

Lipid discovery enabled by sequence statistics and machine learning

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1

February 22, 2024 (this version)

Sent for peer review

December 20, 2023

Posted to preprint server

October 17, 2023

Priya M. Christensen, Jonathan Martin, Aparna Uppuluri, Luke R. Joyce, Yahan Wei, Ziqiang Guan , Faruck Morcos , Kelli L. Palmer 

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA • Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA • School of Podiatric Medicine, University of Texas Rio Grande Valley, Harlingen, TX 78550, USA • Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA • Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, USA • Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080, USA

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Bacterial membranes are complex and dynamic, arising from an array of evolutionary pressures. One enzyme that alters membrane compositions through covalent lipid modification is MprF. We recently identified that *Streptococcus agalactiae* MprF synthesizes lysyl-phosphatidylglycerol (Lys-PG) from anionic PG, and a novel cationic lipid, lysyl-glucosyl-diacylglycerol (Lys-Glc-DAG), from neutral glycolipid Glc-DAG. This unexpected result prompted us to investigate whether Lys-Glc-DAG occurs in other MprF-containing bacteria, and whether other novel MprF products exist. Here, we studied protein sequence features determining MprF substrate specificity. First, pairwise analyses identified several streptococcal MprFs synthesizing Lys-Glc-DAG. Second, a restricted Boltzmann machine-guided approach led us to discover an entirely new substrate for MprF in *Enterococcus*, diglucosyl-diacylglycerol (Glc2-DAG), and an expanded set of organisms that modify glycolipid substrates using MprF. Overall, we combined the wealth of available sequence data with machine learning to model evolutionary constraints on MprF sequences across the bacterial domain, thereby identifying a novel cationic lipid.

eLife assessment

This study reports **important** findings on identifying sequence motifs that predict substrate specificity in a class of lipid synthesis enzymes. It sheds light on a mechanism used by bacteria to modify the lipids in their membrane to develop antibiotic resistance. The evidence is **convincing**, with a careful application of machine learning methods, validated by mass spectrometry-based lipid analysis experiments. This interdisciplinary study will be of interest to computational biologists and to the community working on lipids and on enzymes involved in lipid synthesis or modification.

1 Introduction

Lipids are ubiquitous and diverse group of hydrophobic and amphiphathic compounds that are critical for fundamental biological processes, including the formation of cell membranes, protection, and cargo delivery; energy storage; and cell signaling pathways [1]. They are relatively smaller than other complex biomolecules, such as proteins, thereby allowing a larger portion of their surface to interact with other macromolecules. Chemical modifications to lipids, including changes in fatty acid tails (saturated and unsaturated) [2], modifications of head groups [3], and alterations to lipid charge [4], impact these interactions, as well as the physicochemical properties of membranes. Commonly, bacteria synthesize lipids that are negatively charged or neutral [5]. Some bacteria modify the negatively charged lipids with amino acids to make them positively charged to confer resistance to cationic antimicrobial peptides and cationic antibiotics [6, 7]. The bacterial domain is a rich source of natural lipids that are likely co-evolved for specific interactions with host tissues and other cells, yet the full extent of this chemical diversity remains unexplored. Finding bacteria that possess unique lipids can be useful for biotechnology. For example, natural lipids found in bacteria can be used for drug delivery. A study by [8] used low cytotoxic outer membrane vesicles (OMV) that contained a modified lipopolysaccharide (LPS) to deliver siRNA to targeted cancer cells. This delivery mechanism had success and demonstrates the possibility of using naturally occurring bacterial lipids for future drug delivery processes to eukaryotic cells.

One enzyme responsible for modifying lipids in bacteria is the multiple peptide resistance factor (MprF). MprF modifies lipids through the transfer of amino acids from charged tRNAs to the head group of the anionic membrane phospholipid, phosphatidylglycerol (PG) [4]. The MprF protein is comprised of two domains: an N-terminal flippase domain responsible for moving aminoacylated lipids from the inner membrane leaflet to the outer leaflet and a C-terminal synthase domain responsible for the aminoacyl transfer from tRNA to lipids [9, 10]. Recently, in *Streptococcus agalactiae* (Group B *Streptococcus*, GBS) it was found that MprF can modify two different substrates with lysine (Lys) – the glycolipid glucosyl-diacylglycerol (Glc-DAG) and the phospholipid, PG [11], generating Lys-Glc-DAG, as well as Lys-PG [11, 12] (**Figure 1**). An *S. agalactiae* mprF deletion mutant no longer synthesizes Lys-Glc-DAG or Lys-PG and expression of *S. agalactiae* mprF in the heterologous host *Streptococcus mitis* conferred Lys-Glc-DAG and Lys-PG synthesis to *S. mitis* [11]. This was the first time MprF was demonstrated to add lysine onto a neutral glycolipid (Glc-DAG).

The synthesis of Lys-Glc-DAG by *S. agalactiae* MprF illustrated a unique lipid substrate and product by the enzyme that had not been characterized before, highlighting the possibility for further exploration into unknown lipids and lipid substrates by MprF of different bacterial species. We sought to investigate the molecular determinants of enzyme specificity and identify other bacterial species that may have this novel specificity. This investigation led us from the standard methodology of BLASTp queries [13], which are based on single residue amino acid substitution frequencies, to a more modern statistical method called a Restricted Boltzmann Machine (RBM) [14], which captures global statistical patterns within natural sequence data. In this work, we develop a general strategy for combining this high-powered statistical model with lipidomic analysis to discover MprF variants with unique enzymatic activity. This process led us to the discovery of an uncharacterized lipid substrate for MprF, diglucosyl-diacylglycerol (Glc₂-DAG), which is present in several strains identified by the parameters of the statistical model. These models also provide interpretable and actionable information for use in further enzyme characterization, and we show that this application is a useful companion tool for experimental work.

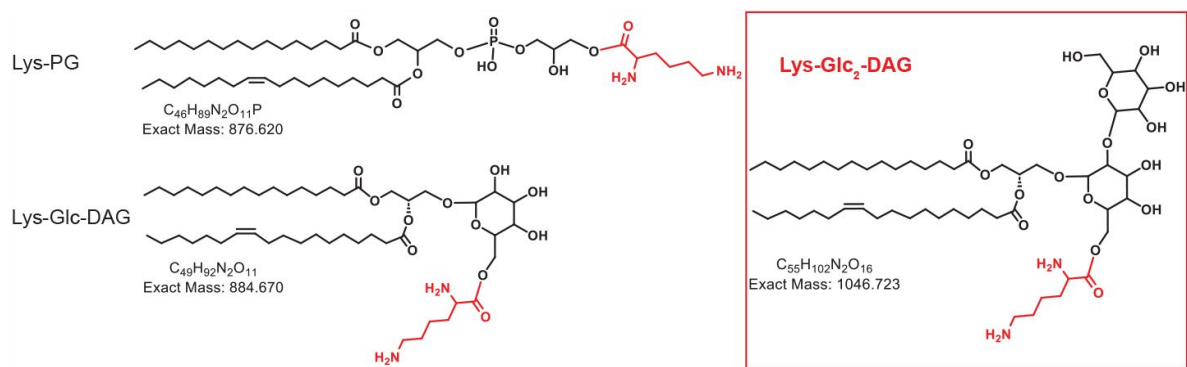


Figure 1.

Chemical structures of lysine lipids. Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc-DAG, lysylglucosyl-diacylglycerol; Lys-Glc₂-DAG, lysyl-diglucosyl-diacylglycerol.

2 Results

2.1 Simple pairwise sequence statistics identify streptococcal MprF enzymes that synthesize Lys-Glc-DAG

Utilizing a method based on the amino acid sequence of *S. agalactiae* MprF, we identified four streptococcal MprFs with high sequence identity to *S. agalactiae* COH1 MprF (WP 000733236.1). A BLASTp analysis found that *Streptococcus sobrinus* ATCC 27352 MprF (WP 019790557.1) had 61.4% amino acid identity to the *S. agalactiae* MprF, *Streptococcus salivarius* ATCC 7073 MprF (WP 002888893.1) had 43.09% amino acid identity, *Streptococcus ferus* MprF (WP 018030543.1) had 61.88% amino acid identity, and *Streptococcus downei* (WP 002997695.1) had 61.03% amino acid identity (**Table 1**). Plasmids were designed to express the *S. sobrinus* *mprF* (pSobrinus), *S. salivarius* *mprF* (pSalivarius), *S. ferus* *mprF* (pFerus), and *S. downei* *mprF* (pDownei) genes. The plasmids were transformed into the expression host *S. mitis* NCTC12261 (SM61). The empty vector (pABG5Δ *phoZ*, referred to as pABG5 here) was also transformed (**Figure 2a**). *S. mitis* does not natively encode *mprF* but synthesizes lipids (Glc-DAG; PG) that are substrates for *S. agalactiae* MprF [15, 16, 11], making it an appropriate heterologous host for expressing streptococcal MprFs. Previously, a plasmid expressing *S. agalactiae* *mprF* (pGBSMprF) [11] was generated and transformed into *S. mitis*. Lys-Glc-DAG and Lys-PG presence in SM61 was only possible with the expression of pGBSMprF (Figure2b) [11].

S. mitis lipids of strains expressing the plasmids pSobrinus (**Table 1**), pSalivarius (**Figure 2c**), pFerus (**Figure 2d**) and pDownei (**Table 1**) were analyzed. Lipidomic analysis found that pSobrinus, pDownei, and pFerus conferred synthesis of Lys-Glc-DAG, but not Lys-PG in *S. mitis*. Notably, pFerus confers synthesis of Lys-Glc-DAG at a similar level to *S. agalactiae* MprF (**Figures 2b, d**). In contrast, pSalivarius conferred synthesis of Lys-PG only and not Lys-Glc-DAG. Importantly, *S. salivarius* MprF also synthesized Lys-PG at a level most similar to *S. agalactiae* MprF (**Figures 2b, c**). Lysine-modified lipids synthesized by *S. mitis* strains expressing streptococcal MprFs used in this study are summarized in **Table 1**. In general, streptococcal MprF enzymes with higher amino acid percent (> 60%) to *S. agalactiae* MprF conferred Lys-Glc-DAG synthesis to *S. mitis*, but only *S. agalactiae* MprF conferred both Lys-Glc-DAG and Lys-PG synthesis in *S. mitis*.

2.2 Restricted Boltzmann Machines can provide sensitive, rational guidance for sequence classification

We set out to understand which specific MprF amino acid residues are involved in enzyme specificity for either the PG or Glc-DAG substrate. To this end, we employed a Restricted Boltzmann Machine (RBM), a probabilistic graphical model which aims to model the probability of data in a data set through statistical connections between the features of the data and a set of hidden units (weights) [17]. This class of models is closely related to another model, the Boltzmann Machine, which has been very successful in elucidating important residues in protein structure and function [18], one example being the careful tuning of enzyme specificity in bacterial two-component regulatory systems [19, 20]. The Boltzmann Machine formulation from [18], termed Direct-Coupling Analysis (DCA), stores patterns in the form of a pairwise coupling matrix and local field matrix (shown in Equation 2), which has been useful in understanding how likely pairs of residues are to interact in physical space. We hypothesize that *S. agalactiae* MprF and other MprF enzymes that utilize glycolipid substrates have unique patterns of interacting residues relative to those enzymes that use only PG as a substrate.

While modeling how pairs of residues covary is powerful, many important features of enzyme function like allostery or enzyme specificity (which is the purpose of this analysis) may involve more than two residues [21], and in these situations it can be difficult to connect sets of residues using solely pairwise parameters [22]. Therefore, a more global approach is required to address

Bacteria Strains	% Amino Acid Identity (Similarity)	Lys-PG	Lys-Glc-DAG
SM61(pGBSMprF)	100.0 (100.0)	✓	✓
SM61(pFerus)	61.9 (79.0)	X	✓
SM61(pSobrinus)	61.4 (79.0)	X	✓
SM61(pDownei)	61.0 (79.0)	X	✓
SM61(pSalivarius)	43.1 (66.0)	✓	X
SM61(pABG5)	- (-)	X	X

Table 1

Summary of different MprF variants expressed in *S. mitis* and the lysine lipids they produce.

Percentage of amino acid identity and similarity compared to *S. agalactiae* COH1 MprF; data obtained from BLASTp. The lipids each strain synthesize are denoted by a checkmark or an x. Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc-DAG, lysyl-glucosyl-diacylglycerol.

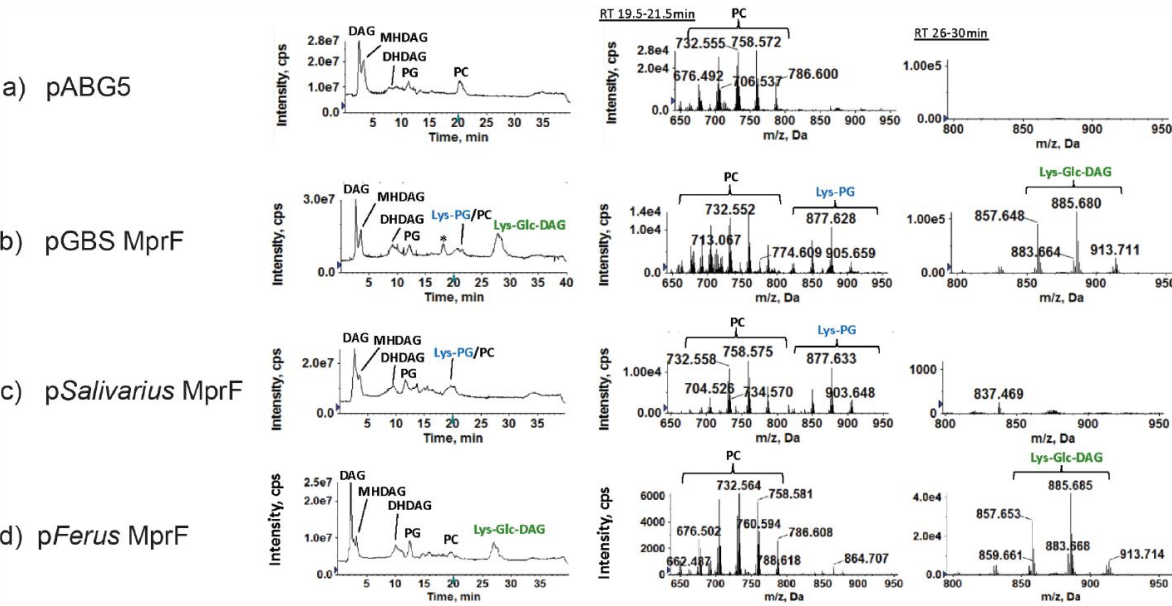


Figure 2.

Synthesis of lysine lipids (Lys-PG and Lys-Glc-DAG) in *S. mitis* expressing *mprFs* from *S. agalactiae*, *S. salivarius*, and *S. ferus*. (a) *S. mitis* NCTC12261 with empty vector control (pABG5) lacks lysine lipids; (b) *S. agalactiae* *mprF* (pGBSMprF) produces both Lys-PG and Lys-Glc-DAG; (c) *S. salivarius* *mprF* produces only Lys-PG; (d) *S. ferus* *mprF* produces only Lys-Glc-DAG. Left panels: total ion chromatograms (TIC); middle panels: mass spectra of retention time 19.5–21.5 min showing Lys-PG and PC; right panels: mass spectra of retention time 26–30 min showing Lys-Glc-DAG. Note: “*” is an extraction artifact due to chloroform used. DAG, diacylglycerol; MHDAG, monohexosyldiacylglycerol; DHDAG, dihexosyldiacylglycerol; PG, phosphatidylglycerol; Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc-DAG, lysyl-glucosyl-diacylglycerol; PC, phosphatidylcholine.

our hypothesis. To this end, a RBM has recently been developed for use in the learning of protein sequence data, with special attention given to producing a model which produces sparse, interpretable, and biologically meaningful representations of these higher-order statistical couplings within the hidden units [14]. The design of this RBM (see the marginal distribution form shown in Equation 3) allows the following useful properties: higher-order couplings between residues (sets of coupled residues instead of pairs); bimodal hidden unit outputs given the training data set (Equation 4); sparse, interpretable configurations of the weights ($w_{i\mu}$) (Equation 5) and a compositional representation of weights [23], where input sequences are modeled largely through combinations of hidden units instead of single, highly activating units.

The compositional representation is a critical feature of this method, allowing us to find independently coupled sets of residues which in combination describe these protein sequences, and it is achieved through the particular choices of training parameters. More details on the training parameter choices are found in *Methods and Materials*. In particular, we focus on the scoring of sequences using individual weights $w_{i\mu}$:

$$I_{\mu}(\mathbf{v}) = \sum_i \mathbf{w}_{i\mu}(\mathbf{v}_i) \quad (1)$$

where $\mathbf{v}$ is an input sequence vectorized through one-hot encoding, to produce an output score (a single number) which depends on the input sequence residues at position i and the value of the weight matrix for that residue at that position.

We used this methodology to study the sequences in the family of MprF, which is a large protein composed of independent flippase and synthase domains [9, 24]. In our analysis, we restrict ourselves to the cytosolic region (see yellow box in Figure 3a), also known as the aminoacyl-PG synthase domain, which constituted the Pfam family DUF2156 (now renamed LPG synthase C) [25]. One weight in particular is highlighted, where we find bimodal activation (Figure 3b) of the hidden unit (Figure 3c) when Equation 1 is evaluated with the training sequences. Interpreting the sequence histograms, a positive value output in Figure 3b corresponds to a sequence containing combinations of the residues listed in the positive portion of the weight diagram in Figure 3c, whereas a negative output corresponds to residues in the negative portion. For display purposes, the residues shown in all of our weights correspond to the positions where the magnitude of the visible positions $w_{i\mu}$ are greater than some threshold proportional to the largest magnitude position (position 212); the weight has values for all possible amino acids in the entire protein length, but they are considerably smaller than this threshold and generally very low.

Two positions in particular, 152 and 212, in the positive portion of the weight correspond to residues 684 and 742, which were experimentally demonstrated to be relevant for aminoacylated-PG production by the *Bacillus licheniformis* and *Pseudomonas aeruginosa* cytosolic synthase domain variants [26]; in *B. licheniformis* the residues which form a complex with the lysine from Lys-tRNA^{Lys} are shown in Figure 3c. We have highlighted these two aforementioned variants as well as the *S. agalactiae* variant for demonstration of sequence scoring through Equation 1 in Figure 3a. We used this weight to create a filtered data set; only sequences that contained this necessary motif were of interest to us, and we subsequently trained a new model using only sequences with a catalytic weight output greater than 2. We outline this general workflow in Figure 4.

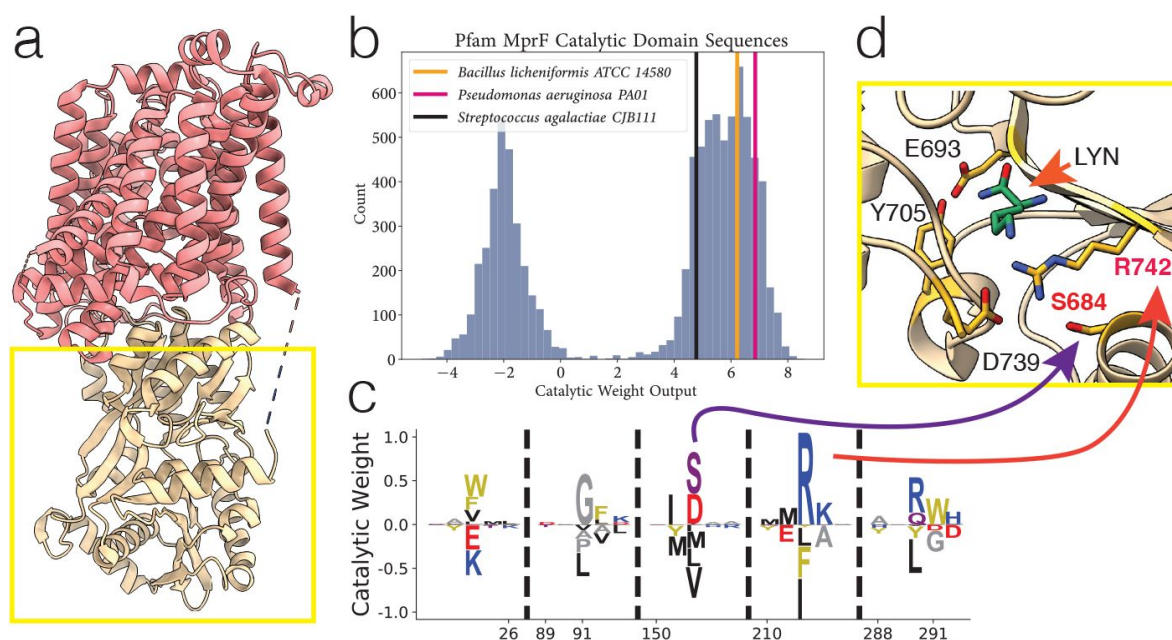


Figure 3.

Example of hidden unit analysis and usage. (a) The structure of PDB:7DUW, with the red colored region being the transmembrane flippase domain and the yellow boxed region the cytosolic domain which we focus on. (b) The scores produce by inputting a sequence into a hidden unit, producing a single number as output which corresponds to a summation of negatively and positively weighted residues. Performed on entire training set (histogram in blue), highlighting sequences corresponding to predominantly positive weighted residues. (c) Hidden unit from an RBM trained on the Pfam DUF2156 domain. The MSA positions 152 and 212 correspond to residues S684 and R742, respectively. (d) Residues (in yellow) in the MprF cytosolic domain which form the binding pocket for Lys-tRNA^{Lys} (the ligand analogue L-lysine amide shown in green), from PDB:4V36. LYN, L-lysine amide.

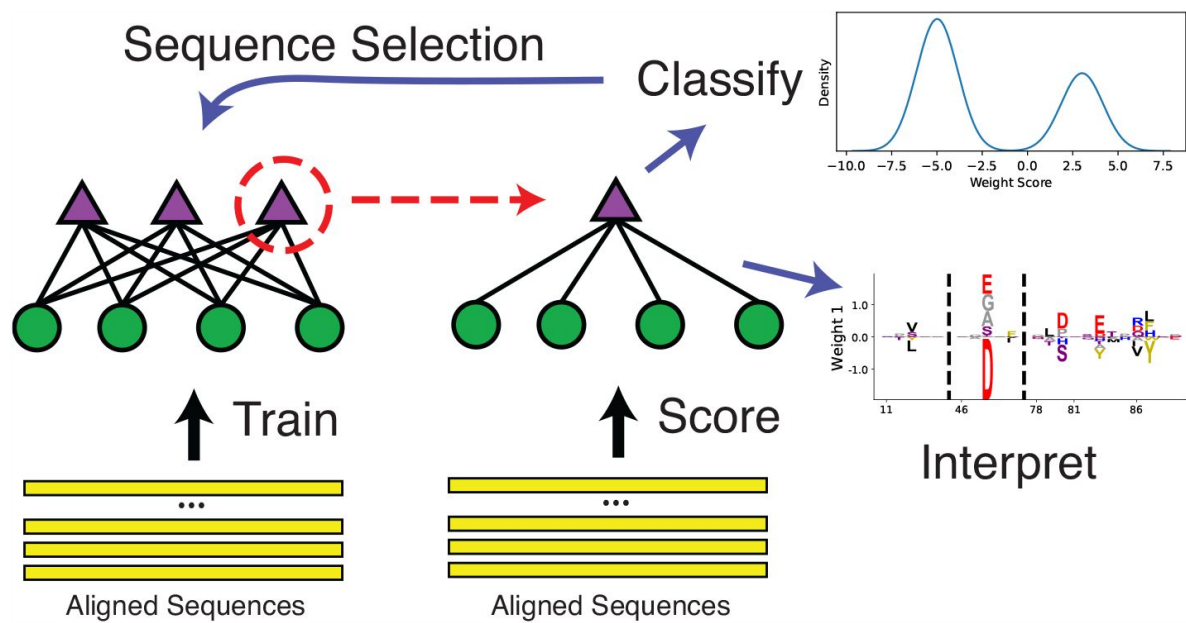


Figure 4.

Schematic of the RBM methodology. An aligned set of protein sequences is first used to learn a hidden unit representation that best describes the statistics of the sequence dataset given restrictions on the hidden unit representation. Then, the individual hidden units can be studied to find particular units which allow useful enzyme classification, and additionally, these weights can be meaningfully interpreted as statistically co-varying sequence configurations. Additionally, the classification can be used to create filtered datasets to train more models.

2.3 RBM weights identify plausible sequence residues for functional characterization

After filtering the data set and training another model, we set out to find weights which could describe MprF specificity. To this end, we identify weights which are sparse, have high overall magnitude, and involve residues which could be plausibly linked to the lipid binding specificity (see Methods for more details). Through this search method we identified the two hidden units shown in **Figure 5a**. They both have high sparsity and high magnitude relative to the majority of other weights in the model, and importantly they correspond to residues localized to regions of the catalytic domain that had been previously identified through Autodock experiments using a PG molecule with side chains C5:0/C8:0 [26] (highlighted in **Figure 5a**). The scoring of the training set of sequences with these weights is shown in **Figure 5b**, and we see a clear separation between the cytosolic domain sequences which produces Lys-Glc-DAG (from *S. agalactiae*) and two domain variants which produce only aminoacylated-PG (from *P. aeruginosa* and *B. licheniformis*). Importantly, these output scores depend on a greatly reduced subset of the full sequence and do not correspond to clustering on total sequence identity (**Figure S1**).

Using these two weights as sequence classifiers with a potential link to the substrate specificity of cytosolic domain variants, we used **Figure 5b** to propose sequences in quadrant three (Q3), the quadrant occupied by the *S. agalactiae* MprF variant with this novel glycolipid-modifying function. We selected a sequence which was distant in terms of sequence identity from the *Streptococcus* genus (**Table S1**) and which has been implicated in human pathologies [27], *Enterococcus dispar* MprF.

2.4 Enterococcal MprF enzymes synthesize Lys-Glc-DAG, Lys-PG, and a novel cationic glycolipid Lys-Glc₂-DAG

Enterococcus dispar ATCC 51266 was identified as a possible candidate for novel lipid synthesis through the analyses described above. Remarkably, our lipidomic analysis found that *E. dispar* synthesizes a novel cationic lipid, lysyl-diglucosyl(Glc₂)-diacylglycerol (Lys-Glc₂-DAG) (**Figure 1**, **Figure 6a,b**), as well as Lys-Glc-DAG and Lys-PG (**Figure 6c**, **Figure S2**). The structural identification of Lys-Glc₂-DAG was supported by exact mass measurement ($[M+H]^+$ observed at m/z 1047.730; calculated m/z 1047.731) and tandem MS (**Figure 6b**). Lys-Glc₂-DAG is unusually polar and charged for a lipid; chromatographically, Lys-Glc₂-DAG is highly retentive on a silica-based normal phase column and elutes off the column at the very end of the gradient (**Figure S2**). The use of an amino-based HPLC column led to much improved elution profiles for Lys-Glc₂-DAG and other glyco- and lysine lipids (**Figure 6**).

The discovery of Lys-Glc₂-DAG in *Enterococcus dispar* led us to analyze other *Enterococcus* strains: *E. faecium* 1,231,410, *E. faecalis* T11 and OG1RF, *E. gallinarium* EG2, *E. casseliflavus* EC10, and *E. raffinosus* Er676. All *Enterococcus* strains examined were found to synthesize Lys-PG, Lys-Glc-DAG and Lys-Glc₂-DAG at varying levels (**Figure 7a** and **Table 2**). MprF-dependent Lys-PG production by *E. faecalis* and *E. faecium* was previously reported [28, 29], but cationic glycolipids production has not been previously reported. *E. gallinarium* EG2 and *E. casseliflavus* EC10 encode a single copy of *mprF* gene, while the other enterococci we tested encode two distinct *mprF* genes (*mprF* 1 and *mprF* 2) [28, 29]. When these MprF sequence variants are plotted using our chosen RBM weights, we see that the two single copy strains plot in quadrant two (Q2), while for the two copy sequences, each MprF variant maps to separate coordinates (**Figure S3**). Both *E. raffinosus* Er676 MprF variants plot to quadrant two (with the lowest Weight 2 value being 0.34). The *E. faecium* and *E. faecalis* variants have copies plotting directly in quadrant one and variants plotting farther out from the sequence cluster; we note that the two displaced variants are both lacking the characteristic aspartic acid and glutamic acid amino acids at positions 100 and 48 in the weights (**Figure 5a**), which are the dominant contributions for determining quadrant one occupancy. To clarify which variants were determinants of the novel cationic lipid

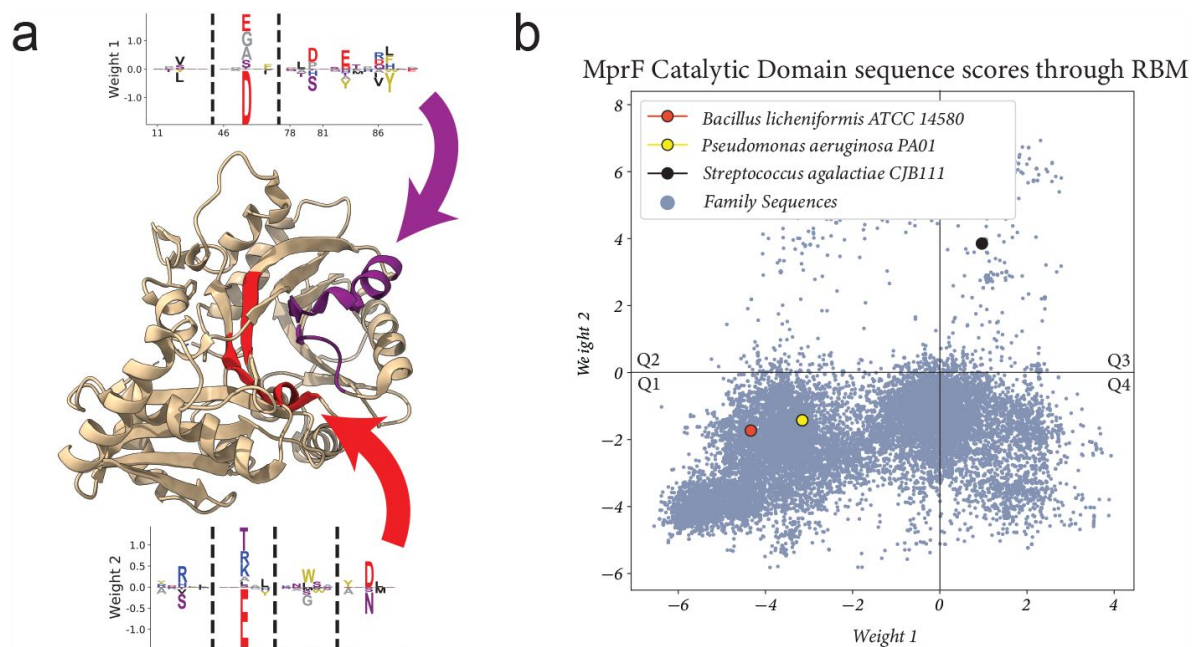


Figure 5.

Proposed set of weights for classifying lipid specificity. (a) two hidden units found in an RBM trained on the filtered Pfam dataset. The hidden unit residues are highlighted in the PDB:4V34 structure, with arrows pointing to their corresponding residue sets. (b) the outputs of the weights when scoring the sequences used in the training set and sequences from NCBI not used during training (N=23,138). *S. agalactiae* produces Lys-Glc-DAG and Lys-PG, while *B. licheniformis* and *P. aeruginosa* produces only aminoacylated-PG. Q1-Q4 are quadrant labels which we refer to throughout the paper.

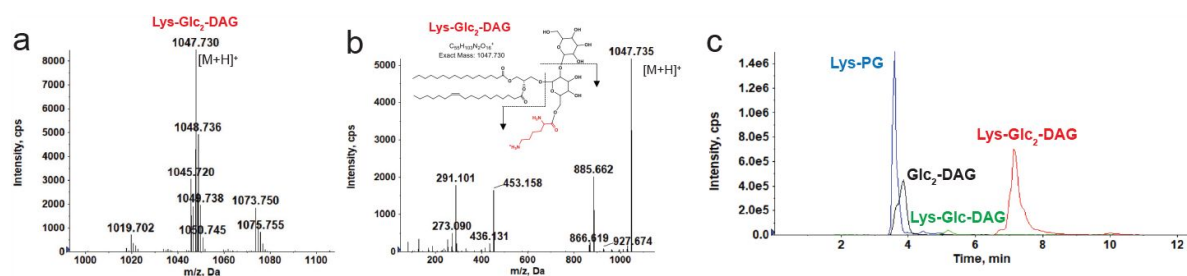


Figure 6.

Identification of Lys-Glc₂-DAG in *E. dispar*. (a) Positive ion mass spectrum of Lys-Glc₂-DAG species in *E. dispar*. (b) MS/MS product ions and fragmentation scheme of Lys-Glc₂-DAG. (c) Extracted ion chromatograms of LC/MS of lysine lipids (Lys-PG, Lys-Glc-DAG, Lys-Glc₂-DAG) and Glc₂-DAG separated on an amino HPLC column. Glc₂-DAG, diglucosyl-diacylglycerol; Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc-DAG, lysyl-glucosyl-diacylglycerol; Lys-Glc₂-DAG, lysyl-diglucosyl-diacylglycerol.

production, lipidomic analysis on an *E. faecalis* OG1RF mini-mariner (EfaMarTn) transposon mutant [30] with a transposon insertion within OG1RF 10760 *mprF* (referred to as *mprF* 2 in previous studies [28]; OG1RF 10760::Tn) revealed that Lys-PG, Lys-Glc-DAG and the newly identified Lys-Glc₂-DAG were absent when the synthase domain of the gene was disrupted (**Figure 7b**, **Table S1**). This led us to the conclusion that *E. faecalis* *mprF* 2 is necessary for the synthesis of the three Lys-lipids in *E. faecalis*. When the empty vector (pABG5) was transformed into OG1RF 10760::Tn, no Lys-lipids synthesis was observed as expected (**Figure 7c**). This phenotype was complemented by expression of *E. faecalis* *mprF* 2 from a plasmid (pOGMprF2) (**Figure 7d**). Additionally, expression of *E. faecium* *mprF*1 (EFTG 00601) (pEfMprF1) restored Lys-PG and Lys-Glc₂-DAG synthesis to the *E. faecalis* OG1RF 10760::Tn strain (**Figure 7e**), while *E. faecium* *mprF*2 (EFTG 02430) did not rescue Lys-lipid synthesis (**Figure 7f**). Taken together, these data indicate that *E. faecalis* MprF2 (OG1RF 10760) and *E. faecium* MprF1 (EFTG 00601), like *S. agalactiae* MprF, act on both phospholipid and glycolipids substrates, additionally synthesizing the novel cationic glycolipid, Lys-Glc₂-DAG.

2.5 Weight 2 is correlated with Glc_N-DAG substrate activity in MprF

When we started our RBM analysis, we had substrate specificity information for only the quadrants one and three, and through identifying the novel substrate in *E. dispar* we expanded our testing to quadrant two by testing *Enterococcus* strains. To expand our RBM analysis, we tested specific sequences which were found in quadrant four (Q4). We conducted lipid analysis on *Ligilactobacillus salivarius* ATCC 11741, *Levilactobacillus brevis* ATCC 14869, *Lactocaseibacillus rhamnosus* ATCC 7469, *Lactobacillus casei* ATCC 393, and *Lactocaseibacillus paracasei* ATCC 25302. All strains tested synthesize Lys-PG (**Table 2**). Only *Levilactobacillus brevis* and *Lactocaseibacillus paracasei* synthesized Lys-Glc₂-DAG, *L. paracasei* synthesized low levels of Lys-Glc₂-DAG. An additional quadrant one (Q1) strain *Exiguobacterium acetylicum* UTDF19-27C, was analyzed by lipidomics. It was found to synthesize only Lys-PG. This MprF falls near *Staphylococcus aureus* RN4220 and *Bacillus subtilis* 168 in our plot, both of which synthesize only Lys-PG [31, 24] (and experimentally confirmed in this study). This further supports that strains located in quadrant one of **Figure 8** are capable of Lys-PG synthesis.

Summaries of all species/variants tested and the Lys-lipids they produced are shown in **Table 2**, and plotted in the two RBM weights in **Figure 8**. We list for each MprF variant which of the three tested lipids it acts on, and in **Figure 8** we indicate whether a variant has only PG as a substrate or if it operates on Glc_N-DAG as a substrate irrespective of its PG activity. With this plot, we can see that sequences which plot with positive values in Weight 2 are more likely to have Glc_N-DAG as a substrate (8/9), and sequences with negative values predominantly use PG (7/11), with a Fisher's exact test p-value of 0.028 (**Table S2**). Weight 1, however, is not correlated with Glc_N-DAG specificity, with a Fisher's exact p-value of 0.67. Therefore we conclude that Weight 2, and thus the high magnitude sequence positions identified by it, are correlated with Glc_N-DAG specificity.

3 Discussion

The utilization of the *S. agalactiae* MprF amino acid sequence as a query for BLASTp was a simple method to identify various streptococcal MprFs and further our understanding of MprF specificity. This allowed us to identify three MprFs that can synthesize the highly cationic Lys-Glc-DAG. Importantly, although highly similar to *S. agalactiae* MprF, this method did not identify other streptococcal MprFs capable of lysine addition to both glycolipid and phospholipid substrates (generating Lys-Glc-DAG and Lys-PG) nor did it lead to the identification of a novel cationic glycolipid found in *Enterococcus*. Sequence logos [32] may identify important sequence features, but largely in instances where amino acids are highly conserved. When amino acids

	Bacterial Strains	Lys-Glc-DAG	Lys-Glc ₂ -DAG	Lys-PG
Q1	<i>Bacillus subtilis</i> 168	X	X	✓
	<i>Bacillus licheniformis</i> ATCC 14580	X	X	✓
	<i>Staphylococcus aureus</i> RN4220	X	X	✓
	<i>Exiguobacterium acetylicum</i> UTDF19-27C	X	X	✓
	<i>Enterococcus faecalis</i> T11/OG1RF**	✓	✓	✓
	<i>Enterococcus faecium</i> 1,231,410	✓	✓	✓
Q2	<i>Enterococcus raffinosus</i> Er676	✓	✓	✓
	<i>Enterococcus gallinarum</i> EG2	✓	✓	✓
	<i>Enterococcus casseliflavus</i> EC10	✓	✓	✓
Q3	<i>Streptococcus salivarius</i> * ATCC 7073	X	X	✓
	<i>Enterococcus dispar</i> ATCC 51266	✓	✓	✓
	<i>Streptococcus agalactiae</i> CJB111/COH1	✓	X	✓
	<i>Streptococcus sobrinus</i> * ATCC 27352	✓	X	X
	<i>Streptococcus downei</i> * (WP_002997695.1)	✓	X	X
	<i>Streptococcus ferus</i> * (WP_018030543.1)	✓	X	X
Q4	<i>Ligilactobacillus salivarius</i> ATCC 11741	X	X	✓
	<i>Lacticaseibacillus rhamnosus</i> ATCC 7469	X	X	✓
	<i>Lactobacillus casei</i> ATCC 393	X	X	✓
	<i>Levilactobacillus brevis</i> ATCC 14869	X	✓	✓
	<i>Lacticaseibacillus paracasei</i> ATCC 25302	X	✓	✓

Table 2

Table of all strains studied, the quadrant they occupy, and the lipids they synthesize.

** trace amounts Lys-Glc-DAG present in lipid extractions. * indicates heterologous expression of *mprF* in *Streptococcus mitis*.
Lys-Glc-DAG, lysyl-glucosyl-diacylglycerol; Lys-Glc₂-DAG, lysyl-diglucosyl-diacylglycerol; Lys-PG, lysyl-phosphatidylglycerol.

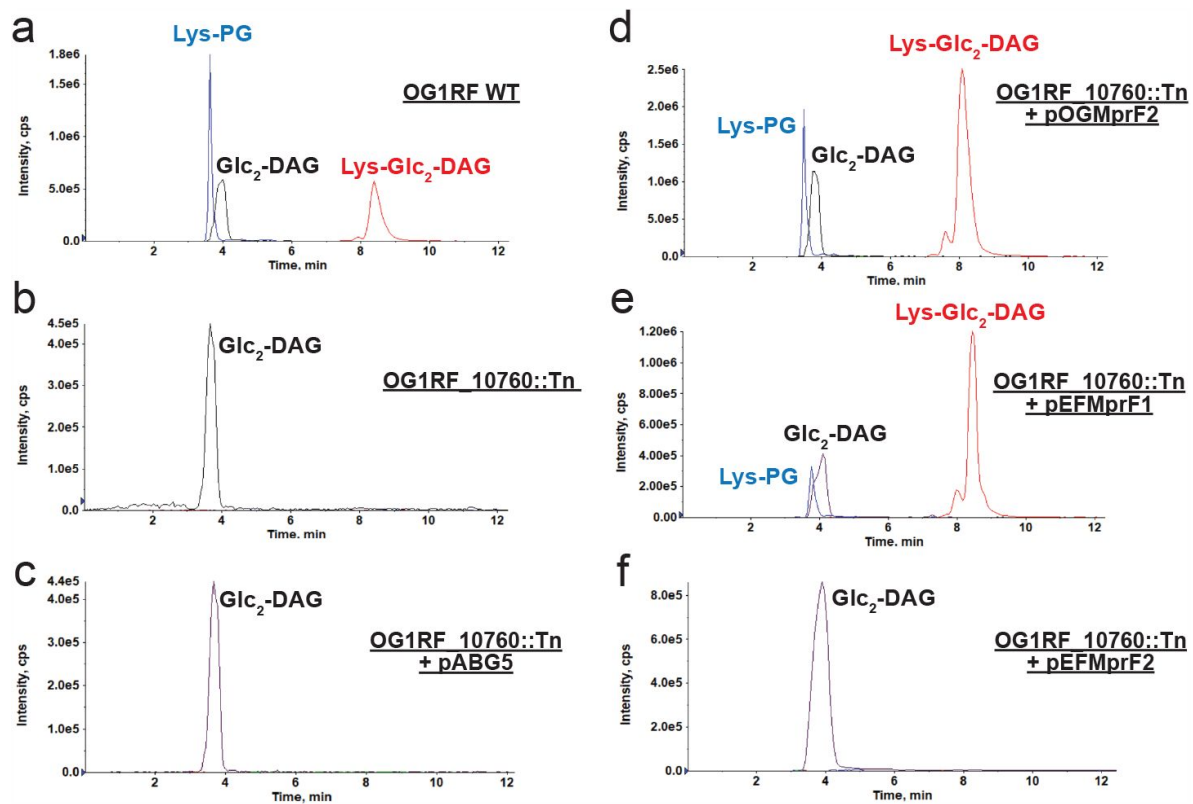


Figure 7.

E. faecalis MprF2 and *E. faecium* MprF1 confer Lys-Glc₂-DAG synthesis. (a) OG1RF-WT; (b) OG1RF 10760::Tn lacks lysine lipids; (c) OG1RF 10760::Tn + pABG5 lacks lysine lipids; (d) OG1RF 10760::Tn + pOGMprF2 restores lysine lipids; (e) OG1RF 10760::Tn + pEFMprF1 restores lysine lipids; (f) OG1RF 10760::Tn + pEFMprF2 lacks lysine lipids. Expression of OGMprF2 and EFMprF1 in OG1RF 10760 Tn mutant restore Lys-Glc₂-DAG synthesis. Shown are the extracted ion chromatograms of lysine lipids (Lys-PG, Lys-Glc₂-DAG) and Glc₂-DAG separated on an amino HPLC column. Note: Lys-Glc₂-DAG was found in trace amounts or missing from lipid extractions. Glc₂-DAG, diglucosyl-diacylglycerol; Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc₂-DAG, lysyl-diglucosyl-diacylglycerol.

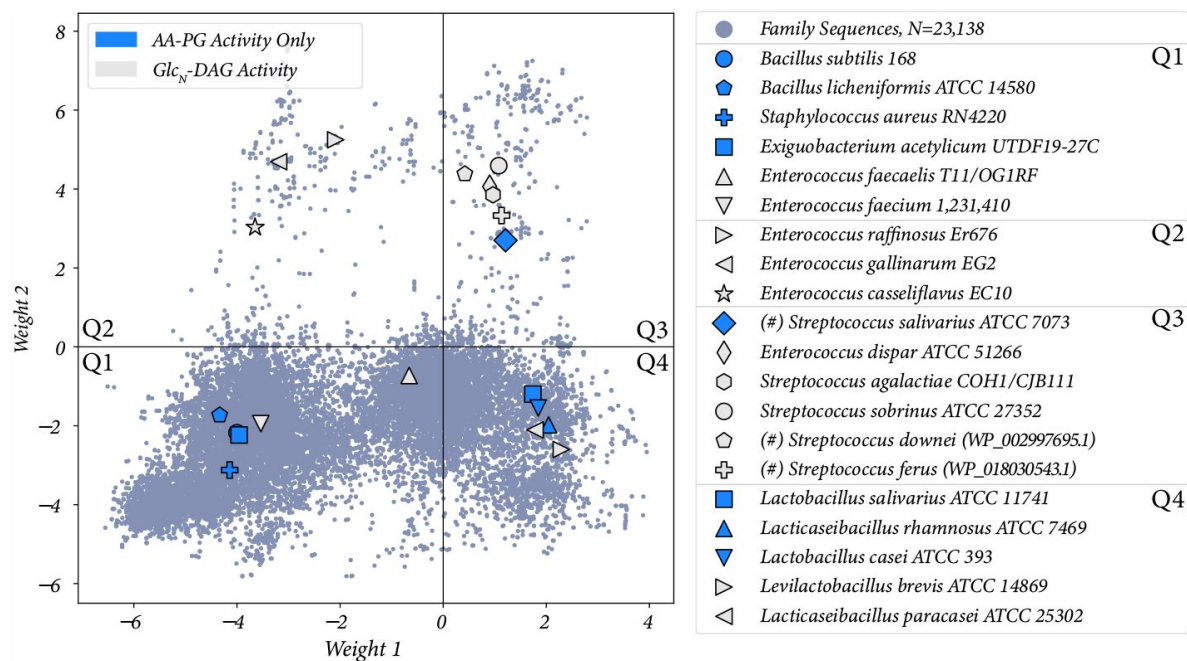


Figure 8.

The two proposed hidden unit outputs with all of the sequences from **Table 2** labeled. Protein ID of highlighted sequences are listed in **Table S1**. (#) indicates the lipid activity was confirmed through heterologous expression.

must interact to produce a specific function, yet the exact identity of those amino acids is not a requirement, methods which model the strength of amino acid coevolution in sequence sets can find important signals which in the sequence logo might appear to be negligible. For this reason we used the Restricted Boltzmann Machines (RBM) method [14], which can model statistical connectivity of residues at multiple sites in the protein sequence.

The RBM can be used in a number of ways to study sequence data. Firstly, it allowed us to filter the sequence data using hidden units (**Figure 3**); the two weights which formed the basis of our specificity analysis were not found in models trained on the Pfam dataset when it was initially pulled from Pfam, and this was potentially due to the presence of sequences without relevant catalytic function adding noise to the dataset. The most prominent feature for our purposes is the ability to use the hidden units learned through RBM training to classify sequences. The bimodal nature of this particular model provides a relatively simple interpretation which is straightforward to use for prediction. We show in **Figure 5** that the combination of clustering and interpretable weight structure allows us to identify a small subset of residues within a structure, grounding our statistical clustering in MprF's structural features. The combination of experimental evidence and evolutionary information provided a strong rationale for selecting organisms for further study, ultimately leading to the discovery of novel cationic lipid biosynthesis in *Enterococcus dispar* (**Figure 6**, **Figure S2**).

One goal of this study was to find the structural determinants of Glc_N-DAG activity by MprF, and we found that Weight 2 from our RBM analysis was predictive of this. However, there are notable exceptions such as the *E. faecium* allele EGTF 00601, which has sequence features placing it in quadrant one yet has Glc_N-DAG activity. Therefore, identifying the exact sequence determinants of Glc_N-DAG specificity remains to be fully elucidated. One explanation could be our exclusion of the N-terminal flippase domain and the non-cytosolic region of the C-terminal domain in the RBM analysis, where important residue interactions determining specificity could occur and is under further investigation. Additionally, a challenge in experimental validation will be to standardize the method of studying *mprF* alleles from diverse organisms. When using native organisms, lipids produced may vary during the growth phase of the organisms, the media they are grown in, and whether a membrane stressor is present (i.e. inducible production of specific lipids). For this study we used stationary phase cultures of equal volume and bacteria were grown in their respective standard laboratory media. The use of a heterologous host for expression of *mprF* alleles is a method of standardization, but may result in not identifying novel lipids that would be found in their native bacteria since the appropriate lipid substrates may not be present [16, 12]. A combination of both approaches (native organisms and heterologous hosts) may be required. Ultimately, mutational studies of the residues identified by the RBM will allow for identification of the residues critical for specificity, though likely the residues will depend in part on the primary sequence being altered.

The general method of RBM-guided exploration has applications in enzyme design, for example, in enhancing methods like directed evolution [33]. Concretely, the residues identified through weights can be thought of as templates for mutational exploration, greatly reducing the search space for assessing which residues are critical for specificity determination. Enzymes like MprF that produce lipids of modified charge have potential applications in biotechnology and therapeutics, where the design of lipid nanoparticles with different charge and chemical properties of the head group have important physiological implications [34]. Notably, the identification of Lys-Glc₂-DAG opens new research avenues, particularly in the fields of antibiotic resistance and host-pathogen interactions, wherein lipid modification and cationic lipids play key roles. With respect to human health, it would be relevant to investigate whether Lys-Glc₂-DAG plays a role in enterococcal resistance to the last-line membrane-active antibiotic, daptomycin.

4 Methods and Materials

4.1 Bacterial Strains and Growth Conditions

See [Tables S1](#) and [S3](#) for a full list of bacterial strains used in this study. *Streptococcus mitis* NCTC12261 (referred to as SM61 here) [\[35\]](#) was grown in Todd Hewitt Broth (THB) at 37°C with 5% CO₂. *Escherichia coli* DH5α [\[36\]](#) and *Bacillus subtilis* 168 [\[37\]](#) was grown at 37°C with shaking at 225 rpm in lysogeny broth (LB). *Staphylococcus aureus* RN4220 [\[38\]](#) was grown at 37°C with shaking at 220 rpm in Tryptic Soy Broth (TSB). *Streptococcus agalactiae* COH1 ATCC BAA-1176 [\[39\]](#) and CJB111 [\[40\]](#), [\[41\]](#) were grown in THB at 37 °C. *Enterococcus faecium* 1,231,410 [\[42\]](#), *Enterococcus faecalis* T11 [\[43\]](#) and OG1RF [\[44\]](#), *Enterococcus gallinarum* EG2 [\[42\]](#), *Enterococcus casseliflavus* EC10 [\[42\]](#), and *Enterococcus raffinosus* Er676 [\[45\]](#) were grown at 37°C in Brain Heart Infusion (BHI). *Enterococcus dispar* ATCC 51266 [\[46\]](#) was grown at 30°C in BHI. *Exiguobacterium acetylicum* UTDF19-27C was grown in BHI at 37 °C. *Ligilactobacillus salivarius* ATCC 11741 [\[47\]](#), *Lactocaseibacillus rhamnosus* ATCC 7469 [\[48\]](#), *Lactobacillus casei* ATCC 393 [\[49\]](#) and *Lactocaseibacillus paracasei* ATCC 25302 [\[48\]](#) were grown in Lactobacilli MRS Broth at 37°C and 5% CO₂. *Levilactobacillus brevis* ATCC 14869 [\[50\]](#) was grown in Lactobacilli MRS Broth at 30°C, 5% CO₂. *Bacillus licheniformis* ATCC 14580 was grown in nutrient broth at 37°C with shaking at 225 rpm [\[51\]](#). A transposon mutant of *E. faecalis* OG1RF from [\[30\]](#) with a mini mariner transposon (EfaMarTn) insertion in OG1RF 10760 (OG1RF 10760::Tn) was grown on BHI supplemented with chloramphenicol at a concentration of 15 μg/mL. The mutant was confirmed to be erythromycin-sensitive. The transposon mutant location was confirmed through Sanger sequencing performed by the Genome Center at the University of Texas at Dallas (Richardson, TX).

Where appropriate for plasmid selection, kanamycin (Sigma-Aldrich) was added. For *E. coli*, a concentration of 50 μg/mL was used, for *S. mitis* a concentration of 300 μg/mL, and for *E. faecalis* a concentration of 500 μg/mL was used.

4.2 Routine Molecular Biology Procedures

Genomic DNA was extracted as done previously in [\[52\]](#) and [\[53\]](#). All PCR reactions used Phusion polymerase (Thermo Fisher Scientific) and Phusion 5x HF buffer (Thermo Fisher Scientific) in a Veriti PCR machine (Applied Biosystems). List of primers used are found in [Table S4](#). Gibson assemblies were completed using a 2x HI-FI Assembly master mix following the manufacturer's protocol (New England Biolabs). PCR clean-up was done using the GeneJET PCR Purification Kit (Thermo Fisher Scientific) per manufacturer protocol. Plasmid extractions were performed per manufacturer protocol using the GeneJET Plasmid miniprep kit (Thermo Fisher Scientific). All plasmid constructs were confirmed by Sanger sequencing at the Massachusetts General Hospital CCIB DNA Core facility or by Illumina sequence at SeqCenter. DNA concentrations were measured using Nanodrop (Thermo Fisher Scientific) or Qubit 2.0 (Invitrogen by Life Technologies). Optical Density at 600 nm (OD_{600nm}) was measured in a disposable cuvette (Thermo Fisher Scientific) using a spectrophotometer (Thermo Scientific Genesys 30).

Gibson assembly was performed using pABG5 as previously described [\[11\]](#). Transformation of plasmids into *E.coli*, *S. mitis*, and *E.faecalis* was performed as previously described in [\[11\]](#), [\[16\]](#), [\[54\]](#), [\[55\]](#).

4.3 Acidic Bligh-Dyer Lipid Extractions

Bacteria were grown for approximately 16 hours overnight in 15 mL of an appropriate culture medium. After growth, OD_{600nm} measurements were taken, and cells were pelleted at 4,280 x g for 5 minutes at room temperature in a Sorvall RC6+ centrifuge. The supernatant was tipped out, and the cell pellet was washed and resuspended in 1X Phosphate Buffered Saline (PBS). The cells were

pelleted again, and all the supernatant was aspirated out. The cell pellet was stored at -80°C until lipid extraction. An acidic Bligh-Dyer lipid extraction was performed as previously reported [16, 12]. Briefly, cell pellets were resuspended in 0.8 mL of 1x Dulbecco's PBS (Sigma-Aldrich). Cells were transferred to a 9 mL glass tube with a Teflon-lined cap (Pyrex). 1 mL chloroform (MilliporeSigma) and 2 mL methanol (Thermo Fisher Scientific) were added to create the single-phase Bligh-Dyer. Tubes were vortexed every 5 minutes for 20 minutes, then centrifuged at $500 \times g$ for 10 minutes at room temperature. The supernatant was transferred to a new 9 mL glass tube. 100 μL of Hydrochloric acid (Thermo Fisher Scientific) was added followed by 1 mL of chloroform and 0.9 mL of 1x Dulbecco's PBS to create the two-phase Bligh-Dyer. Tubes were gently mixed and centrifuged at $500 \times g$ for 5 minutes at room temperature. After centrifuging, the bottom layer was extracted to a new tube and dried under nitrogen gas and stored at -80°C prior to lipidomic analysis. Lipid analyses were repeated in biological triplicate aside, from OG1RF 10760::Tn extractions, which were performed once.

4.4 Analysis of lysine lipids by an amino column-based Liquid Chromatography-Electrospray Ionization-Mass Spectrometry (LC-ESI MS): A new method

An amino column-based normal phase LC-ESI MS was performed using an Agilent 1200 Quaternary LC system coupled to a high-resolution TripleTOF5600 mass spectrometer (Sciex, Framingham, MA). A Unison UK-Amino column ($3 \mu\text{m}$, $25 \text{ cm} \times 2 \text{ mm}$) (Imtakt USA, Portland, OR) was used. Mobile phase A consisted of chloroform/methanol/aqueous ammonium hydroxide (800:195:5, v/v/v). Mobile phase B consisted of chloroform/methanol/water/aqueous ammonium hydroxide (600:340:50:5, v/v/v/v). Mobile phase C consisted of chloroform/methanol/water/aqueous ammonium hydroxide (450:450:95:5, v/v/v/v). The elution program consisted of the following: 100% mobile phase A was held isocratically for 2 min and then linearly increased to 100% mobile phase B over 8 min and held at 100% B for 5 min. The LC gradient was then changed to 100% mobile phase C over 1 min and held at 100% C for 3 min, and finally returned to 100% A over 0.5 min and held at 100% A for 3 min. The total LC flow rate was $300 \mu\text{L}/\text{min}$. The MS settings were as follows: Ion spray voltage (IS) = 5000 V, Curtain gas (CUR) = 20 psi, Ion source gas 1 (GS1) = 20 psi, De-clustering potential (DP) = 50 V, and Focusing Potential (FP) = 150 V. Nitrogen was used as the collision gas for MS/MS experiments. Data acquisition and analysis were performed using Analyst TF1.5 software (Sciex, Framingham, MA).

4.5 Generation of *mprF* expression plasmids

The *S. downei mprF* nucleotide sequence (Locus tag HMPREF9176 RS03810) and *S. ferus mprF* nucleotide sequence (Locus tag A3GY RS0106165) were used to design synthetic geneblocks (GeneWiz from Azenta Life Sciences). A 20 nucleotide 5' extension and a 20 nucleotide 3' extension complementary to the pABG5 plasmid were added to enable Gibson assembly of the inserts with pABG5. Sequences for the geneblocks can be found in Table S5. The *S. salivarius mprF* (Locus tag SSAL8618 04345) was amplified from *Streptococcus salivarius* ATCC 7073. A 20 nucleotide 5' extension and a 20 nucleotide 3' extension complementary to the pABG5 plasmid was added to the insert. The same protocol was used for generation of pSobrinus, pEFM-prF1, pEFMprF2, and pOGMprF2 with slight modifications. *S. sobrinus mprF* (Locus tag DLJ52 05040) was amplified from *Streptococcus sobrinus* ATCC 27352. *E. faecium mprF1* (Locus tag EFTG 00601) and *E. faecium mprF2* (Locus tag EFTG 02430) were amplified from *Enterococcus faecium* 1, 231, 410. *E. faecalis mprF2* (Locus tag OG1RF 10760) was amplified from *E. faecalis* OG1RF.

4.6 Boltzmann Machines and Restricted Boltzmann Machines for Protein Sequences

Formulation of models

The Boltzmann Machine, as described in [18], is defined as:

$$P(\mathbf{v}) = \frac{1}{Z} \exp\left\{\sum_{i<j} e_{ij}(v_i, v_j) + \sum_i h_i(v_i)\right\} \quad (2)$$

which is a probability distribution defined on the pairwise interactions (the couplings e_{ij}) between the positions (v_n) in an input sequence ($\mathbf{v}$) and a local field term h_i . The pairwise coupling matrix is akin to a covariance matrix, and the local fields are similar to single site frequency measures.

The joint probability distribution of the Restricted Boltzmann Machine as described in [14] is as follows:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp\left(\sum_{i=1}^N g_i(v_i) - \sum_{\mu=1}^M \mathcal{U}_{\mu}(h_{\mu}) + \sum_{i\mu} h_{\mu} w_{i\mu}(v_i)\right) \quad (3)$$

Here, the vector $\mathbf{v}$ again represents the input sequence data, $g(v_i)$ is a local field which controls the conditional probability of the input data, $\mathcal{U}_{\mu}$ is the hidden unit potential/activation, and the terms h_i and $w_{i\mu}(v)$ couple the input variables with the hidden unit variables. In this way, interactions between residues are mediated by a relatively smaller set of hidden units which do not interact directly with each other, leading to a bipartite structure as opposed to the Boltzmann machine's fully connected pair based formulation.

Particularly important is the potential function $\mathcal{U}$ which in this work is the dReLU function defined as:

$$\mathcal{U}_{\mu}(h) = \frac{1}{2}\gamma_{\mu,+}h_+^2 + \frac{1}{2}\gamma_{\mu,-}h_-^2 + \theta_{\mu,+}h_+ + \theta_{\mu,-}h_- \quad \text{where} \quad (4)$$

$$h_+ = \max(h, 0), \quad h_- = \min(h, 0)$$

The four parameters $\gamma_{\mu,\pm}$, $\theta_{\mu,\pm}$ are learned through the inference procedure, and allow the positive and negative components of the weights w_{μ} to be separately gated, which allows learning of bimodal distributions for the output h_{μ} . Another key feature is the sparsity regularization on the weights applied during model inference:

$$L2/L1 = \frac{\lambda_1^2}{2qN} \sum_{\mu} \left(\sum_{i,v} |w_{i\mu}(v)| \right)^2 \quad (5)$$

where q is the number of amino acids plus a gap character (21) and the hyperparameter λ_1^2 can be increased to induce sparsity and lower weight magnitude. Equations 4 and 5 together guide the representation of the hidden units to bimodal outputs with interpretable features.

Dataset acquisition and preprocessing

Multiple sequence alignment used for model training was the DUF2156 domain acquired from Pfam [25] (in later revisions renamed LPG synthase C). Starting with this MSA, it was processed to include only amino acid characters and gaps, excluding sequences with ambiguous or non-standard characters. Additionally, sequences with contiguous strings of gaps greater than 20% of

the full sequence's length were removed. This left 11,507 sequences with a length of 298. After the data cleaning process described in **Figure 3**, a total of 7,890 sequences were used to train our final model.

Model training

The model previously described [14] was utilized, specifically the Python 2.7 version freely available on Github (<https://github.com/jertubiana/ProteinMotifRBM>). Extensive testing of the combinations of the number of hidden units, $L2/L1$ regularization, and learning rate were performed. Models were trained in triplicate with different random seeds, as the training procedure converges to slightly different weight values depending on their initial random configuration. Training was considered successful when the model consistently produced similar sets of weights across different random seeds, and the weights had the previously mentioned sparsity and magnitude indicating compositional representation. The final model used in this work was trained for 2000 epochs with a learning rate of 0.1, a learning rate decay of 0.33, $L2/L1$ parameter λ_1^2 set to 1.0, and 300 hidden units.

Model Weight selection

To find weights which were relevant to function we computed the $L1$ norm of each weight individually, then ranked them. The weights with the largest magnitude were looked at first, and these were typically the sparsest of all weights. Weights were chosen which involved sequence coordinates implicated in our function of interest. Specifically, locations identified through Autodock [26] where the lipid was likely to interact, and a small radius around this region to select a small set of coordinates.

Software Accessibility

DUF2156 domain sequences were retrieved from Pfam33 [25]. Raw sequence datasets are available as Supplementary Material.

Acknowledgements

J.M. and F.M. acknowledge support from the National Institutes of Health (NIH R35GM133631). F.M. acknowledges support from the National Science Foundation CAREER award (MCB-1943442). We acknowledge support from the National Institutes of Health grant R01AI178692 (Z.G., F. M. and K.P.) and R01AI148366 (Z.G. and K.P.). K.P. acknowledges support from the Cecil H. and Ida Green Chair in Systems Biology Science. L.R.J acknowledges support from the American Heart Association (23POST1013835)

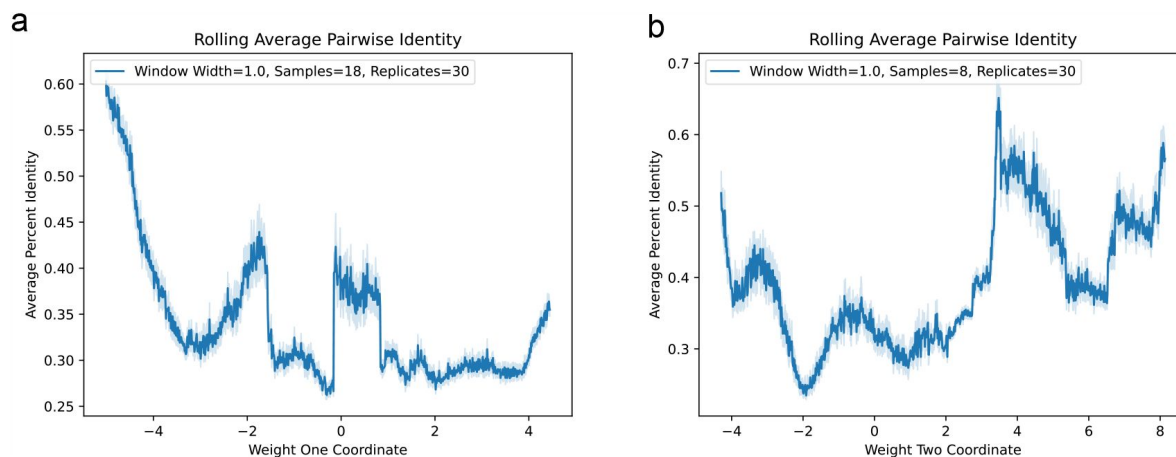


Figure S1

Assessment of RBM weight's reliance on total sequence identity in classification. A sliding window average is performed, where a window of width 1 is slid from the negative end to the positive of the weight, where at each position sequence coordinates are sampled from the window. These sampled sequences are compared pairwise between themselves, computing the Hamming distance across their entire sequence length to produce an average. This sampling procedure is repeated 30 times for each window, with the light blue shading representing the 95% confidence interval.

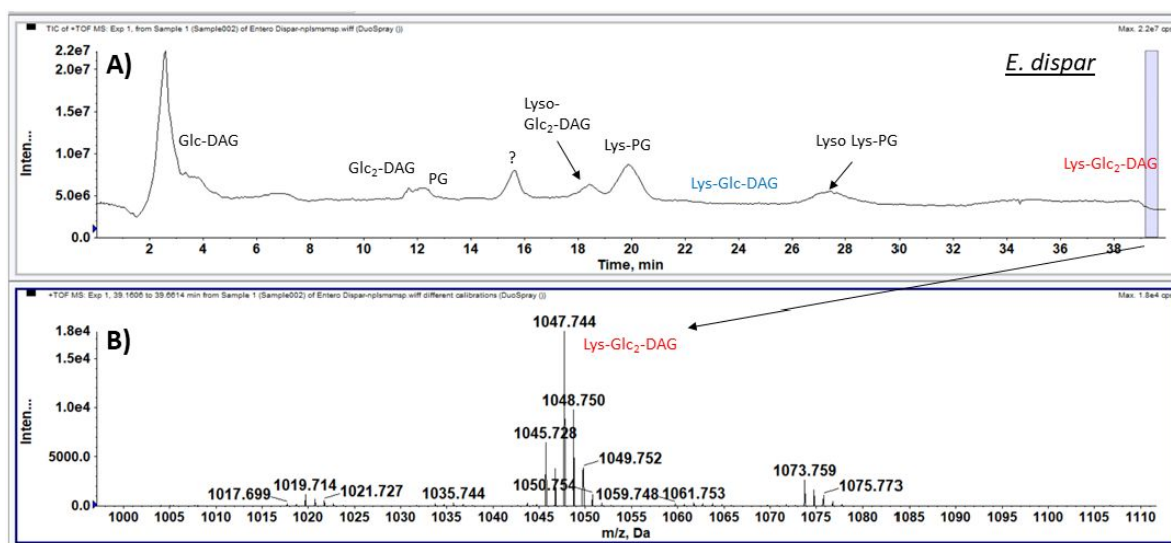


Figure S2

Identification of Lys-Glc₂-DAG which is highly retentive on a silica HPLC column. A) the total ion chromatogram of normal phase LC/MS of *E. dispar* lipids separated on a silica HPLC column. B) the positive ion mass spectrum of Lys-Glc₂-DAG eluting at the end of the LC gradient. Glc-DAG, glucosyl-diacylglycerol; Glc₂-DAG, diglucosyl-diacylglycerol; PG, phosphatidylglycerol; Lyso-Glc₂-DAG, lyso diglucosyl-diacylglycerol; Lys-PG, lysyl-phosphatidylglycerol; Lys-Glc-DAG, lysyl-glucosyl-diacylglycerol; Lyso Lys-PG, lyso lysyl-phosphatidylglycerol; Lys-Glc₂-DAG, lysyl-diglucosyl-diacylglycerol.

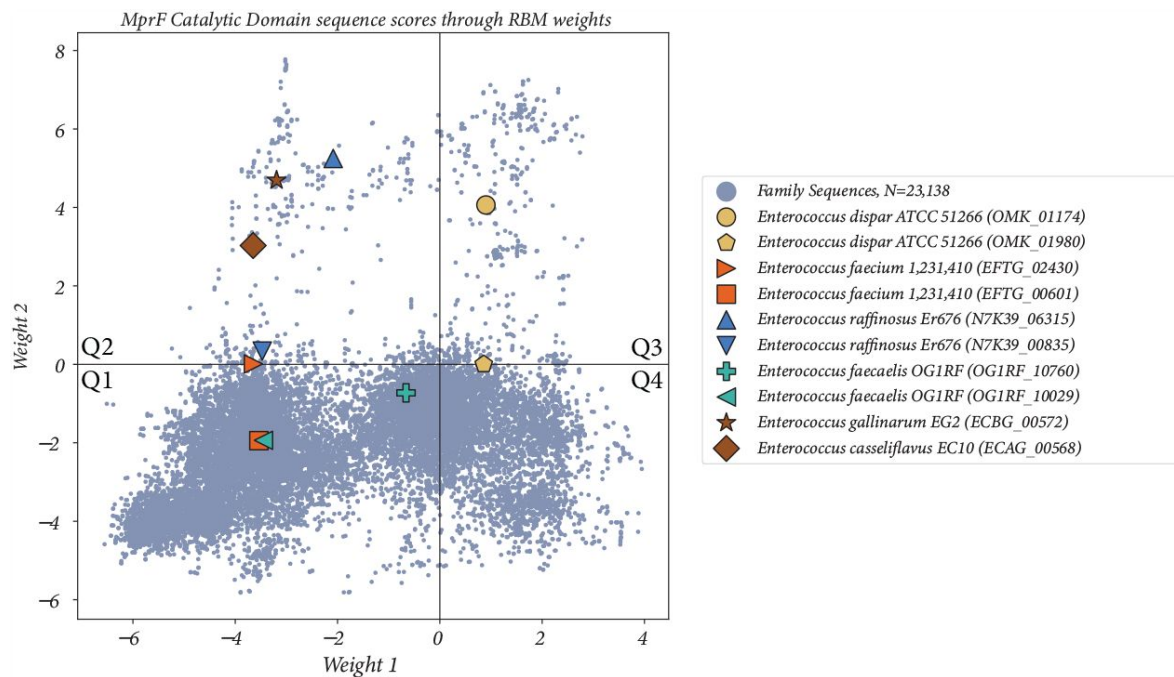


Figure S3

All *Enterococcus* sequences analyzed in the course of this study.

Bacterial Strain	Protein ID	Locus ID	Description	Reference
<i>Enterococcus dispar</i> ATCC 51266 (<i>mprF1</i>)	WP_016173119	OMK_01174	Wild-type <i>E. dispar</i> strain	Collins et al. (1991)
<i>Enterococcus dispar</i> ATCC 51266 (<i>mprF2</i>)	WP_016173252	OMK_01980	Wild-type <i>E. dispar</i> strain	Collins et al. (1991)
<i>Enterococcus gallinarum</i> EG2	WP_195971496	ECBG_00572	Wild-type <i>E. gallinarum</i> strain	Palmer et al. (2010)
<i>Enterococcus casseliflavus</i> EC10	WP_201705909	ECAG_00568	Wild-type <i>E. casseliflavus</i> strain	Palmer et al. (2010)
<i>Enterococcus faecalis</i> T11 (<i>mprF1</i>)	EEU90403	EFMG_00309	Wild-type <i>E. faecalis</i> strain	McBride et al. (2007)
<i>Enterococcus faecalis</i> T11 (<i>mprF2</i>)	EEU90403	EFMG_01835	Wild-type <i>E. faecalis</i> strain	McBride et al. (2007)
<i>Enterococcus faecalis</i> OG1RF (<i>mprF1</i>)	WP_002414160	OG1RF_10029	Wild-type <i>E. faecalis</i> strain	Gold et al. (1975)
<i>Enterococcus faecalis</i> OG1RF (<i>mprF2</i>)	WP_002358907	OG1RF_10760	Wild-type <i>E. faecalis</i> strain	Gold et al. (1975)
<i>Enterococcus faecium</i> 1,231,410 (<i>mprF1</i>)	WP_002294379	EFTG_00601	Wild-type <i>E. faecium</i> strain	Palmer et al. (2010)
<i>Enterococcus faecium</i> 1,231,410 (<i>mprF2</i>)	WP_002287340	EFTG_02430	Wild-type <i>E. faecium</i> strain	Palmer et al. (2010)
<i>Enterococcus raffinosus</i> Er676	UXJ96875	N7K39_06315	Wild-type <i>E. raffinosus</i> strain	Palmer et al. (2010)
<i>Streptococcus agalactiae</i> COH1	CDN67397	GBSCOH1_1931	Wild-type <i>S. agalactiae</i> strain, serotype III	Sharon et al. (2022)
<i>Streptococcus agalactiae</i> CJB111	QOW76674	ID870_10050	Wild-type <i>S. agalactiae</i> strain, serotype V	Kuyper et al. (1989)
<i>Staphylococcus aureus</i> RN4220	QWF27327	NMK49_RS06325	Wild-type <i>S. aureus</i> strain	Faralla et al. (2014); Spencer et al. (2021)
<i>Bacillus subtilis</i> 168	QJF41380	BSU_08425	Wild-type <i>B. subtilis</i> strain	Kreiswirth et al. (1983)
<i>Bacillus licheniformis</i> ATCC 14580	WP_003179900	EJ992_04410	Wild-type <i>B. licheniformis</i> strain	Kunst et al. (1997)
<i>Erwinia chrysanthemi</i> UTDF19-27C	WP_238238782	K6T22_02680	Wild-type <i>E. chrysanthemi</i> strain	Chester (1901)
<i>Ligilactobacillus salivarius</i> ATCC 14580	EEJ74510	HMPREF0545_0714	Wild-type <i>L. salivarius</i> strain	Drucker (1979)
<i>Lactococcus lactis</i> rhamnosus ATCC 7469		NHLLHDL_02198	Wild-type <i>L. rhamnosus</i> strain	Collins et al. (1989)
<i>Lactobacillus casei</i> ATCC 393	BAN75198	LBCZ_2030	Wild-type <i>L. casei</i> strain	Hansen and Lesse (1971)
<i>Leuconostoc brevis</i> ATCC 14869	ERK41244	HMPREF0495_02376	Wild-type <i>L. brevis</i> strain	Rogosa and Hansen (1971)
<i>Lactococcus lactis</i> paracasei ATCC 25302	TDG80985	C5L26_003005	Wild-type <i>L. paracasei</i> strain	Collins et al. (1989)
<i>Streptococcus mitis</i> SM61 + pFerus	WP_018030543	A3GY_RS0106165	Expression of <i>S. ferus mprF</i> in <i>S. mitis</i>	This work
<i>Streptococcus mitis</i> SM61 + pDownei	WP_002997695	HMPREF9176_1523	Expression of <i>S. downei mprF</i> in <i>S. mitis</i>	This work
<i>Streptococcus mitis</i> SM61 + pSobrinus	WP_019781626	DLJ52_05040	Expression of <i>S. sobrinus mprF</i> in <i>S. mitis</i>	This work
<i>Streptococcus mitis</i> SM61 + pSalivarius	WP_181671155	SSAL8618_04345	Expression of <i>S. salivarius mprF</i> in <i>S. mitis</i>	This work
<i>Enterococcus faecalis</i> OG1RF.10760::Tn + pABG5	WP_002358907	OG1RF_10760	OG1RF with a mini-mariner transposon (EfaMarTn) inserted into <i>MprF2</i> at position 787608	Dale et al. (2018)
<i>Enterococcus faecalis</i> OG1RF.10760::Tn + pOGMprF2	WP_002358907	OG1RF_10760	OG1RF.10760::Tn with the empty vector pABG5 expressed	This work
<i>Enterococcus faecalis</i> OG1RF.10760::Tn + pEMMprF1	WP_002294379	EFTG_00601	Complementation of the transposon. pOG.MprF2 expresses the OG1RF.10760 gene on the pABG5 vector	This work
<i>Enterococcus faecalis</i> OG1RF.10760::Tn + pEMMprF2	WP_002287340	EFTG_02430	Expression of <i>MprF1</i> from <i>E. faecium</i> 1, 231, 410 into the OG1RF.10760::Tn	This work
			Expression of <i>MprF2</i> from <i>E. faecium</i> 1, 231, 410 into the OG1RF.10760::Tn	This work

Table S1

Sequence Locus IDs for all sequences listed in Table 2 [↗](#).

Bold denotes confirmed *mprF* allele.

	Weight 2 Positive	Weight 2 Negative
Glc_N-DAG	8	4
PG Only	1	7

Table S2

Fisher’s exact test for determining whether positive values of Weight 2 are predictive of Glc_N-DAG specificity. p=0.028.

<i>E.coli</i> and Plasmid	Description	Reference
<i>E.coli</i> DH5α (pFerus)	<i>E. coli</i> with <i>S. ferus mprF</i> in pABG5	This work
<i>E.coli</i> DH5α (pDownei)	<i>E. coli</i> with <i>S. downei mprF</i> in pABG5	This Study
<i>E.coli</i> DH5α (pSalivarius)	<i>E. coli</i> with <i>S. salivarius</i> ATCC 7073 <i>mprF</i> in pABG5	This Study
<i>E.coli</i> DH5α (pSobrinus)	<i>E. coli</i> with <i>S. sobrinus</i> ATCC 27352 <i>mprF</i> in pABG5	This Study
<i>E.coli</i> DH5α (pEFmprF1)	<i>E. coli</i> with <i>E. faecium mprF1</i> (EFTG_00601)	Joyce et al. (2022)
<i>E.coli</i> DH5α (pEFmprF2)	<i>E. coli</i> with <i>E. faecium mprF2</i> (EFTG_02430) in pABG5	This Study
<i>E.coli</i> DH5α (pOGMprF2)	<i>E. coli</i> with <i>E. faecalis mprF2</i> (OG1RF_10760) in pABG5	This Study
<i>E.coli</i> DH5α (pABG5)	<i>E. coli</i> with the empty vector (pABG5)	Joyce et al. (2022)

Table S3

***E.coli* and plasmids used**

Primer Name	Sequence	Use	Reference
pABG5-5'	GGAAAGGGACCTCTCTTCTCTAAAC	Linearize pABG5 with pABG5-3'	Joyce et al. (2022)
pABG5-3'	GATAAAGGTATTGGTAAATAACAAA	Linearize pABG5 with pABG5-5'	Joyce et al. (2022)
Faec_MprF1_F	GAGAGGTCCCTTTCC TTGTTAAAAAATACCATAACAATG	Primer for amplification of <i>E. faecium</i> <i>mprF1</i> with Faec-MprF1.R contains complementary sequence to pABG5 for Gibson assembly	Joyce et al. (2022)
Faec_MprF1_R	ACCAATACCTTTATC TTAATACTTCTTCGTATCC	Primer for amplification of <i>E. faecium</i> <i>mprF1</i> with Faec-MprF1.R	Joyce et al. (2022)
Faec_MprF2_F	GAGAGGTCCCTTTCC GTGAAAAATAACGATACTTG	Primer for amplification of <i>E. faecium</i> <i>mprF2</i> with Faec-MprF2.R contains complementary sequence to pABG5 for Gibson assembly	This Study
Faec_MprF2_R	ACCAATACCTTTATC TTATAATTCTGTGAGCACTTC	Primer used for amplification of <i>E. faecium</i> <i>mprF2</i> with Faec-MprF2.R	This Study
OG1RF_MprF2_F	GAGAGGTCCCTTTCC ATGAAACAAATTATTCATTGGATC	Primer used for amplification of <i>E. faecalis</i> <i>mprF2</i> with OG1RF_MprF2.R	This Study
OG1RF_MprF2_R	ACCAATACCTTTATC TTAGTCAATATTTTTTTTGATCGAC	Primer used for amplification of <i>E. faecalis</i> <i>mprF2</i>	This Study
SalMprF_R	ACCAATACCTTTATC CTATTTTCCAAAGACGTC	Primer used for amplification of <i>S. salivarius</i> <i>mprF</i> with SalMprF_F	This Study
SalMprF_F	GAGAGGTCCCTTTCC ATGAAGCGCTCTTTTAGAG	Primer used for amplification of <i>S. salivarius</i> <i>mprF</i>	This Study
SobMprF_F	GAGAGGTCCCTTTCC TTGAAAGCGCTTTCTTG	Primer used for amplification of <i>S. sobrinus</i> <i>mprF</i> with SobMprF_R	This Study
SobMprF_R	ACCAATACCTTTATC TTAATTCCGAGTCACAC	Primer used for amplification of <i>S. sobrinus</i> <i>mprF</i>	This Study
EF1.S1	GAATAACGCTGATCAAAAGT	Primer used for sequencing pEFMprF1 insert with pABG5Fup2	Joyce et al. (2022)
EF1.S2	TGCCAAGAGAAATAGTC	Primer used for sequencing pEFMprF1 insert with EF1.S3	Joyce et al. (2022)
EF1.S3	ACAATCTCTTCGCTTG	Primer used for sequencing pEFMprF1 insert	Joyce et al. (2022)
EF1.S4	CCAACTGTTCTTCTCCAA	Primer used for sequencing pEFMprF1 insert with pABG5Fdown2	Joyce et al. (2022)
EF2.S1	CTCCTGTATGTGTGATGAAC	Primer used for sequencing pEFMprF2 insert with pABG5Fup2	This Study
EF2.S2	GGAATGGGGACTTTTGA	Primer used for sequencing pEFMprF2 insert with EF2.S3	This Study
EF2.S3	ACTGATTCGTTGTGTC	Primer used for sequencing pEFMprF2 insert	This Study
EF2.S4	CTTCGCTTGTGAGTTAC	Primer used for sequencing pEFMprF2 insert with pABG5Fdown2	This Study
OG2.S1	TTGTAAAGAACAAACAGCC	Primer used for sequencing pOG_MprF2 insert with pABG5Fup2	This Study
OG2.S2	GCAACTTAGGTGTGCG	Primer used for sequencing pOG_MprF2 insert with OG.S3	This Study
OG2.S3	GCCGAAACGATCTGT	Primer used for sequencing pOG_MprF2 insert	This Study
OG2.S4	CGTTGTGATGGGCAAC	Primer used for sequencing pOG_MprF2 insert with pABG5Fdown2	This Study
Sal.S1	GAAGGGAATAACGTAG	Primer used for sequencing pSalivarius insert with pABG5Fup2	This Study
Sal.S2	TATCTTCTCTATGATTCC	Primer used for sequencing pSalivarius insert with Sal.S3	This Study
Sal.S3	TCTTCAAGAAAGATTCCCA	Primer used for sequencing pSalivarius insert	This Study
Sal.S4	CTCTTTTGGTATCGTGTC	Primer used for sequencing pSalivarius insert with pABG5Fdown2	This Study
Sob.S1	AGGAATGACATAATAAGCCAAG	Primer used for sequencing pSobrinus insert with pABG5Fup2	This Study
Sob.S2	CGGTATTATTTCTCTGAT	Primer used for sequencing pSobrinus insert with Sob.S3	This Study
Sob.S3	TTCTACATCGGCCTCAT	Primer used for sequencing pSobrinus insert	This Study
Sob.S4	TTAGCCTTTCTGGGAGAC	Primer used for sequencing pSobrinus insert with pABG5Fdown2	This Study
Downei.S1	GAACATAATAGGCCAGGC	Sequencing primer for pDownei insert with pABG5Fup2	This Study
Downei.S2	CGGTGGGGATTATTTCC	Sequencing primer for pDownei insert with Downei.S3	This Study
Downei.S3	CTGCTTGGTTACCAGCAG	Sequencing primer for pDownei insert	This Study
Downei.S4	CAAAAGCCTCTTCTG	Sequencing primer for pDownei insert with pABG5Fdown2	This Study
Ferus.S1	CCGGCAAAGAATGGCAGAATA	Sequencing primer for pFerus insert with pABG5Fup2	This Study
Ferus.S2	CACATAATCCCTGGCGG	Sequencing primer for pFerus insert with Ferus.S3	This Study
Ferus.S3	CAATGAAGGCTCTTGTAGC	Sequencing primer for pFerus insert	This Study
Ferus.S4	GTATCAGGAGGAGGGAC	Sequencing primer for pFerus with pABG5Fdown2	This Study
pABG5Fup2	CAAAAGGTTTCGACTTTTCACC	Primer used for sequencing with pABG5Fup2	Joyce et al. (2022)
pABG5Fdown2	CCAATAATAATGACTAGAGAAG	Primer used for sequencing	Joyce et al. (2022)

Table S4

Primers used in this study.

Red indicates sequence complementarity to pABG5.

Name	Sequence
<i>S. Downei mprF</i> (Locus tag HMPREF9176_RS03810)	AGGAAGAGAGGTCCCCTTTCC TTGAAAGCCCTTCTTGAAAAATCAAGCCCTTACCTCCATCTTAA ACCGTCTTCTTTGTTTCATCATCTCTTAATCATCATGGAAATCCTCAATCTGAGGCGGACCAATTCTGTGTCGCAACTCAAGG CGGCATTGGGTGGGTGTCTCCTATCAATATTCTCTGATTTTTATCATTGGAGCCCTGGCCGTCTACCAACGACGGGCTATGA TTTTGTCTCAATAAAATTTGGGTACCAATAAGTCCAAGTGGTATATTTTGACAGACCTCTGGTGTATCAATACCTTTAATAAT TTATCGGGTTTTGGCGGTCTGATTGATATTGGTCTGGCTTTGGCTTCTACGGGCAAAAGGAGAAGAGGAGCGCGACCTCCAGC AGGTCACGACCTTCCCTTATCTGATTTGGGGCTTGCTTTTGTGAGTCTGGCTTCTCTGGGTCTGACCTACGCTATGGAGT CAACCGCTCTGGTCTTCCCTATTATGATATTGTTCTGATTGGAGCTAGTCTCTATTTCTCTATTCTGTTGACGGGACGC AAGACCAAGAATTACTTTGCCAATATGCCTGTCAAGACCAGGAACCGTCTAGGGATTGTGCTCTTCTCGAGTGGTCTTGTGCG CGGACGCTTTATCATTATTGGCCGTCTGATGGGCATTGATATGCCGATTTCGGTCTGCTTGGCCCTCTTTGGCATTGGTTGTG GGTGGGATTATTTCCTGATTCTGGGGCTCTAGGGAGTTTGACTTTGTCGTCTGACAGGTCTGTCTGCCACGGTTTGGCT AAGGAAACGGTTCTAGCCTGGCTCTTGTCTTACCGCCTGGCTATTATGTTCTCCCTTTCTTTGTCGGGATTATTTCTTTATCA GACATCTGGGTGGAAGATTAAATGAGAAATCAATGCCGTGCCAAGCAGATGCTGGAAGGGGACATTACAGTCACTATCCG TCTGATTCTGCTGGTTGGGATTCTTAGCCCTTTCCACCGCTCTTTTGAAAAATCTGGATTATACACCTGGCTAAGGCCCTTT AACCCCTTGCAAGCAGATTATTGGCAATTTCCAGGTCTCTTGTGGGATTGTTTCTAGCCAGGACCATCGAAAA AGAAGGTCAAGCTGGCCTATCCTCTGACCATGATTGGTTGGTCTGACTCTGATTATGTGCAATTTTGGAGCTATTAGTTGGGG CTTTTCCATCGCCATTTTCTTAGCTATAATTGCCCTCAATTTGGTCTGACCTCAGCTCTATAAGAAAGCAGTTTATTATTCCTGG GAGCTAGAACCAAGGATAGCCTGATTATTGCTCTGACCTGCTTGTGATTTTCACTTGGCTGGAGTGGTTCCCTGTTCCGG CCCATGAGCTCCATGAACCTGGAGCGGCATAAAACCATCAGTCCAGAAAGCTACCTCTAACCGGGAATAATCGTTTAAATCAT TACGGTGATTATCACGGCTCTTACTGGATTTTGTATGCGGATTGTTCAAGGACGCAAGGTTAAGTTGGGCAACGCCCTTCCAGGCT GACCGCTACCGTAAGCTCTGGAACCTATGGTGGTCAAGGGGATAGCGCACTTGTCTTCTAGGTGACAAACGCCCTCTCTGGT ATCAAAAGGAAGCGAAGACCGGTTGCCCTTCAATTGCTATCCGCAACAACAAATGTCTGGTATGGGCCAACCTGCTGGTAA CCAAGCAGATGTGGAAGAAGCAGATTGCTTATTTCATTGATAAGCCGATCGGCTGGATTAGCAGTCTGATTCTCTATAGTATCGGT CAAAAGATGACCCCTCTACCTGCAATGAATTCGGCTTTGAATTTATGAAGGTGGGGGAAACGCCCTAGTGGATTGGAAAAATTCA CCCTCAGGGGCAATAATACAAGCCTTTCCGCAATGCCCTCAATCGGTGGAGAAGATGCTTATGCTTTGAAGTCTTGGCCAG CCCCATAGTCTGAGCTTTTGGCGGAAGTCAAGGCCGTTTCTGACATCTGGCTGGAGCGGTGATCTGAAAAGGGCTTCTCACTA GGTTATTTTACGAGGCTACTTGGCAAGCTGCTCTGGCAGTCTGTCGAATCAGGAGGCAAACTGGTTGCCCTTTGCCAATG TCATGCTAATTATCAGGAAATATCGCTTCCATTGATGATGCGCTATGACAAGGAGGCCATCCAAACGGGCTCATGGACTA CCTCTTATTTCCCTCTTGTCTATTTCCGAGACCAAGGTATCGGTTATTTGCACTTGGGCATGGCTCCCTATCAGGAGTTGGT CAGGTCGAGGAAAGCTTTATTCAGAGCGTTTGGCCTATCTGGTCTATCATTTTGGCAGTCAITTTTACTCCTTTGAAGGTCTTT ATCGCTATAAGGAAAAATTCACCCCACTTTGGGACGAGCGCTATATTTCTATTCTAGAACTTCTGGGTGGCTATGCTCTTAT GACCTAGTCTTAGAGGACCGAGTGACCAAGTTAA GATAAAGGTATTGGTAAATA AGGAAGAGAGGTCCCCTTTCC AGGAAGAGAGGTCCTTTTCCCTTGAAAACGATTATTGAGAACTGAAG GCTTTTATGCCTATCATTAAGGTCAATTTTTCACCTCAATCGTGGTCTTAATCATTGTTGAGTTAATGCACCTGAAACAAACC ATATCGGTCAAAAGATTGGGGCAGGTGCTGCAAGGATTTCTTGGCTTAATATCTTCTGATTTTTTATGTCGGCAGCGTGGCC GTCTTTCCAAACACAGGATATGATTTCATTGTCAACAGGATTCTAAAAACCAACCAAGCAAGAACTGATATTTCTGAGAGCAAGC TGGTGTATCAATACCTTTAATAATCTGACTGGTTTTGGGGAAATTATGACATTGGCCTGGCAGTTCGCTTTTACGGCAAAAAA GGCGAGGAAGAAAGCAATTGCAGGAATTACCCGTTTCTGCCCTATCTTATTTACGGGCTGTCTTTATCAGCCTTTGTTTCC CTAGGATTGGTTTCGATCTTTCCCGTCACAAAGGAGTGCCGTTTTTACGATATCATCCTTTTGGCATCAACCCCTTTATTCCTCT CTGATTTATGGTTTTTCCGGCAGAAAAACCACTACTACTTTGGCAATATGCCCGCAAAATCCGGCTGCAGCTTGGTCTTGTCT TCCGCTTTGAATGGACTTTGTGGCGCTTGTCTTTATCATCATTGGCTATCTGATGGGAATTGAATTTGCCCTCTATAAAAAG CTACCCCTCTTTAGTATCGCCTGTCTGCAGGGATTGTGCTACTAATCCCTGGCGGGTTGGGAAGTTTGAAGTGGTCTTGTGTT ACCGGCTTGTCTGCCGAAGGCTTGGCCAAAGAAACAGTCTGCTGGCTCTTACTCTATCTGTCTGGCTTATTATTTCTGCTGCA TTCTTTTGGCGGTATTTATTTCTTTTGGCCATCTTGGACATCAGATTAAATCAAAGGTATCAAGGGGCTCTAGGAGGCTCTCT TCAAGCAGTCTACACAGTGTCTATACACGCTGTAGGTGGCTGGGAATCTCACTTGTCTTTTCAACCCCTCTTTTGAAGAA ATCTCTCATCTCAGCTGGCTAGACGCTTTTGGTCCCTTACAAAAACAGCTTATCTGGCAATTTCCAGGGCTTTTAAATCGGGGT TGCTTTATCCTGCTGGCCAGAGCCATTGCTAATGGGGTTAAGAGGGCTTATTCTCTTATGTTTTTGGGTCTGGCAGCAACCCCT GTGTACATTAATACTGGATCTATCAGTTGGCAATGGTCACTCGCTATTGTCTTACTTGGATTGGGAGTGTCTCTATCCGAACA CACCTCTATAAAGCCCAATTTATCTATTCTGGGAAGGACGCAACCAAGATATCATCTTGATTTCCTCCATGCTCTTTATCTTA TTCAACCTTGCCAGTCTGATTTTCCATACCGAGCCTATGAAATGAATCAGCACCTTGGCTTCTCTGTTACTTTTGGAGCTGG AAACATATCCTGCTGGCAGCAGCTTGGTCACCTAGCCTATGCTGTCTGTTAGGTATCTCTACGGTGTAAACCTTTATTG GGTGAGACTTTTGGACGATCGCTACCAAGCGTTTATTGAGCGAATATGGCGGACGCGCAGCAGCGGACTCGCCTTTATAAAA AACAAAGCTCTCTTCTGATATCAGGAGGAGGACTTGACAAGGAGCCTTCCAATTTTCAACTATCAATAACAAAGGCTGTTGTC ATGGGAGAGCCGGCTGGAATCCGGAACCTTTAAAAATGCTACAAGAGCCTTCAATGAAGTGTGATCGGCTCAATATGAC CTCGTCTTTTACAGTATCGGCAAGATATGACCCCTTTTTCATGAATATGGCTTTGAATTCATGAAGTGGGGGAAATGCC CTAGTTGATTTAGATAGCTTCCATCTTAAAGGTAAATAACAAAGCCTTTCCGAAACGCTCTCAACCGGTCTTAAAAAGAGGA TTACCTTTGAAATAAGTCTCAGCCTCACTCGAAAGTCTTATTAGATGAACAGAGACTATTTCAAACAAATGGCTGGAAAAA AGACAGGAAAAAGATTCTCACTCGGTTCTTTGATAGAGACTATTTCGAAGGACCGCTGGCGCTGGTTAAAAATAAGGAA CAGGAAATTTGTCCTTTGGCAATATTATGCCAATCACCAGAACGGTATCGTCTCCATGCACTTGTGCGCTACGATGCCAAA AAAAATCCAAATGGGCTCATGACTACCTCTTCTTATCTCTTTTATCCATTTTAAAGGAAGAGGACTGGCTCTTACCTTTGACCT GGAATGGCACCCCTTGTGGGGGTTGGGCAAGTCAAGGAAAGTTTCTCCAGCAACGGCTATCTGCTCTACTCAATTTCCGA ACCCATTTTATCCCTTGAAGGACTGCAACGATACAAAAAGAAATCAACCCGATTGGGAGGAACGCTATGCTCTGTGTTGCC AAATCCTCTTGGCTGTCTACGCCATCTTGGGCAATTTCTTACTGGATACCAACTGTTTAAACAAAAAGAGATCCTCTTATG GATAAAGGTATTGGTAAATA
<i>S. ferus mprF</i> (Locus tag A3GY_RS0106165)	AGGAAGAGAGGTCCCCTTTCC AGGAAGAGAGGTCCTTTTCCCTTGAAAACGATTATTGAGAACTGAAG GCTTTTATGCCTATCATTAAGGTCAATTTTTCACCTCAATCGTGGTCTTAATCATTGTTGAGTTAATGCACCTGAAACAAACC ATATCGGTCAAAAGATTGGGGCAGGTGCTGCAAGGATTTCTTGGCTTAATATCTTCTGATTTTTTATGTCGGCAGCGTGGCC GTCTTTCCAAACACAGGATATGATTTCATTGTCAACAGGATTCTAAAAACCAACCAAGCAAGAACTGATATTTCTGAGAGCAAGC TGGTGTATCAATACCTTTAATAATCTGACTGGTTTTGGGGAAATTATGACATTGGCCTGGCAGTTCGCTTTTACGGCAAAAAA GGCGAGGAAGAAAGCAATTGCAGGAATTACCCGTTTCTGCCCTATCTTATTTACGGGCTGTCTTTATCAGCCTTTGTTTCC CTAGGATTGGTTTCGATCTTTCCCGTCACAAAGGAGTGCCGTTTTTACGATATCATCCTTTTGGCATCAACCCCTTTATTCCTCT CTGATTTATGGTTTTTCCGGCAGAAAAACCACTACTACTTTGGCAATATGCCCGCAAAATCCGGCTGCAGCTTGGTCTTGTCT TCCGCTTTGAATGGACTTTGTGGCGCTTGTCTTTATCATCATTGGCTATCTGATGGGAATTGAATTTGCCCTCTATAAAAAG CTACCCCTCTTTAGTATCGCCTGTCTGCAGGGATTGTGCTACTAATCCCTGGCGGGTTGGGAAGTTTGAAGTGGTCTTGTGTT ACCGGCTTGTCTGCCGAAGGCTTGGCCAAAGAAACAGTCTGCTGGCTCTTACTCTATCTGTCTGGCTTATTATTTCTGCTGCA TTCTTTTGGCGGTATTTATTTCTTTTGGCCATCTTGGACATCAGATTAAATCAAAGGTATCAAGGGGCTCTAGGAGGCTCTCT TCAAGCAGTCTACACAGTGTCTATACACGCTGTAGGTGGCTGGGAATCTCACTTGTCTTTTCAACCCCTCTTTTGAAGAA ATCTCTCATCTCAGCTGGCTAGACGCTTTTGGTCCCTTACAAAAACAGCTTATCTGGCAATTTCCAGGGCTTTTAAATCGGGGT TGCTTTATCCTGCTGGCCAGAGCCATTGCTAATGGGGTTAAGAGGGCTTATTCTCTTATGTTTTTGGGTCTGGCAGCAACCCCT GTGTACATTAATACTGGATCTATCAGTTGGCAATGGTCACTCGCTATTGTCTTACTTGGATTGGGAGTGTCTCTATCCGAACA CACCTCTATAAAGCCCAATTTATCTATTCTGGGAAGGACGCAACCAAGATATCATCTTGATTTCCTCCATGCTCTTTATCTTA TTCAACCTTGCCAGTCTGATTTTCCATACCGAGCCTATGAAATGAATCAGCACCTTGGCTTCTCTGTTACTTTTGGAGCTGG AAACATATCCTGCTGGCAGCAGCTTGGTCACCTAGCCTATGCTGTCTGTTAGGTATCTCTACGGTGTAAACCTTTATTG GGTGAGACTTTTGGACGATCGCTACCAAGCGTTTATTGAGCGAATATGGCGGACGCGCAGCAGCGGACTCGCCTTTATAAAA AACAAAGCTCTCTTCTGATATCAGGAGGAGGACTTGACAAGGAGCCTTCCAATTTTCAACTATCAATAACAAAGGCTGTTGTC ATGGGAGAGCCGGCTGGAATCCGGAACCTTTAAAAATGCTACAAGAGCCTTCAATGAAGTGTGATCGGCTCAATATGAC CTCGTCTTTTACAGTATCGGCAAGATATGACCCCTTTTTCATGAATATGGCTTTGAATTCATGAAGTGGGGGAAATGCC CTAGTTGATTTAGATAGCTTCCATCTTAAAGGTAAATAACAAAGCCTTTCCGAAACGCTCTCAACCGGTCTTAAAAAGAGGA TTACCTTTGAAATAAGTCTCAGCCTCACTCGAAAGTCTTATTAGATGAACAGAGACTATTTCAAACAAATGGCTGGAAAAA AGACAGGAAAAAGATTCTCACTCGGTTCTTTGATAGAGACTATTTCGAAGGACCGCTGGCGCTGGTTAAAAATAAGGAA CAGGAAATTTGTCCTTTGGCAATATTATGCCAATCACCAGAACGGTATCGTCTCCATGCACTTGTGCGCTACGATGCCAAA AAAAATCCAAATGGGCTCATGACTACCTCTTCTTATCTCTTTTATCCATTTTAAAGGAAGAGGACTGGCTCTTACCTTTGACCT GGAATGGCACCCCTTGTGGGGGTTGGGCAAGTCAAGGAAAGTTTCTCCAGCAACGGCTATCTGCTCTACTCAATTTCCGA ACCCATTTTATCCCTTGAAGGACTGCAACGATACAAAAAGAAATCAACCCGATTGGGAGGAACGCTATGCTCTGTGTTGCC AAATCCTCTTGGCTGTCTACGCCATCTTGGGCAATTTCTTACTGGATACCAACTGTTTAAACAAAAAGAGATCCTCTTATG GATAAAGGTATTGGTAAATA

Table S5

mprF sequences synthesized by Genewiz.

Red indicates sequence complementarity pABG5.

References

- [1] Fahy Eoin, Cotter Dawn, Sud Manish, Subramaniam Shankar (2011) **Lipid classification, structures and tools** *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* **1811**:637–647
- [2] Wang Lang-Hong, Zeng Xin-An, Wang Man-Sheng, Brennan Charles S, Gong Deming (2018) **Modification of membrane properties and fatty acids biosynthesis-related genes in escherichia coli and staphylococcus aureus: Implications for the antibacterial mechanism of naringenin** *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1860**:481–490
- [3] Roy Hervé, Dare Kiley, Ibba Michael (2009) **Adaptation of the bacterial membrane to changing environments using aminoacylated phospholipids** *Molecular microbiology* **71**:547–550
- [4] Roy Hervé, Ibba Michael (2008) **Rna-dependent lipid remodeling by bacterial multiple peptide resistance factors** *Proceedings of the National Academy of Sciences* **105**:4667–4672
- [5] Sohlenkamp Christian, Geiger Otto (2016) **Bacterial membrane lipids: diversity in structures and pathways** *FEMS microbiology reviews* **40**:133–159
- [6] Weidenmaier Christopher, Peschel Andreas, Kempf Volkhard AJ, Lucindo Natalie, Yeaman Michael R, Bayer Arnold S (2005) **Dltabcd-and mprf-mediated cell envelope modifications of staphylococcus aureus confer resistance to platelet microbicidal proteins and contribute to virulence in a rabbit endocarditis model** *Infection and immunity* **73**:8033–8038
- [7] Joyce Luke R, Doran Kelly S (2023) **Gram-positive bacterial membrane lipids at the host-pathogen interface** *PLoS Pathogens* **19**
- [8] Gujrati Vipul, Kim Sunghyun, Kim Sang-Hyun, Min Jung Joon, Choy Hyon E, Kim Sun Chang, Jon Sangyong (2014) **Bioengineered bacterial outer membrane vesicles as cell-specific drug-delivery vehicles for cancer therapy** *ACS nano* **8**:1525–1537
- [9] Ernst Christoph M., Staubitz Petra, Mishra Nagendra N., Yang Soo-Jin, Hornig Gabriele, Kalbacher Hubert, Bayer Arnold S., Kraus Dirk, Peschel Andreas (2009) **The bacterial defensin resistance protein mprf consists of separable domains for lipid lysinylation and antimicrobial peptide repulsion** *PLOS Pathogens* **5**:1–9
- [10] Staubitz Petra, Neumann Heinz, Schneider Tanja, Wiedemann Imke, Peschel Andreas (2004) **Mprf-mediated biosynthesis of lysylphosphatidylglycerol, an important determinant in staphylococcal defensin resistance** *FEMS microbiology letters* **231**:67–71
- [11] Joyce Luke R., Manzer Jéssica da Haider S., Mendonça C., Villarreal Ricardo, Nagao Prescilla E., Doran Kelly S., Palmer Kelli L., Guan Ziqiang (2022) **Identification of a novel cationic glycolipid in streptococcus agalactiae that contributes to brain entry and meningitis** *PLOS Biology* **20**:1–15
- [12] Joyce Luke R, Guan Ziqiang, Palmer Kelli L (2021) **Streptococcus pneumoniae, s. pyogenes and s. agalactiae membrane phospholipid remodelling in response to human serum** *Microbiology* **167**

- [13] Camacho Christiam, Coulouris George, Avagyan Vahram, Ma Ning, Papadopoulos Jason, Bealer Kevin, Madden Thomas L. (2009) **BLAST+: architecture and applications** *BMC Bioinformatics* **10**
- [14] Tubiana Jérôme, Cocco Simona, Monasson Rémi (2019) **Learning protein constitutive motifs from sequence data** *eLife* **8**
- [15] Adams Hannah M, Joyce Luke R, Guan Ziqiang, Akins Ronda L, Palmer Kelli L (2017) **Streptococcus mitis and s. oralis lack a requirement for cdsa, the enzyme required for synthesis of major membrane phospholipids in bacteria** *Antimicrobial agents and chemotherapy* **61**:e02552–16
- [16] Joyce Luke R, Guan Ziqiang, Palmer Kelli L (2019) **Phosphatidylcholine biosynthesis in mitis group streptococci via host metabolite scavenging** *Journal of bacteriology* **201**:e00495–19
- [17] Smolensky P. (1986) **Information processing in dynamical systems: foundations of harmony theory** :194–281
- [18] Morcos Faruck, Pagnani Andrea, Lunt Bryan, Bertolino Arianna, Marks Debora S., Sander Chris, Zecchina José Riccardo, Onuchic N., Hwa Terence (2011) **Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families** *Proceedings of the National Academy of Sciences* **108**:E1293–E1301
- [19] Cheng Ryan R., Morcos Faruck, Levine Herbert, Onuchic N. (2014) **Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information** *Proceedings of the National Academy of Sciences* **111**:E563–E571
- [20] Jiang Xian-Li, Dimas Rey P., Chan Clement T. Y., Morcos Faruck (2021) **Coevolutionary methods enable robust design of modular repressors by reestablishing intra-protein interactions** *Nature Communications* **12**
- [21] Goodey Nina M., Benkovic Stephen J. (2008) **Allosteric regulation and catalysis emerge via a common route** *Nature Chemical Biology* **4**:474–482
- [22] Figliuzzi Matteo, Barrat-Charlaix Pierre, Weigt Martin (2018) **How pairwise coevolutionary models capture the collective residue variability in proteins?** *Molecular Biology and Evolution* **35**:1018–1027
- [23] Tubiana Jérôme, Cocco Simona, Monasson Rémi (2019) **Learning Compositional Representations of Interacting Systems with Restricted Boltzmann Machines: Comparative Study of Lattice Proteins** *Neural Computation* **31**:1671–1717
- [24] Slavetinsky Christoph J, Peschel Andreas, Ernst Christoph M (2012) **Alanyl-phosphatidylglycerol and lysyl-phosphatidylglycerol are translocated by the same mprf flippases and have similar capacities to protect against the antibiotic daptomycin in staphylococcus aureus** *Antimicrobial agents and chemotherapy* **56**:3492–3497
- [25] Mistry Jaina *et al.* (2021) **Pfam: The protein families database in 2021** *Nucleic acids research* **49**:D412–D419
- [26] Hebecker Stefanie, Krausze Joern, Hasenkampf Tatjana, Schneider Julia, Groenewold Maike, Reichelt Joachim, Jahn Dieter, Heinz Dirk W., Moser Jürgen (2015) **Structures of two bacterial resistance factors mediating tRNA-dependent aminoacylation of phosphatidylglycerol with lysine or alanine** *Proceedings of the National Academy of Sciences* **112**:10691–10696

- [27] Mundy L. M., Sahm D. F., Gilmore M. (2000) **Relationships between enterococcal virulence and antimicrobial resistance** *Clinical Microbiology Reviews* **13**:513–522
- [28] Bao Yinyin, Sakinc Tuerkan, Laverde Diana, Wobser Dominique, Benachour Abdellah, Theilacker Christian, Hartke Axel, Huebner Johannes (2012) **Role of mprf1 and mprf2 in the pathogenicity of enterococcus faecalis** *PLoS One* **7**
- [29] Roy Hervé, Ibba Michael (2009) **Broad range amino acid specificity of rna-dependent lipid remodeling by multiple peptide resistance factors** *Journal of Biological Chemistry* **284**:29677–29683
- [30] Dale Jennifer L *et al.* (2018) **Comprehensive functional analysis of the enterococcus faecalis core genome using an ordered, sequence-defined collection of insertional mutations in strain og1rf** *Msystems* **3**:e00062–18
- [31] Peschel Andreas *et al.* (2001) **Staphylococcus aureus resistance to human defensins and evasion of neutrophil killing via the novel virulence factor mprf is based on modification of membrane lipids with l-lysine** *The Journal of experimental medicine* **193**:1067–1076
- [32] Schneider Thomas D, Stephens R Michael (1990) **Sequence logos: a new way to display consensus sequences** *Nucleic acids research* **18**:6097–6100
- [33] Arnold Frances H. (2018) **Directed Evolution: Bringing New Chemistry to Life** *Angewandte Chemie (Inter-national Ed. in English)* **57**:4143–4148
- [34] Carrasco Manuel J. *et al.* (2021) **Ionization and structural properties of mRNA lipid nanoparticles influence expression in intramuscular and intravascular administration** *Communications Biology* **4**:1–15
- [35] Kilian Mogens, Mikkelsen Lena, Henrichsen Jørgen (1989) **Taxonomic study of viridans streptococci: description of streptococcus gordonii sp. nov. and emended descriptions of streptococcus sanguis (white and niven 1946), streptococcus oralis (bridge and sneath 1982), and streptococcus mitis (andrewes and horder 1906)** *International Journal of Systematic Bacteriology* **39**:471–484
- [36] Taylor Robin G, Walker David C, McInnes RR (1993) **E. coli host strains significantly affect the quality of small scale plasmid dna preparations used for sequencing** *Nucleic acids research* **21**
- [37] F Kunst N Ogasawara, Moszer I, Albertini AM, Alloni GO, Azevedo Vasco, Bertero MG, Bessieres Philippe, Bolotin A, Borchert Sea, *et al.* (1997) **The complete genome sequence of the gram-positive bacterium bacillus subtilis** *Nature* **390**:249–256
- [38] Kreiswirth Barry N, Löfdahl Sven, Betley Marsha J, O'reilly Mary, Schlievert Patrick M, Bergdoll Merlin S, Novick Richard P (1983) **The toxic shock syndrome exotoxin structural gene is not detectably transmitted by a prophage** *Nature* **305**:709–712
- [39] Kuypers JANE M, Heggen LAURA M, Rubens CRAIG E (1989) **Molecular analysis of a region of the group b streptococcus chromosome involved in type iii capsule expression** *Infection and immunity* **57**:3058–3065

- [40] Faralla Cristina, Metruccio Matteo M, De Chiara Matteo, Mu Rong, Patras Kathryn A, Muzzi Alessandro, Grandi Guido, Margarit Immaculada, Doran Kelly S, Janulczyk Robert (2014) **Analysis of two-component systems in group b streptococcus shows that rgfac and the novel fspsr modulate virulence and bacterial fitness** *MBio* **5**:e00870–14
- [41] Spencer Brady L, Chatterjee Anushila, Duerkop Breck A, Baker Carol J, Doran Kelly S (2021) **Complete genome sequence of neonatal clinical group b streptococcal isolate cjb111** *Microbiology resource announcements* **10**:e01268–20
- [42] Palmer Kelli L *et al.* (2010) **High-quality draft genome sequences of 28 enterococcus sp. isolates** *Journal of bacteriology* **192**:2469–2470
- [43] McBride Shonna M, Fischetti Vincent A, LeBlanc Donald J, Jr Robert C Moellering, Gilmore Michael S (2007) **Genetic diversity among enterococcus faecalis** *PloS one* **2**
- [44] Gold Olga G., Jordan H.V., van Houte J. (1975) **The prevalence of enterococci in the human mouth and their pathogenicity in animal models** *Archives of Oral Biology* **20**:473–IN15
- [45] Sharon Belle M, Hulyalkar Neha V, Zimmermann Philippe E, Palmer Kelli L, De Nisco Nicole J (2023) **Inter-species diversity and functional genomic analyses of closed genome assemblies of clinically isolated, megaplasmid-containing enterococcus raffinosus er676 and atcc49464** *Access Microbiology*
- [46] Collins MD, Rodrigues UM, Pigott NE, Facklam RR (1991) **Enterococcus dispar sp. nov. a new enterococcus species from human sources** *Letters in Applied Microbiology* **12**:95–98
- [47] Drucker D. B. (1979) **3: Sweetening agents in food, drinks and medicine: Cariogenic potential and adverse effects** *International journal of food sciences and nutrition* **33**:114–124
- [48] Collins Matthew D, Phillips Brian A, Zannoni Paolo (1989) **Deoxyribonucleic acid homology studies of lactobacillus casei, lactobacillus paracasei sp. nov., subsp. paracasei and subsp. tolerans, and lactobacillus rhamnosus sp. nov., comb. nov** *International Journal of Systematic and Evolutionary Microbiology* **39**:105–108
- [49] Hansen PA, Lessel Erwin F (1971) **Lactobacillus casei (orla-jensen) comb. nov** *International Journal of Systematic and Evolutionary Microbiology* **21**:69–71
- [50] Rogosa Morrison, Hansen P Arne (1971) **Nomenclatural considerations of certain species of lactobacillus beijerinck: Request for an opinion** *International Journal of Systematic and Evolutionary Microbiology* **21**:177–186
- [51] Chester Frederick Dixon (1901) **A manual of determinative bacteriology** *Macmillan*
- [52] Adams Hannah M, Li Xiang, Mascio Carmela, Chesnel Laurent, Palmer Kelli L (2015) **Mutations associated with reduced surotomycin susceptibility in clostridium difficile and enterococcus species** *Antimi-crobial Agents and Chemotherapy* **59**:4139–4147
- [53] Manson Janet M, Keis Stefanie, Smith John MB, Cook Gregory M (2003) **A clonal lineage of vanatype enterococcus faecalis predominates in vancomycin-resistant enterococci isolated in new zealand** *Antimicrobial agents and chemotherapy* **47**:204–210
- [54] Salvadori Gabriela, Junges Roger, Morrison Donald A, Petersen Fernanda C (2016) **Overcoming the barrier of low efficiency during genetic transformation of streptococcus mitis** *Frontiers in microbiology* **7**

- [55] Shepard Brett D, Gilmore Michael S (1995) **Electroporation and efficient transformation of enterococcus faecalis grown in high concentrations of glycine** *Electroporation Protocols for Microorganisms* :217–226

Article and author information

Priya M. Christensen

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
ORCID iD: [0000-0003-4790-9839](https://orcid.org/0000-0003-4790-9839)

Jonathan Martin

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
ORCID iD: [0000-0003-0946-3864](https://orcid.org/0000-0003-0946-3864)

Aparna Uppuluri

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
ORCID iD: [0009-0005-6334-8375](https://orcid.org/0009-0005-6334-8375)

Luke R. Joyce

Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
ORCID iD: [0000-0002-9346-671X](https://orcid.org/0000-0002-9346-671X)

Yahan Wei

School of Podiatric Medicine, University of Texas Rio Grande Valley, Harlingen, TX 78550, USA
ORCID iD: [0000-0002-6372-7237](https://orcid.org/0000-0002-6372-7237)

Ziqiang Guan

Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA
For correspondence: ziqiang.guan@duke.edu
ORCID iD: [0000-0002-8082-3423](https://orcid.org/0000-0002-8082-3423)

Faruck Morcos

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA,
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, USA, Center
for Systems Biology, University of Texas at Dallas, Richardson, TX 75080, USA
For correspondence: faruckm@utdallas.edu
ORCID iD: [0000-0001-6208-1561](https://orcid.org/0000-0001-6208-1561)

Kelli L. Palmer

Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
For correspondence: kelli.palmer@utdallas.edu
ORCID iD: [0000-0002-7343-9271](https://orcid.org/0000-0002-7343-9271)

Copyright

© 2024, Christensen et al.

This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and

source are credited.

Editors

Reviewing Editor

Anne-Florence Bitbol

Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

Senior Editor

Qiang Cui

Boston University, Boston, United States of America

Reviewer #1 (Public Review):

The basic approach is that the authors first train an RBM on all MprF sequences, and then use this analysis to identify a subset of the family that catalyzes the addition of amino acids to PG. Then a second RBM is trained on this subset.

In the initial RBM training a particular hidden unit is identified that has a sparse and bimodal activation in response to the input sequences. The contribution of individual residues is shown in Figure 3c, which highlights one of the strengths of this RBM implementation - it is interpretable in a physically meaningful way. However, there are several decisions here, the justification of which is not entirely clear.

- i) Some of the residues in Fig 3c are stated as "relevant" for aminoacylated PG production. But is this the only such hidden unit? Or are there others that are sparse, bimodal, and involve "relevant" AA?
- ii) In order to filter the sequences for the second stage, only those that produce an activation over +2.0 in this particular hidden unit were taken. How was this choice made?
- iii) How many sequences are in the set before and after this filtering? On the basis of the strength of the results that follow I expect that there are good reasons for these choices, but they should be more carefully discussed.
- iv) Do the authors think that this gets all of the aminoacylated PG enzymes? Or are some missed?

The authors show that they can classify members of the family by training a second RBM on the filtered sequences. They do this by identifying two hidden unit activations in particular (Figure 5b) which seem to be useful for determining lipid substrate specificity, and they test several variants that obtain different responses of these two hidden units by experimentally determining what lipids they produce (Table 2). However, some similar criticisms from the last point occur here as well, namely the selection of which weights should be used to classify the enzymes' function. Again the approach is to identify hidden unit activations that are sparse (with respect to the input sequence), have a high overall magnitude, and "involve residues which could be plausibly linked to the lipid binding specificity."

- i) Two hidden units are identified as useful for classification, but how many candidates are there that pass the first two criteria? Indeed, how many hidden units are there?
- ii) The criterion "involve residues which could be plausibly linked to the lipid binding specificity" is again vague. Do all of the other candidate hidden units *not* involve significant contributions from substrate-binding residues? Maybe one of the other units does a better job of discriminating substrate specificity. (As indicated in Figure 8, there are examples of enzymes that confound the proposed classification.) Why combine the activations of two units for the classification, instead of 1 or 3 or...?

<https://doi.org/10.7554/eLife.94929.1.sa2>

Reviewer #2 (Public Review):

In "Lipid discovery enabled by sequence statistics and machine learning" Christensen et al. address an important question: how can bacteria modify lipid charges to produce cationic lipids, prone to confer resistance to cationic antibiotics? One of the enzymes involved in this process is MprF, which can, through the transfer of amino acids, in particular, lysine, from charged tRNA modify the charge of anionic membrane phospholipid from negative to positive. Recent works have shown that MprF can also modify another substrate, glycolipid glucosyl-diacylglycerol, which is neutral. These findings immediately raise two questions: what are the determinants in the MprF sequence controlling the lipid substrates it can modify? Are there other substrates for MprF, so far unknown?

Christensen et al. address both of these questions in an elegant way, combining sequence analysis with machine-learning methods and experimental characterisation of the enzymatic products through mass spectrometry. Using restricted Boltzmann machines (RBM), an unsupervised architecture extracting statistical features from the sequence data, they identify putative amino-acid motifs along the MprF sequences possibly related to the substrate identity, select some bacterial species whose wild-type sequence contains those motifs, and validate the biological role of the motifs by identifying the produced lipids. Remarkably, with this approach, the authors find a novel cationic lipid with two glucosyl groups.

Besides these new results on MprF and its operation, the present work is appealing, as it shows that the functional characterisation of a very small number of proteins (here, three!) combined with the guided classification of homologous sequence data with appropriate machine-learning methods can lead to the discovery of new functionalities.

<https://doi.org/10.7554/eLife.94929.1.sa1>

Reviewer #3 (Public Review):**Summary:**

After the previous identification that the *Streptococcus agalactiae* MprF enzyme can synthesize also lysyl-glucosyl-diacylglycerol (Lys-Glc-DAG), besides the already known lysyl-phosphatidylglycerol (Lys-PG), the authors aim for the current manuscript was to investigate the molecular determinants of MprF lipid substrate specificity in a variety of bacterial species.

Strengths:

- In general, the manuscript is well constructed and easy to follow, especially taking into account the multidisciplinary aspect of it (computational machine learning combined with lipid biology).
- The added value of the Restricted Boltzmann machines (RBM) approach, in comparison to standard computational pairwise sequence statistics, becomes evident. This is exemplified by a successful, although not perfect, classification and categorization of MprF activity.
- The MS analysis (monoisotopic mass, plus fragmentation pattern), convincingly shows the identification of a novel lipid species Lys-Glc2-DAG.

Weaknesses:

- In many of the analyzed strains, the presence of the lipid species Lys-PG, Lys-Glc-DAG, and Lys-Glc2-DAG is correlated to the presence of the MprF enzyme(s), but one should keep in mind that a multitude of other membrane proteins are present that in theory could be involved in the synthesis as well. Therefore, there is no direct evidence that the MprF enzymes are linked to the synthesis of these lipid species. Although, it is unlikely that other

enzymes are involved, this weakens the connection between the observed lipids and the type of MprF.

-Related to this, in a few cases MprF activity is tested, but the manuscript does not contain any information on protein expression levels. Heterologous expression of membrane proteins is in general challenging and due to various reasons, proteins end up not being expressed at all. As an example, the absence of activity for the *E. faecalis* MprF1 and *E. faecium* MprF2 could very well be explained by the entire absence of the protein.

Overall, the authors largely achieved their goals, as the applied RBM approach led to specific sequence determinants in MprF enzymes that could categorize the specificity of these enzymes. The experimental data could largely confirm this categorization, although a stronger connection between synthesized lipids and enzyme activity would have further strengthened the observations.

The work now focuses only on MprF enzymes, but could in theory be expanded to other categories of lipid-synthesizing enzymes. In other words, the RBM approach could have an impact on the lipid synthesis field, if it would be a tool that is easily applicable. Moreover, the lipids synthesized by MprF (Lys-PG, but also other cationic lipids) play an important role in bacterial resistance against certain antibiotics.

<https://doi.org/10.7554/eLife.94929.1.sa0>