Understanding Rumen Methanogen Interactions in Sheep Using Machine Learning

Katharina Dost[1,2] (✉), Steffen Albrecht[2], Paul Maclean[3], Jörg Wicker[2], and Sandeep K Gupta[3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia katharina.dost@ijs.si
[2] University of Auckland, Auckland, New Zealand
{steffen.albrecht,j.wicker}@auckland.ac.nz
[3] AgResearch, Grasslands, New Zealand
{paul.maclean,sandeep.gupta}@agresearch.co.nz

Abstract. Methane emissions from livestock pose a significant challenge globally, particularly in countries with a strong farming industry dominated by sheep farming, such as Aotearoa, New Zealand (NZ). Chemical inhibitors such as feed additives or vaccines help to decrease methane emissions. However, their successful development has been hindered by a limited understanding of the complex interactions among the microorganisms in the rumen (forestomach). This study serves as a proof-of-concept to explore the potential of using metatranscriptome data to understand the genetic basis of microbial interactions in the rumen and identify potential inhibitor targets. We analyzed a small but carefully curated dataset of 10 sheep emitting different levels of methane. We employed various statistical and machine learning techniques to uncover new contigs (continuous sequences of DNA) linked to high levels of methane output. Despite the limited sample size, our findings revealed new insights into microbial mechanisms, validated by domain experts. These preliminary results suggest that expanding the dataset and integrating machine learning can enhance our understanding of the complex microbial interactions in the rumen, ultimately contributing to the development of effective strategies to reduce methane emissions in livestock.

Keywords: Livestock Methane Emission · Chemical Inhibitors · Microbial Interactions · Applied Machine Learning.

1 Introduction

Methane emissions from farmed animals pose a significant environmental challenge, contributing to global warming [23]. In Aotearoa New Zealand (NZ), these emissions are particularly pronounced, with livestock methane accounting for $\sim 35\%$ of the country's greenhouse gas output [13].

To reduce the methane output in livestock, researchers have tested targeted breeding, leading to successfully lowered emissions; however, the genetic long-term impacts are uncertain [13]. An alternative is the administration of chemical inhibitors – substances that slow down or completely stop chemical reactions or

biological processes – via feed additives [21] or vaccines [1] that specifically target the growth of methanogens. Methanogens are the main organisms responsible for producing methane in ruminants, but they rely on other microorganisms in the rumen for their survival. Therefore, understanding their interactions with other microorganisms in the rumen will help develop new ways to target them.

Developing an effective vaccine or feed additive is now a key goal for scientists, industry, and the government in NZ, but it has proven challenging due to limited knowledge of the complex interactions of the methanogens with the other microbial population in the rumen. Machine learning could help in dissecting these complex interactions and identify the specific genes of methanogens responsible for their interactions with other microbes in the rumen. The understanding of these complex genetic interactions can lead to the development of novel avenues to target methane production in ruminants.

As a proof-of-concept, we use a sheep rumen metatranscriptome dataset [26] gathered in NZ (10 sheep x 2 sampling days, yielding a total of 20 samples) to enhance our understanding of rumen microbial interactions and to identify promising contigs, continuous sequences of DNA, for further investigation. Although small for machine learning tasks, the sample size is considered large in the field, and it is sufficient as the involved sheep have been hand-selected for this task – the dataset contains sheep with low, intermediate and high methane output, which enables us to investigate the differences in interactions between contigs. This is the first study using metatranscriptome data in a sheep rumen context, but it has shown great potential for the human microbiome [28]. To the best of our knowledge, it is also the first study to apply advanced machine learning approaches to analyze metatranscriptome data from the rumen of low/intermediate/high methane-yielding sheep in general. Particularly, this is also the first study to analyze our dataset. Upon success, this study may lead to a substantially larger sample collection, yielding the foundation for inhibitor development.

Particularly, we seek to answer the following questions:

1. Hypothesis and data validation: Is there a connection between contig counts and methane output in sheep? Is it manifested in our sample?
2. Narrowing down the search: Which contigs play a role in methane production beyond methanogens?
3. Understanding patterns: Are there groups of contigs that act together?
4. Identifying causal relationships in the rumen: Are some of the identified relationships causal?

To address these questions, we employ various statistical and machine learning techniques to uncover potential drivers responsible for low or high methane production in the same breed of sheep. Despite challenges due to the small sample size, our analysis managed to provide interesting and promising insights. Due to intellectual property (IP) restrictions, only anonymized contigs without annotations can be made publicly available alongside the paper. However, we acknowledge the need for additional data and research to validate these findings and to obtain more robust results.

Section 3 describes our dataset, Section 2 reviews related approaches in the literature, and Section 4 uses both to answer the above questions. Section 5 concludes the paper.

2 Related Research

A large body of research has been dedicated to understanding genetic interactions and revealing genetic functions in different organisms in complex communities. We provide a brief, non-exhaustive overview of approaches related to this project.

Analyzing metatranscriptome data: Metatranscriptome data is obtained via RNA sequencing and captures gene expression profiles of organisms within a complex microbial community. It is typically analyzed by mapping to reference genomes [24,20] or assembly [12], which provide, among other benefits, a natural grouping, an on-gene distance metric, or insights into some specific functionalities [28,25]. However, we are only provided with contig counts but no contig meta-information, such as genetic annotations, making such a mapping infeasible for our dataset.

Finding genetic interactions: To investigate contig-contig interactions, we use concepts from gene interaction or co-expression networks in which nodes are typically defined as genes, and edges are the interaction strength between adjacent nodes. This interaction strength can be defined via correlation [22], assembly graph similarities [16], or structural similarities [7]. Given our dataset, correlation is the only option as it does not require auxiliary contig information, and we include it in our analysis. Cui *et al.* [6] detect genetic interactions by capturing them in a neural network using Shapley Taylor interaction indices. We include an adapted version in our analysis.

Investigating the rumen microbiome in livestock: Söllinger *et al.* [27] used quantitative metatranscriptomics with gas and volatile fatty acid profiling to investigate methanogen interactions and effects within the rumen of Holstein cows. Rather than observing natural differences between animals, the authors designed a targeted experiment allowing them to observe abundance fluctuations over time that can be linked to a specific feeding pattern. Their work follows a different path to identifying active methanogens and is not applicable to our dataset. Li *et al.* [18] investigate the breed effect on the rumen microbiome in beef cattle using metagenomics and metatranscriptomics. The authors compare observed abundancies with statistical analysis using t-tests and link these differences to feed efficiency. Our analysis extends beyond this approach using a machine learning perspective.

3 Dataset Description

The dataset used in this study consists of metatranscriptome data, which represents the collection of RNA sequences from the microbial community in the

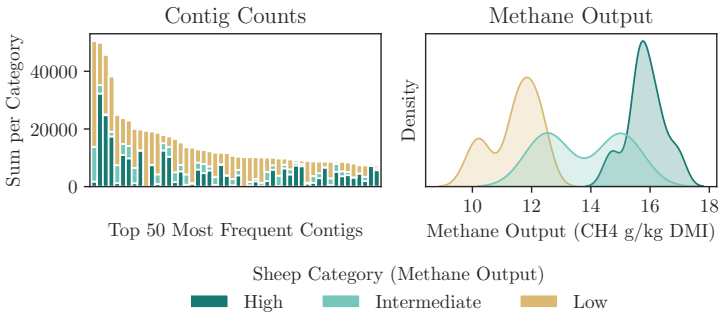

Fig. 1: Basic dataset statistics: contig counts per 1M contigs per sample, summed up per methane output category (left) and methane output distribution per sample (right)

sheep rumen. This data helps us understand which genes are active and what functions the microbes are performing in relation to methane production.

To obtain this data, 10 sheep with varying methane outputs were sampled on two distinct days. Specifically, RNA was extracted from the rumen of low (4), high (4), and intermediate (2) methane-yielding sheep, sampled on two dates with a 14-day gap in New Zealand. This RNA is then sequenced, producing "reads," short fragments of RNA sequences that serve as snapshots of the gene expression activity within the microbial community at the time of sampling.

The raw sequencing reads often contain errors or low-quality segments, which are first trimmed[4]. After cleaning, the reads are assembled into longer, contiguous sequences ("contigs")[5]. Contigs provide a clearer picture of which genes are being expressed and can then be used to explore how microbial activity in the rumen contributes to methane production, offering insights that could inform predictive models or strategies to reduce methane emissions in livestock. As Figure 1 (left) highlights, some contigs are found in the samples with high frequency, while others are rarely found. Overall, there are differences in the abundance of specific contigs for sheep with different methane output levels.

To assign equal weight to all samples in subsequent tasks, we normalize the raw contig counts per sample and express them as "counts per million". Contigs with counts per million less than one for all samples were subsequently removed, leaving $686,456$ contigs.

The contigs could further be annotated with corresponding genes, biological roles and functions, or (groups of) organisms from which the contig originated by

[4] Reads are trimmed with Trimmomatic version 0.39

[5] Trimmed reads are assembled into contigs using MEGAHIT version 1.2.9 with default parameters. The alignment of trimmed reads from each metatranscriptome sample to the MEGAHIT assembly was performed using the bwa aligner version 0.7.17-r1188. Aligned reads with a mapping quality of 30, indicating a 1 in 1000 chance of misalignment, were extracted using Samtools version 1.17.

Table 1: Overview of datasets used in this project after preprocessing

Table "Methane Output" Columns (20×5-dimensional)

SampleID	Unique identifier per sample that matches the SampleID in other tables
Sheep#	Unique identifier per sheep used for training/test splits
CH_4	Average daily methane emission (in g)
CH_4 / DMI	Average daily methane emission (in g) per Dry Matter Intake (in kg) (feed consumed per day on a moisture-free basis; in kg)
Methane Class	Categorization based on methane output (low/intermediate/high)

Table "Contigs" Columns ($686,456 \times 21$-dimensional)

ContigID	Unique identifier per contig
Counts SampleID_1	Contig counts for sample with ID_1
Counts ...	...
Counts SampleID_{20}	Contig counts for sample with ID_{20}

comparing the sequences to known databases. However, due to IP restrictions, we cannot disclose the annotated contigs but use unique identifiers instead. Using these ContigIDs, we are able to make this dataset, as well as the code for our analysis, publicly available alongside the paper in our repository[6].

In addition to the RNA samples, we measure the sheep's methane output, CH_4, by placing them in separate sealed chambers (respiration chambers) where their breath is monitored over a day. Multiple measurements (two to three days) per sheep were taken to mitigate measurement errors, and we averaged these results. Standard deviations were found to be very low, justifying the choice of averaging. Since the methane output is highly correlated with the sheep's food intake, we also monitor the Dry Matter Intake (DMI) of the sheep during their stay in the respiration chambers. Subsequently, we use the raw methane output in grams per kilogram DMI for our analysis, i.e., CH_4 g / kg DMI, and refer to it as methane output. This entire sampling procedure was repeated for the same sheep two weeks later to rule out anomalies, leading to a total of 20 samples (10 sheep $\times$ 2 measurement rounds). Figure 1 (right) illustrates the distribution of methane output per sample.

The two measurement rounds per sheep are generally considered separate training instances in this analysis due to the small dataset size. When splitting the data into training and test sets for evaluation, however, we make sure to randomly select sheep, not the individual measurements. After filtering the relevant columns, we obtain the datasets described in Table 1 that can be joined on a shared ID.

The number of contigs counted per sample varies, reflecting differences in the sizes of the rumen samples. This variation may introduce bias when analyzing the

[6] Our repository: https://github.com/KatDost/Sheep_Methane_Paper

data. To address this issue, we employ counts per million (CPM) normalization, which normalizes the counts per sample, mitigating the imbalance between samples. Furthermore, as is common in the field, the counts are then rounded to the nearest integer, suppressing measurement noise. This normalization technique ensures that each sample's contribution to the overall analysis is proportional to the contigs' share, not their absolute count, facilitating fair comparisons across samples.

4 Methods and Results

After preprocessing our dataset, we address the questions listed in the introduction, drawing inspiration and incorporating approaches from the related research projects discussed above.

4.1 Hypothesis and Data Validation

As a first step, we validate the existence of a connection between contig counts and methane output and the presence of meaningful signals within our dataset. We use various regression models to predict the methane output from the contig counts and evaluate their performance.

To identify suitable hyperparameters for each model while guarding against overfitting, we employ a rough hyperparameter grid search methodology using HalvingGridSearchCV [14,19]. Based on the grid search results, we decided to include the following regressors in the test: Linear Regression, Lasso Regression with $\alpha = 9339.46$, *Support Vector Regression (SVR)* with the nonlinear RBF kernel, $\gamma = 0.01$ and $C = 1000$, *Decision Tree (DT)* with different maximum depths (3 and 4), *Random Forest (RF)* with varying maximum depths (3 and 4) and 20 trees, and XGBoost [4] with 20 trees and learning rate = 0.1.

Employing 5-fold cross-validation, we assess the predictive capability of these models on our dataset based on *Mean Absolute Error (MAE)* for the sake of its interpretability. Mean Absolute Percentage Error (MAPE) showed similar patterns and is hence excluded. These metrics were computed for both training and test sets to gauge model performance and are presented in Figure 2 (left).

We observe that despite our efforts in hyperparameter tuning, all models overfit the training data. This overfitting can be attributed to the stark disparity between the small sample size and the vast number of features. The limited number of samples relative to the high dimensionality of the feature space poses a significant challenge for the models to generalize effectively.

The machine learning models generally demonstrate MAE values below the baseline model, always predicting the average methane output, with the exception of SVR. However, it is worth noting that there is a significant standard deviation among folds, indicating variability in model performance that can be traced back to the small test set sizes in each fold. Linear regression performs better than the tree-based methods, leading us to suspect an adverse effect due to many highly correlated features (see our repository for Pearson correlations).

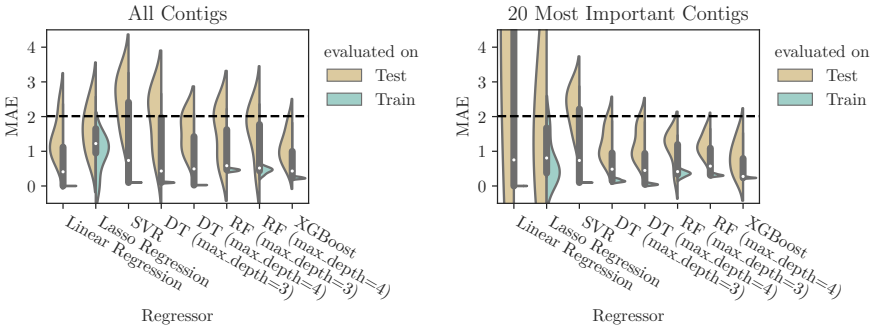

Fig. 2: MAE for multiple regressors predicting methane output trained on all features/contigs (left) and a selection of the 20 most important features based on a pre-trained RF (right) for training and test set individually. The dashed black line serves as a baseline (always predict the average methane output).

To disentangle the high correlation among contigs, we train a RF on each fold's training set and use its feature importance to select the most important features. This approach chooses a set of features (in our case, contigs) that is highly informative for the model, which mitigates high correlations by design as they would carry duplicate information. We observe a natural drop in feature importance after the 20 most important features for each fold and drop the rest before repeating the above experiment. Figure 2 (right) shows the results. While the feature selection harms the regression methods, the tree-based ones benefit largely, which may be attributed to the tree-based feature importance. We further observe that the feature selection decreases the test error substantially more than the training error, which confirms that the size of the input space is significantly contributing to the overfitting, in addition to the small dataset size. Although the models still overfit, they demonstrate a performance well below baseline, indicating that there is indeed a relationship between contig counts and methane output, validating our hypothesis.

In conclusion, while our analysis suggests the presence of a signal in the data, the small sample size, in contrast to the large number of contigs, imposes limitations on our machine learning approach to analyzing the dataset.

4.2 Identifying Essential Contigs

We can expect our dataset to contain a large number of contigs with auxiliary functions that do not play a role in methane production and are, therefore, irrelevant to this study. However, there may be non-methanogen contigs that do contribute to methane production by interacting with the methanogens in the rumen. These are the contigs we aim to identify as they provide new insights.

As evident from the previous section, predicting the raw methane output as a regression task is challenging. This can be attributed to the small sample size:

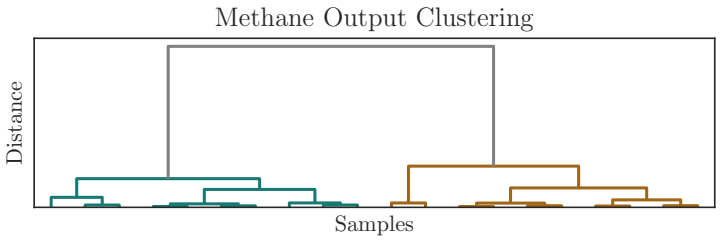

Fig. 3: Hierarchical clustering (Ward linkage) of samples based on their methane output

Variations in methane output can either be due to (i) measurement noise (sheep are not particularly compliant with our scientific endeavors) and the sample size is insufficient to obtain a fair estimate of the underlying distribution, or (ii) the variations are true signals, and the sample size is too small to capture these signals accurately. We choose to simplify the prediction task by converting it to a binary problem to alleviate the impact of the above issues.

To obtain binary labels, we cluster the samples hierarchically based on their methane output and observe two clearly defined groups as illustrated in Figure 3: high and low methane output samples. Note that these groups do not match the "Methane Class" categorization the dataset was originally annotated with (see Table 1 – "Methane Output"). Our hierarchical clustering reveals that there is no natural third group with "medium" methane output.

Upon training an initial decision tree classifier, we uncover decision stumps that can perfectly distinguish between low and high methane output. One example is shown in Figure 4 (left), where a single contig suffices to discern between the two output categories, a surprising discovery. We adopt an iterative approach to tally the number of contigs with this distinguishing property, sequentially removing the contig used for the stump and retraining a new stump. This method identifies 348 contigs capable of perfectly differentiating between low and high methane output.

Motivated by these findings, we scrutinize whether these contigs represent statistically significant discoveries or mere chance occurrences. To this end, we conduct pairwise t-tests for each contig, comparing the corresponding contig counts between samples with high and low methane output. Figure 4 (right) showcases the number of contigs exhibiting significantly different values for low and high methane output samples under specified p-values. We denote these contigs as *"supercontigs"* for brevity. Subsequently, we typically limit our analysis to supercontigs.

In conclusion, we have identified a set of contigs that play a substantial role in the sheeps' methane production. These findings are statistically significant under specified significance levels. Interestingly, our set of supercontigs contains methanogens as well as non-methanogens.

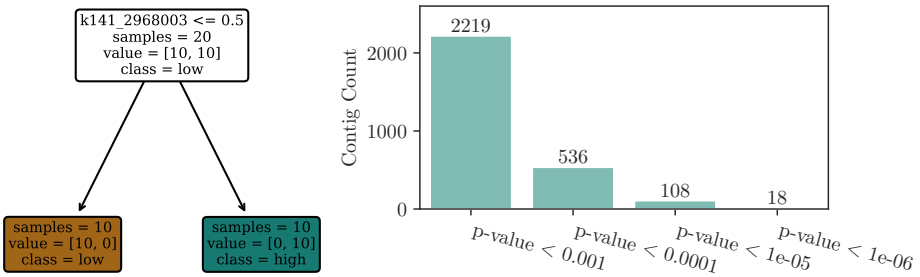

Fig. 4: **Left:** Decision Tree to predict high (10) and low (10) methane output using all contigs. There are 348 contigs, such as k141_2968003, that can distinguish **perfectly** between both classes. **Right:** Zoom-in on p-values for pairwise t-tests on contig counts for the cohorts high/low methane output: Displayed is the number of contigs for which the p-value lies below a specific threshold.

4.3 Understanding Patterns

In the previous sections, we narrowed down the set of contigs that are involved in the methanogen cycle in sheep, but we have also observed high correlations among contigs. Naturally, we seek to investigate which of the relevant contigs act together, and which ones drive different mechanisms. We explore three different approaches to find groups of contig interactions, i.e., community search in a correlation network, non-negative matrix factorization, and neural networks with Shapley Taylor interaction index [6,9] values. These approaches are not to be seen as competing but as different perspectives on the same question. The groups identified by different approaches will likely be different but can all be of interest to a domain expert and collectively help the understanding of rumen methanogen mechanisms.

Community Search in a Contig Network We construct a contig-contig interaction network as follows: Each contig is a node. Each pair of contigs is connected by an edge indicating the p-value of a pairwise t-test between the counts of the contigs represented by the adjacent nodes. Using the Louvain method [2], we detect communities of highly interacting contigs within the network. Figure 5 shows an example. Although the displayed interactions are statistically significant, we observe an unwelcome chain effect: If $C_1 \leftrightarrow C_2$ and $C_2 \leftrightarrow C_3$ are significant interactions, we frequently observe a community containing C_1, C_2, C_3, although $C_1 \not\leftrightarrow C_3$ is *not* necessarily a significant interaction.

Matrix Factorization We employ *non-negative matrix factorization (NMF)* [17] to uncover latent structures and patterns within high-dimensional data, facilitating the identification of groups of similarly acting contigs and aiding in the interpretation of complex relationships between contig counts and methane output levels.

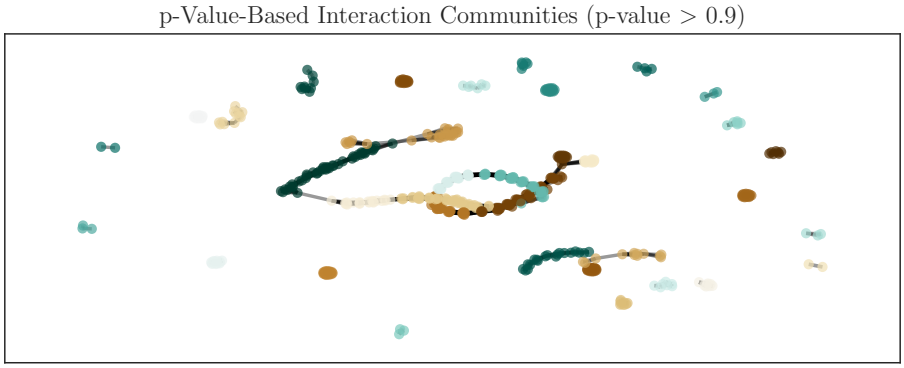

Fig. 5: Louvain-communities (colors) in a pairwise t-test-based contig interaction network. Edges with p-values < 0.9 as well as isolated nodes have been removed. Annotations are omitted to enhance readability.

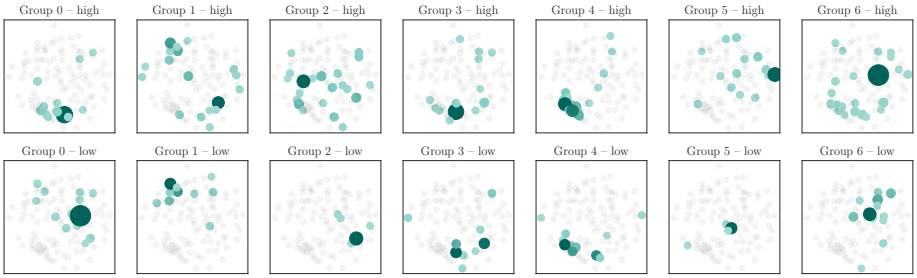

Fig. 6: NMF has been applied separately for samples with high and low methane output. Each point corresponds to a contig in MDS space. We included only supercontigs with $p < 10^{-4}$. Group members are colored; darker colors correspond to stronger group membership.

Given a matrix X, matrix factorization aims to find two matrices of lower dimensionality whose product approximates X as closely as possible. We use NMF to find mechanisms of contig behavior, i.e., groups of contigs that operate together and exhibit similar patterns by factorizing the Sample x Contig Count matrix into A (of dimension #samples$\times l$) and B (of dimension $l\times$ #contigs). The number of groups, l, is often referred to as the latent dimension and is a parameter that needs to be tuned.

Since both factors typically have a smaller dimension, the input matrix usually cannot be reconstructed perfectly, and the factors necessarily have to focus on the most important information and neglect minor variations, suppressing noise. Following the definition of matrix multiplication, the factors group rows with similar patterns since they trigger the same columns in the corresponding factor and vice versa. These groups can overlap (which distinguishes matrix fac-

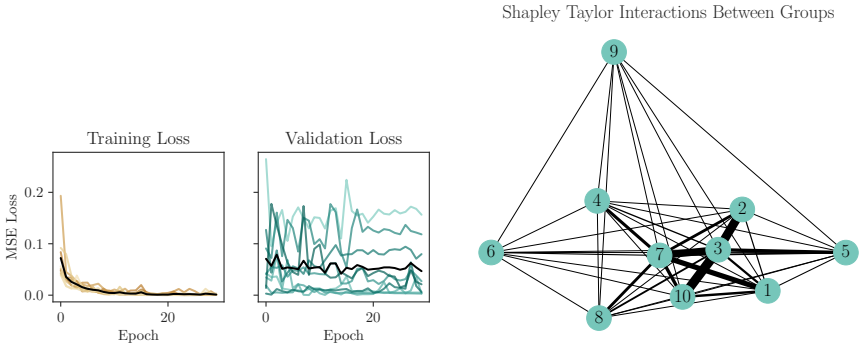

Fig. 7: Neural Network using contig groups. **Left:** Training and validation loss show poor learning performance – **the results of this neural network cannot be trusted! Right:** Nodes are groups of contigs. Edges indicate interactions between connected groups. Edge thickness displays the strength of the interaction.

torization from standard clustering and our network community search). NMF offers nonnegative numbers expressing the strength of group memberships.

We restrict our analysis to contigs that exhibit significant differences for high and low methane outputs at a significance level of 1e-05, acknowledging the adjustability of this parameter for future experiments. Each sample is rescaled independently such that its L2 norm equals one. This normalization step is crucial to ensure each sample is considered equally.

Next, we tune the latent dimension l by maximizing the cophenetic correlation coefficient – a measure of how faithfully a dendrogram preserves pairwise distances in the original data—and select $l = 7$. We construct separate sample×contig count matrices for high and low-methane-outputting sheep and perform matrix factorization on both.

To visually represent the identified contig groups, we train a 2-dimensional embedding space using *multi-dimensional scaling (MDS)* [8,3]. An embedding space is a low-dimensional representation of high-dimensional data that preserves its inherent structure and relationships. We use MDS to position the contigs in a plot using their pairwise correlation as a similarity measure ($1 - |\text{correlation}|$ as a distance). That means close contigs are highly correlated (positively or negatively), whereas distant contigs have a low correlation.

The identified contig groups are illustrated in Figure 6, where distinct interactions are observed for low and high methane-outputting sheep. We can observe changes in group memberships in high and low methane producing sheep and that these identified groups often stretch beyond clusters of highly correlated contigs, which has the potential to reveal interesting insights. Domain experts investigated annotated versions of these plots (that we have to omit due to IP restrictions) and confirmed this.

Neural Networks with Shapley Taylor Interaction Index To find different contig groups and uncover interactions between these groups, we adapt the methodology proposed by Cui *et al.* [6]. The authors proposed a framework for detecting genetic interactions by considering all single nucleotide polymorphisms (SNPs) within selected genes and their complex relationships. They developed a deep learning architecture that captures these interactions using Shapley scores between hidden nodes representing genes. Their approach successfully identified significant interactions in real-world datasets and hence offers a promising avenue for us. However, since the authors use an existing mapping from SNPs to genes, the approach cannot directly be transferred to this study.

Instead, we allow the model to learn the contig grouping instead of initializing it with prior knowledge, as we lack predefined mappings. Our model architecture consists of a sparse layer, a linear layer with softplus activation, and a linear output layer. The sparse layer creates a bottleneck to group contigs, akin to mapping single nucleotide polymorphisms (SNPs) to genes in the original framework.

We present the training and validation loss per epoch in Figure 7 (left). We observe that the network does not train properly, as can be seen from the validation loss. We attribute this to the small sample size but include the results regardless since the approach is promising and can be reused in larger datasets.

We emphasize that results derived from this trained model cannot be considered reliable! However, for the sake of demonstration, we search for interactions between contig groups post-training using the *Shapley Taylor Interaction Index* [9]. Attribution or feature importance for neural networks generally measures the contribution of individual features to a prediction. The Shapley value measures the change in model prediction when a specific feature is included or omitted. The Shapley Taylor index identifies to what extent a set of features exert influence in conjunction as opposed to independently. Thereby, we obtain strong interactions between pairs (or higher-order groups) of features. Figure 7 (right) shows the identified interactions between groups. We have to omit the annotated version of the plot due to IP limitations.

We conclude that this network is too complex given our small sample size, and decide to drop the sparse layer. This way, we will find contig-contig interactions instead of group interactions. Although not a grouping per se, these interactions also provide us with information on contigs acting together. This network is training well – Training and validation loss are converging conjointly on all leave-one-sheep-out runs (see our repository for details). We can, therefore, expect more reliable results than previously.

As before, we carry out a post-training interaction search using the Shapley Taylor interaction index. The identified interaction network is presented in Figure 8. We suggest filtering interactions involving specific contigs of interest when analyzing this result. We observe a densely interacting group of contigs. The surrounding ones, however, interact with only a few contigs.

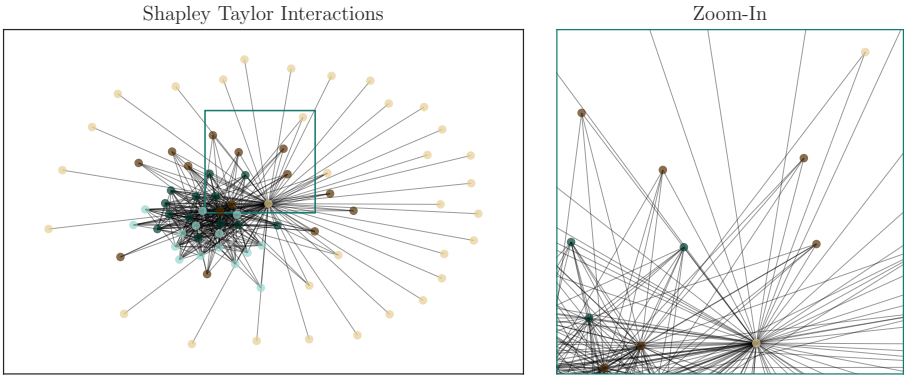

Fig. 8: Interactions between contigs identified using the Shapley Taylor interaction index with interaction strength > 0.5. The right image shows a zoom-in on a part of the left figure to reveal more detailed interactions. We included only supercontigs with $p < 10^{-4}$.

4.4 Identifying Causal Relationships in the Rumen

Lastly, we seek to find stronger relationships than interactions – we are searching for causal relationships between contigs as well as between contigs and the methane output. Contigs causing an increase or decrease in methane production can be the key to designing injections that mitigate methane production in ruminants.

Initially, we normalize the data such that each sample's contig counts add up to 1. We then establish a graph skeleton using the *Graphical LASSO (GLASSO)* [10] technique. This skeleton serves as a scaffold for subsequent causal inference and reduces the computational burden.

To find causal relationships within the constructed skeleton, we employ a number of algorithms from the Causal Discovery Toolbox [15] (see the package documentation for details on the methods and the original references): Greedy Equivalence Search, *Peter-Clark (PC) Algorithm* [5], Greedy Interventional Equivalence Search, Linear Non-Gaussian Acyclic model, and Structural Agnostic Model. None of these methods identified any causal relationships within our dataset, except for the PC algorithm. PC is a score-based approach for causal discovery based on conditional tests on variables and sets of variables that is quite popular [11] and allows us to set the significance level α for the individual conditional independence tests and make up for the small sample size.

As depicted in Figure 9, PC identified a number of causal relationships between contigs (left), but also between contigs and the methane output (right), revealing potential drivers of methane production in the rumen microbiome.

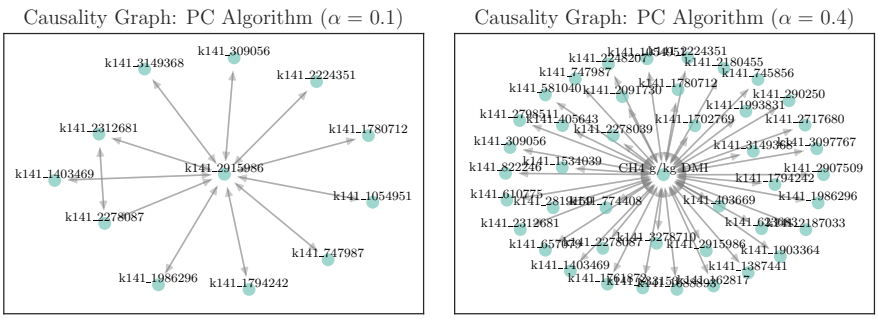

Fig. 9: Causal relationships between contigs (left) as well as between contigs and the methane output (right) using the Peter-Clark Algorithm. We included only supercontigs with $p < 10^{-4}$.

4.5 Validation

Our project is of exploratory nature, aiming to provide new insights to domain experts. Many of our results are backed up by what is already common knowledge in the fields. In addition, we revealed a number of previously unknown mechanisms that will be subject to further experimental and theoretical investigation by domain experts on the path to developing new chemical inhibitors.

5 Conclusion

As a proof-of-concept on the way to developing new chemical inhibitors for livestock to reduce their methane emissions, in this project, we aimed to explore and understand the genetic basis of the complex interplay between rumen microbes and methane production in livestock using machine learning.

Our findings are based on a small rumen metatranscriptome dataset gathered in-house from 10 sheep on two sampling days by domain experts, yielding 20 samples. Although small, the sample size is sufficient to indicate whether there is a signal in the data or not. We approached this project with a number of different statistical and machine learning techniques, identifying potential molecular drivers of methane production, with several contigs emerging as strong candidates for further investigation. Our domain experts confirmed that our results are reasonable and reveal new and interesting mechanisms.

Looking ahead, to augment our understanding of rumen microbial interactions and methane production, a larger-scale subsequent experiment could benefit from additional samples (particularly from a larger number of individual animals, but also from other ruminant species such as cattle and deer). Increasing the dataset size will enhance the reliability of machine learning models and support the training process, making it possible to train more complex models than were used in this study.

Integrating metadata on the contigs, such as their molecular structure or existing relationships, could help define a graph and open the field for graph neural networks, even on relatively small datasets.

In conclusion, this project is a step towards utilizing machine learning approaches to understand the complex interactions affecting methane production in ruminant livestock.

Acknowledgments. This study was funded by the Ministry of Business, Innovation & Employment, New Zealand (MBIE number C10X2201 awarded to SKG). KD and JW have received research funding from AgResearch, New Zealand. Although their identities are unknown to us, we would like to thank the reviewers of this paper. We found great value in the provided comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data and Software Availability Data, software, and experimental scripts able to reproduce all results presented in this article are available in our repository at https://github.com/KatDost/Sheep_Methane_Paper.

References

1. Baca-González, V., Asensio-Calavia, P., González-Acosta, S., Pérez de la Lastra, J.M., Morales de la Nuez, A.: Are vaccines the solution for methane emissions from ruminants? a systematic review. Vaccines **8**(3) (2020). https://doi.org/10.3390/vaccines8030460
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (oct 2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
3. Cannistraci, C.V., Alanis-Lobato, G., Ravasi, T.: Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. Bioinformatics **29**(13), i199–i209 (06 2013). https://doi.org/10.1093/bioinformatics/btt208
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
5. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. Journal of Machine Learning Research **15**(116), 3921–3962 (2014)
6. Cui, T., El Mekkaoui, K., Reinvall, J., Havulinna, A.S., Marttinen, P., Kaski, S.: Gene–gene interaction detection with deep learning. Communications Biology **5**(1), 1238 (2022)
7. Dai, Y., Guo, C., Guo, W., Eickhoff, C.: Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. Briefings in Bioinformatics **22**(4), bbaa256 (10 2020). https://doi.org/10.1093/bib/bbaa256

8. Davison, M.L., Sireci, S.G.: 12 - multidimensional scaling. In: Handbook of Applied Multivariate Statistics and Mathematical Modeling, pp. 323–352. Academic Press, San Diego (2000). https://doi.org/10.1016/B978-012691360-6/50013-6
9. Dhamdhere, K., Agarwal, A., Sundararajan, M.: The shapley taylor interaction index. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR (2020)
10. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (12 2007). https://doi.org/10.1093/biostatistics/kxm045
11. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10** (2019). https://doi.org/10.3389/fgene.2019.00524
12. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.: Full-length transcriptome assembly from rna-seq data without a reference genome. Nature biotechnology **29**(7), 644–652 (2011)
13. Hickey, S.M., Bain, W.E., Bilton, T.P., Greer, G.J., Elmes, S., Bryson, B., Pinares-Patiño, C.S., Wing, J., Jonker, A., Young, E.A., Knowler, K., Pickering, N.K., Dodds, K.G., Janssen, P.H., McEwan, J.C., Rowe, S.J.: Impact of breeding for reduced methane emissions in new zealand sheep on maternal and health traits. Frontiers in Genetics **13** (2022). https://doi.org/10.3389/fgene.2022.910413
14. Jamieson, K., Talwalkar, A.: Non-stochastic best arm identification and hyperparameter optimization. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 51, pp. 240–248. PMLR, Cadiz, Spain (5 2016)
15. Kalainathan, D., Goudet, O., Dutta, R.: Causal discovery toolbox: Uncovering causal relationships in python. Journal of Machine Learning Research **21**(37), 1–5 (2020)
16. Lamurias, A., Sereika, M., Albertsen, M., Hose, K., Nielsen, T.D.: Metagenomic binning with assembly graph embeddings. Bioinformatics **38**(19), 4481–4487 (08 2022). https://doi.org/10.1093/bioinformatics/btac557
17. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. nature **401**(6755), 788–791 (1999)
18. Li, F., Hitch, T.C.A., Chen, Y., Creevey, C.J., Guan, L.L.: Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. Microbiome **7**(6) (2019). https://doi.org/10.1186/s40168-019-0618-5
19. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research **18**(185), 1–52 (2018)
20. Mann, E., Wetzels, S.U., Wagner, M., Zebeli, Q., Schmitz-Esser, S.: Metatranscriptome sequencing reveals insights into the gene expression and functional potential of rumen wall bacteria. Frontiers in Microbiology **9** (2018). https://doi.org/10.3389/fmicb.2018.00043
21. Palangi, V., Lackner, M.: Management of enteric methane emissions in ruminants using feed additives: A review. Animals **12**(24) (2022). https://doi.org/10.3390/ani12243452
22. Park, C., Kim, J., Kim, J., Park, S.: Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. PLOS ONE **13**(7), 1–15 (07 2018). https://doi.org/10.1371/journal.pone.0201056

23. Reisinger, A., Clark, H.: How much do direct livestock emissions actually contribute to global warming? Global Change Biology **24**(4), 1749–1761 (2018). https://doi.org/10.1111/gcb.13975
24. Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.J., Cuenca, M., Field, C.M., Coelho, L.P., Cruaud, C., Engelen, S., et al.: Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell **179**(5), 1068–1083 (2019)
25. Shakya, M., Lo, C.C., Chain, P.S.G.: Advances and challenges in metatranscriptomic analysis. Frontiers in Genetics **10** (2019). https://doi.org/10.3389/fgene.2019.00904
26. Shi, W., Moon, C.D., Leahy, S.C., Kang, D., Froula, J., Kittelmann, S., Fan, C., Deutsch, S., Gagic, D., Seedorf, H., Kelly, W.J., Atua, R., Sang, C., Soni, P., Li, D., Pinares-Patiño, C.S., McEwan, J.C., Janssen, P.H., Chen, F., Visel, A., Wang, Z., Attwood, G.T., Rubin, E.M.: Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Research **24** (2014). https://doi.org/10.1101/gr.168245.113
27. Söllinger, A., Tveit, A.T., Poulsen, M., Noel, S.J., Bengtsson, M., Bernhardt, J., Hellwing, A.L.F., Lund, P., Riedel, K., Schleper, C., Højberg, O., Urich, T.: Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. mSystems **3**(4) (2018). https://doi.org/10.1128/msystems.00038-18
28. Zhang, Y., Thompson, K.N., Branck, T., Yan, Y., Nguyen, L.H., Franzosa, E.A., Huttenhower, C.: Metatranscriptomics for the human microbiome and microbial community functional profiling. Annual Review of Biomedical Data Science **4**(Volume 4, 2021), 279–311 (2021). https://doi.org/10.1146/annurev-biodatasci-031121-103035