann, S., Colman, A. S., Lebedinsky, A. V., Sokolova, T. G., & Robb, F. T. (2012). Evidence for horizontal gene transfer of anaerobic carbon monoxide dehydrogenases. *Frontiers in microbiology*, *3*, 132.

Tissera, S. D., Köpke, M., Simpson, S. D., Humphreys, C., Minton, N. P., & Dürre, P. (2017). Syngas biorefinery and syngas utilization. *Biorefineries*, 247-280.

Voordouw, G. (2002). Carbon monoxide cycling by Desulfovibrio vulgaris Hildenborough. *Journal of bacteriology*, *184*(21), 5903-5911.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., & Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature microbiology*, *1*(9), 1-8.

Wu, M., Ren, Q., Durkin, A. S., Daugherty, S. C., Brinkac, L. M., Dodson, R. J., ... & Eisen, J. A. (2005). Life in hot carbon monoxide: the complete genome sequence of Carboxydothermus hydrogenoformans Z-2901. *PLoS genetics*, *1*(5), e65.

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current protocols in bioinformatics*, *69*(1), e96.

6 Supplemental Data

Higher resolution PDF formatted versions of Figures 9 through 11 are available at https://github.com/isbigcra/ThermalSpringsAndSubwayBenches. Compared to the versions presented here, the PDF versions allow for much more magnification to more fully and effectively explore the breadth of data presented.

Table S2. Number of MAGs containing multiple putative CODHs.

Number of CODHs in MAG	**MAG Count**
1	3018
2	745
3	202
4	63
5	12

Table S3. Proportion of GEMs dataset MAGs containing putative CODHs grouped by their assigned phylum-level taxonomy, sorted by number of CODHs found. Taxonomy was originally assigned to MAGs using GTDB (Chaumeil et al., 2020).

Phylum-level Taxon	**MAG Count**	**CODH Count**	**Proportion of CODHs Found**
Firmicutes A	8376	925	0.110
Desulfobacterota	975	584	0.598
Halobacterota	764	431	0.564
Chloroflexota	1141	294	0.257
Firmicutes C	1222	254	0.207
Crenarchaeota	1040	166	0.159
Firmicutes B	184	86	0.467

Phylum-level Taxon	**MAG Count**	**CODH Count**	**Proportion of CODHs Found**
Actinobacteriota	4009	83	0.020
Desulfobacterota B	87	77	0.885
Euryarchaeota	137	77	0.562
Nitrospirota	167	56	0.335
Firmicutes D	84	53	0.630
Planctomycetota	902	53	0.058
Campylobacterota	364	51	0.140
Bacteroidota	8868	41	0.004
Spirochaetota	533	39	0.073
Thermoplasmatota	621	35	0.056
Bipolaricaulota	72	26	0.361
Omnitrophota	259	26	0.100
Acidobacteriota	742	24	0.032
Aquificota	53	22	0.415
Desulfuromonadota	84	16	0.190
Armatimonadota	155	14	0.090
Verrucomicrobiota	1382	14	0.010
Proteobacteria	10606	11	0.001
Altiarchaeota	26	10	0.384
Firmicutes F	37	9	0.243
RBG-13-61-14	10	9	0.900
Aerophobetota	13	8	0.615
Unspecified Bacteria	39	8	0.205
Fibrobacterota	68	8	0.117

Phylum-level Taxon	**MAG Count**	**CODH Count**	**Proportion of CODHs Found**
Hadesarchaeota	20	8	0.400
JdFR-18	9	7	0.777
WOR-3	105	7	0.066
Elusimicrobiota	62	6	0.096
KSB1	52	6	0.115
UBP3	15	6	0.400
Asgardarchaeota	10	5	0.500
Firmicutes G	142	5	0.035
MBNT15	17	5	0.294
Firmicutes E	45	4	0.088
Latescibacterota	41	4	0.097
Micrarchaeota	130	4	0.030
Myxococcota	269	4	0.014
Nitrospinota	44	4	0.090
UBP10	86	4	0.046
AABM5-125-24	21	3	0.142
Caldisericota	55	3	0.054
CG03	3	3	1.000
Deferribacterota	9	3	0.333
Desantisbacteria	3	3	1.000
Eremiobacterota	10	3	0.300
Hydrogenedentota	46	3	0.065
Nanoarchaeota	258	3	0.011
Synergistota	152	3	0.019

Phylum-level Taxon	**MAG Count**	**CODH Count**	**Proportion of CODHs Found**
TA06	17	3	0.176
DTU030	5	2	0.400
Firestonebacteria	5	2	0.400
Firmicutes	1936	2	0.001
Firmicutes H	27	2	0.074
Methylomirabilota	19	2	0.105
Moduliflexota	2	2	1.000
UBA3054	11	2	0.181
UBA9089	2	2	1.000
UBP7	14	2	0.142
Bdellovibrionota	193	1	0.005
Caldatribacteriota	44	1	0.022
Calditrichota	8	1	0.125
CG2-30-53-67	1	1	1.000
Coprothermobacterota	16	1	0.062
Eisenbacteria	22	1	0.045
GWC2-55-46	3	1	0.333
Margulisbacteria	35	1	0.028
Patescibacteria	2222	1	0.0005
Poribacteria	8	1	0.125
Riflebacteria	15	1	0.066
SAR324	239	1	0.004
Thermotogota	201	1	0.004
UBP1	18	1	0.056

Phylum-level Taxon	**MAG Count**	**CODH Count**	**Proportion of CODHs Found**
UBP18	1	1	1.000
UBP6	21	1	0.047