environmental microbiology

Environmental Microbiology (2019) **21**(8), 2921–2932 doi:10.1111/1462-2920.14685

Cyclo(tetrahydroxybutyrate) production is sufficient to distinguish between *Xenorhabdus* and *Photorhabdus* isolates in Thailand

Nicholas J. Tobias,[1,2*†] César Parra-Rojas,[3†] Yan-Ni Shi,[1] Yi-Ming Shi,[1] Svenja Simonyi,[1] Aunchalee Thanwisai,[4] Apichat Vitta,[4] Narisara Chantratita,[5] Esteban A. Hernandez-Vargas[3*] and Helge B. Bode[1,2,6*]

[1]*Molekulare Biotechnologie, Goethe-Universität Frankfurt, Frankfurt am Main, Germany.*

[2]*LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), 60325, Frankfurt am Main, Germany.*

[3]*Frankfurt Institute for Advanced Studies, Ruth-Moufang-Straße 1, 60438, Frankfurt am Main, Germany.*

[4]*Department of Microbiology and Parasitology, Faculty of Medical Science, Naresuan University, Phitsanulok, 65000, Thailand.*

[5]*Department of Microbiology and Immunology, Faculty of Tropical Medicine, Mahidol University, Bangkok, 10400, Thailand.*

[6]*Buchmann Institute for Molecular Life Sciences, Goethe-Universität Frankfurt, 60438, Frankfurt am Main, Germany.*

Summary

Bacteria of the genera *Photorhabdus* and *Xenorhabdus* produce a plethora of natural products to support their similar symbiotic life cycles. For many of these compounds, the specific bioactivities are unknown. One common challenge in natural product research when trying to prioritize research efforts is the rediscovery of identical (or highly similar) compounds from different strains. Linking genome sequence to metabolite production can help in overcoming this problem. However, sequences are typically not available for entire collections of organisms. Here, we perform a comprehensive metabolic screening using HPLC-MS data associated with a 114-strain collection (58 *Photorhabdus* and 56 *Xenorhabdus*) across Thailand and explore the metabolic variation among the strains, matched with several abiotic factors. We utilize machine learning in order to rank the importance of individual metabolites in determining all given metadata. With this approach, we were able to prioritize metabolites in the context of natural product investigations, leading to the identification of previously unknown compounds. The top three highest ranking features were associated with *Xenorhabdus* and attributed to the same chemical entity, cyclo(tetrahydroxybutyrate). This work also addresses the need for prioritization in high-throughput metabolomic studies and demonstrates the viability of such an approach in future research.

Introduction

Photorhabdus and *Xenorhabdus* are soil dwelling bacteria that are found worldwide in association with nematodes of the genera *Heterorhabditis* and *Steinernema* respectively (Forst *et al*., 1997; Stock *et al*., 2001). The bacteria live in symbiosis with their cognate nematode species and their life cycle involves a pathogenic stage towards invertebrate insects (Han and Ehlers, 2000). Although members of different genera, *Xenorhabdus* and *Photorhabdus* produce a number of shared specialized metabolites (SMs) and occupy very similar ecological niches (Tobias *et al*., 2017). Interestingly, the bacteria have yet to be isolated from the environment as free-living organisms, but instead are always found in association with their respective nematodes. Despite this specificity towards a nematode host, bacteria–nematode pairs may be isolated from the same geographic location.

Recently, we highlighted the extensive chemical diversity present in these genera using high-throughput genomic and metabolomic analyses. It appears that SMs make up a major part of those coding sequences that were acquired and maintained in the genera upon divergence from a common ancestor, namely, members of the Enterobacteriaceae. We proposed that SMs, specifically products of polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs), may be related to the given ecological niche that each strain occupies (Tobias *et al*., 2017). The products of

Received 3 April, 2019; revised 15 May, 2019; accepted 16 May, 2019. *For correspondence. E-mail tobias@bio.uni-frankfurt.de; Tel. (+49) 69 798 29561. E-mail vargas@fias.uni-frankfurt.de; Tel. (+49) 69 798 47533. E-mail h.bode@bio.uni-frankfurt.de; Tel. (+49) 69 79829557. †These authors are the co-first authors

these enzymes in *Photorhabdus* and *Xenorhabdus* have a range of known functions including antibiotic, signalling and assisting in the development of the nematode host, among others (for recent reviews of all known natural products from these genera, see Shi and Bode, 2018; Tobias *et al*., 2018a).

One argument supporting an ecological function for the SMs is the fact that although a few compounds appeared at first to be genus-specific, continued investigations have identified the same clusters in the other genus. Several clear examples of this are xenocoumacin, whose gene cluster was recently found in *Photorhabdus luminescens* PB45.5 (Tobias *et al*., 2016) and xenorhabdin, whose gene cluster has been found in *Photorhabdus asymbiotica* strains (Wilkinson *et al*., 2009). Natural product research is continually encountering the problem of the best way to prioritize research efforts relating to 'new' metabolites. One common way to do this is to find 'new' genera or species that often produce a new subset of SMs (Hoffmann *et al*., 2018). Using their genomic information to identify biosynthetic gene clusters that often produce bioactive compounds, such as PKSs or NRPSs, and subsequently activating 'silent' clusters to specifically stimulate production of the metabolite is a common approach. However, in the absence of genetic information, this becomes increasingly difficult. Tools such as Global Natural Product Molecular Networking Social (GNPS; Wang *et al*., 2016), Sirius (Böcker *et al*., 2009), MZmine (Katajamaa *et al*., 2006; Pluskal *et al*., 2010), DEREPLICATOR+(Mohimani *et al*., 2018) and others have recently been developed for dereplication of MS/MS data. These have also been linked to several databases, which can assist in quickly identifying compounds absent in these databases. However, prioritizing the continued research and development of these unexplored metabolites is still a major problem.

Here, we describe the use of a machine learning model in order to explore the metabolomes of geographically distinct strains of *Photorhabdus* and *Xenorhabdus* from different regions in Thailand. We explored metabolic potential in relation to the environment in which they were collected, identified known compounds and prioritized the structure elucidation of one of the metabolites whose presence was most determining in distinguishing *Xenorhabdus* from *Photorhabdus.* Despite a number of long-standing hypotheses suggesting that metabolite production is specific to each strain (and its respective environment), this is the first time it has been empirically tested.

Results

Strain collection and processing

Strains selected for this study were collected from a variety of areas across central Thailand (Fig. 1, Supporting Information Table S1). Following isolation of the bacteria, each species was identified by sequencing and alignment

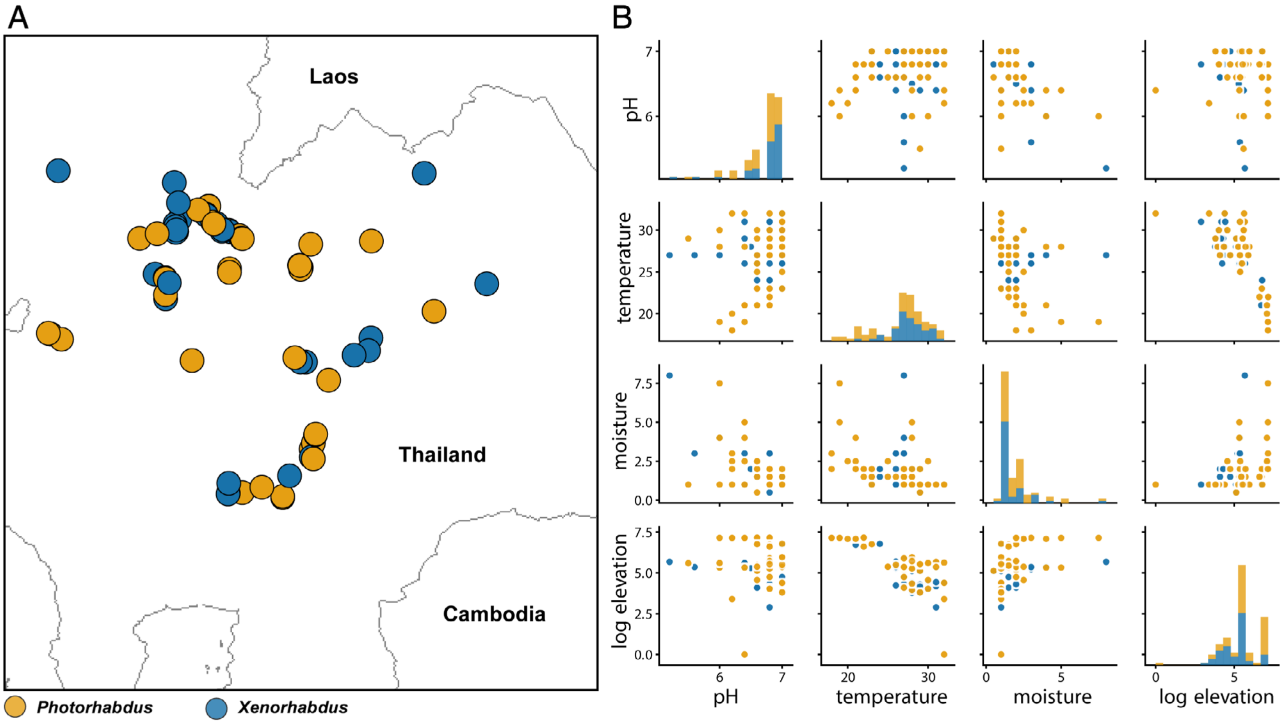

Fig. 1. (A) Location and (B) spread of metadata associated with the 114 *Photorhabdus* and *Xenorhabdus* strains collected from Thailand. For specific metadata values, see Supporting Information Table S1.

of the *recA* coding sequence to the NCBI database (see Supporting Information Table S1 for NCBI accession numbers). Our aim was to explore as big a metabolite repertoire as possible. We therefore cultivated in two different media; Lysogeny broth (LB; nutrient rich) and SF900 (an insect-like medium); extracted each culture independently and combined the final results. Methanol was used to extract the cultures directly in equal volumes, which provided a robust data set on which to perform further analyses. Acetonitrile blanks and media only were used to subtract background masses while *E. coli* (a close relative of *Xenorhabdus* and *Photorhabdus*) was additionally used in order to determine metabolites that were not likely specific to the *Xenorhabdus* and *Photorhabdus.* The combined analysis identified a total of 44,836 molecular features after removing background features (LB, SF900, acetonitrile and *E. coli* in both media). MS data sets can be found under public MassIVE ID: MSV000083378 and the combined network analysis can be downloaded at http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=02057a6b9eb54048847c9dd18746aac9.

Network analysis

Network analysis was performed on the complete collection of strains using GNPS (Wang *et al*., 2016) and Cytoscape (Shannon *et al*., 2003) for visualization (Fig. 2). *Xenorhabdus* had a greater number of unique molecular features (3265) than the *Photorhabdus* (1791). A total of 261 networks with three or more nodes were formed (Fig. 2). Of these, 14 families of compounds could be identified based on previously published studies, leaving a majority of networks still completely unexplored. Use of GNPS and its resulting network analyses revealed a number of networks containing known compounds. These networks group metabolites with structural similarity based on their fragmentation patterns (Wang *et al*., 2016). We assume that all nodes within a given network belong to the same metabolite family. We have shown in *Photorhabdus* and *Xenorhabdus* that this is indeed often the case, as described in our previous work (Tobias *et al*., 2017). Despite providing a broader perspective on the presence and absence of metabolite families, what this fails to address is whether or not these nodes and/or metabolites are important in defining any variables that may be interesting for further investigation.

We have discussed at length the possibility for analogous functions by different *Photorhabdus*- or *Xenorhabdus*-specific compounds (Tobias *et al*., 2017), which would help explain the reasons they live such a similar lifestyle. However, as is clear from Fig. 2, there is still a significant number of metabolite clusters yet to be explored. This begs the question as to where we should focus our research efforts in looking for unknown and important compounds, with respect to both the bacterial ecology and natural product discovery. We therefore decided to utilize machine learning in order to prioritize compounds and their investigations, with an end-goal of researching metabolites that are likely to be both undiscovered and specific.

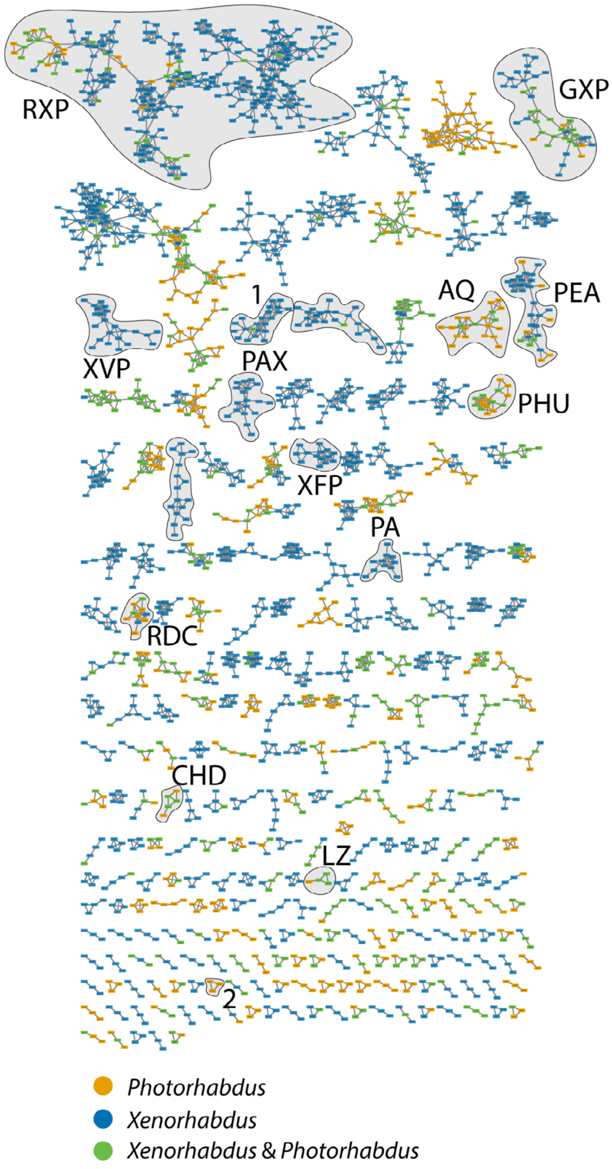

Fig. 2. Network analysis of all 114 isolates. Shown is a summary of all nodes with at least two connections in *Photorhabdus* and *Xenorhabdus.* Known subnetworks are also highlighted: RXP, rhabdopeptide; GXP, GameXPeptide; XVP, xentrivalpeptide; PAX, PAX peptide; AQ, anthraquinone; PEA, phenylethylamide; XFP, xefoampeptide; CHD, cyclohexanedione; LZ, luminizone; RDC, rhabduscin; PA, pyrrolizidine alkaloids; XBN, xenobactins; XMT, xenematide/xenoprotide; **1,** (cyclo)tetrahydroxybutyrate; **2,** network containing signal with *m/z* of 487.18. For a closer view of the network containing **1**, see Supporting Information Fig. S3.

Machine learning to explain metadata

Our data consisted of a total of 114 different strains, coupled to seven abiotic metadata points; two media

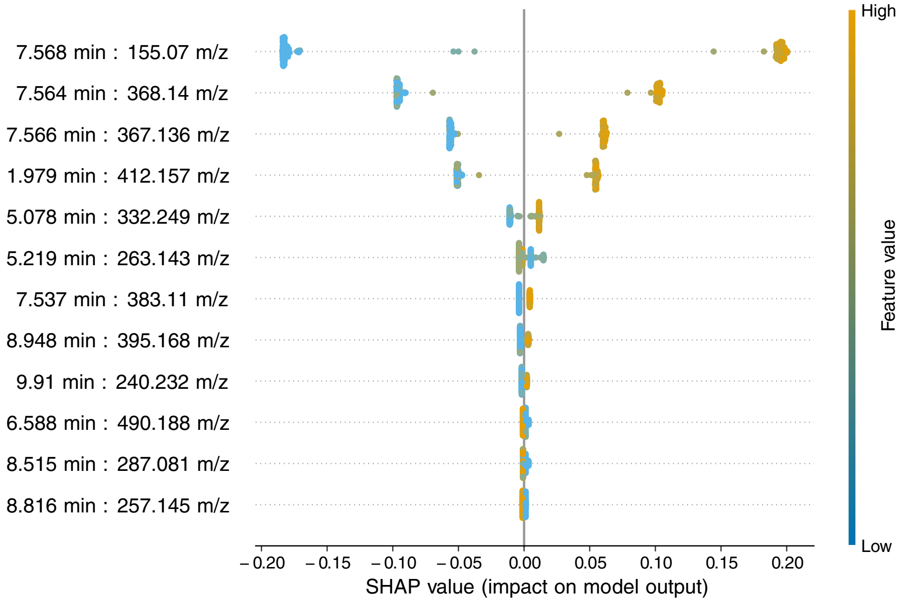

Fig. 3. SHAP output of the GBDT model constructed using intensity values. The value represents the impact of a given feature in determining whether an isolate is *Photorhabdus* or *Xenorhabdus.* The *m/z* ratios and retention times are indicated for the top 10 ranking features.

conditions, four soil types, 10 provinces representing rough geographic relatedness, soil pH, soil temperature, soil moisture and elevation above sea level. In order to explore our data in more detail and determine what, if any, of these abiotic factors could be distinguished by utilizing metabolite production, we turned to machine learning. We utilized a gradient boosting decision tree (GBDT) algorithm in order to train the model on the full data set, as well as a reduced data set consisting of highly correlated signals (see Methods section).

Training the model on the full versus the pruned and clustered datasets (Supporting Information Fig. S1) results in essentially the same performance (Supporting Information Table S2). An initial analysis failed to show any significant impact of the abiotic data on metabolite production. Additionally, both randomizing and removing geographical metadata from the data set did not result in a performance drop. We incorporated SHapley Additive exPlanations (SHAP) values into our model in order to determine the importance of individual features on model output. For both AUC (area under the curve) and intensity data sets with low levels of clustering, we see that a small number of metabolites strongly affect the output of all samples, and seem to do so in a well-delimited fashion (Fig. 3). The impact of a few others is not as strong, but retain the latter property.

Structure elucidation of top-ranking feature(s)

Multiple metabolites seemed to be independently capable of discerning between genera with a high degree of accuracy. In particular, the top three single-feature predictors possessed the same retention times with *m/z* of 155.07, 368.14 and 367.13 respectively (Supporting Information Fig. S2). All three of these metabolites were highly correlated, with the third compound additionally identified in the network analysis (Fig. 2, Supporting Information Fig. S3) and produced in large amounts in a strain of *X. szentirmaii* (see Methods section).

Compound **1**, obtained as a colourless crystal, has the molecular formula $C_{16}H_{24}NaO_8$ as deduced from its HR-ESI-MS at *m/z* 367.1366 $[M + Na]^+$ (calcd for $C_{16}H_{24}NaO_8$, 367.1363) in combination with ^1H and ^{13}C NMR data (Supporting Information Table S3 and Supporting Information Figs. S14–S18). By comparing its spectroscopic and single-crystal X-ray diffraction data with those reported previously in literature, it was identified as (4*R*,8*R*,12*R*,16*R*)-4,8,12,16-tetramethyl-1,5,9,13-tetraoxacyclohexadecane-2,6,10,14-tetrone, a cyclic tetramer of (*R*)-3-hydroxybutyrate (Fig. 4, Supporting Information Table S3) (Plattner *et al*., 2004; Riddell *et al*., 2004). The presence of the signal with an *m/z* of 155.07 (Fig. 4A and B) can also be explained by the structure of **1** (Fig. 4C) with ester bond cleavage followed by the elimination of water (Fig. 4A), while the signal with *m/z* of 368.14 is the ^{13}C isotope of **1**.

Single features are capable of discerning genera with high accuracy

Higher clustering (lower correlation thresholds) of the metabolite data resulted in the signal with an *m/z* of 155.07, being identified as having, by far, the largest influence in model output in all cases (Supporting Information Fig. S4–S9). Focusing on all metabolites belonging to the same cluster as this metabolite, as well as those belonging to the clusters represented by the metabolites ranked second and third by SHAP values,

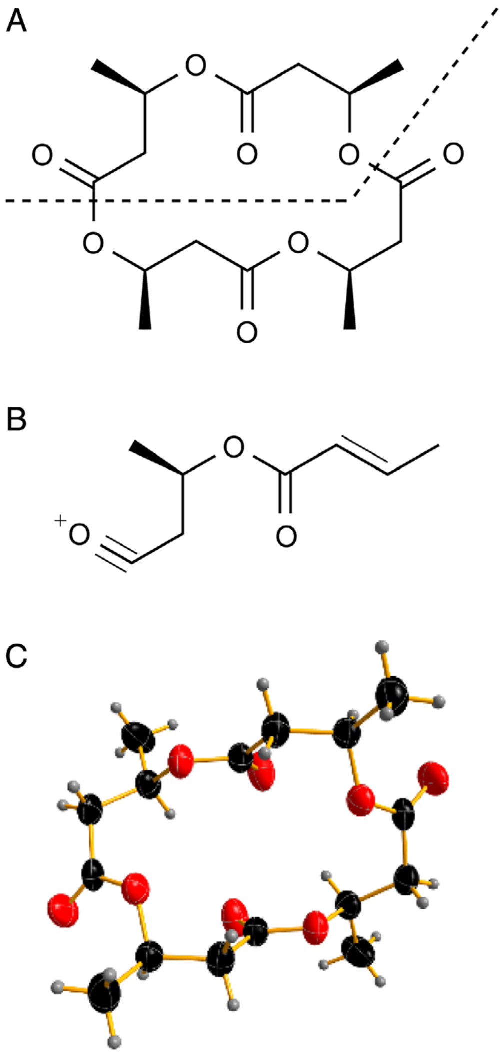

Fig. 4. Structure of (4*R*,8*R*,12*R*,16*R*)-4,8,12,16-tetramethyl-1,5,9,13-tetraoxacyclo hexadecane-2,6,10,14-tetrone (1). The structure (A) and the fragment responsible for the signal at *m*/*z* 155 (B) is indicated as well as the ORTEP representation of its crystal structure (CCDC 1880748) (C).

we proceeded to retrain the model, employing as a feature only one metabolite at a time. We found that the three best single predictors in terms of receiver operating characteristic – area under the curve (ROC-AUC) for both the intensity and AUC data corresponded to signals with an *m*/*z* of 155.07, 368.14 and 367.14 (Fig. 3). These can be used as sole predictors while maintaining a very high performance, equivalent to using the full set of metabolites (Supporting Information Table S4).

To explore whether the three top ranking features, all belonging to the same cluster of signals, significantly impacted the model's performance, we removed all features associated with this cluster and recalculated the model. The resulting top-ranking feature and its highly correlated features were again removed and the model recalculated a third time for comparison. The performance after removing these clusters remained high at 95.2% ± 1.44% and 95% ± 1.3%, respectively, with other signals showing a highly discriminatory effect between *Photorhabdus* and *Xenorhabdus* (Supporting Information Figs. S10 and S11). However, the top three clusters all related to features present in the *Xenorhabdus* and absent in *Photorhabdus.* We therefore identified features that were negatively correlated with the top-ranking cluster and used this as a sole predictor for the genera. In essence, the original model was able to predict a *Photorhabdus* by the absence of the three aforementioned top-ranking features. By using a negative correlation, we aimed to identify compounds that were present in a majority of *Photorhabdus*, but absent in *Xenorhabdus.* This resulted in the identification of a signal with an *m*/*z* of 487.19 (predicted sum formula: $C_{26}H_{25}N_5O_5$), whose fragmentation pattern suggests it might be a peptide (Supporting Information Fig. S12). Additionally, this metabolite was also detected in the network analysis, albeit in a much smaller cluster of nodes (Fig. 2). Using this feature as a sole predictor of genus resulted in a performance of 92.9% ± 2.99%.

Model testing on unseen data

Fourteen *Photorhabdus* and 15 *Xenorhabdus* were randomly selected from the strain collection used for generating the original model, grown and extracted from both media types, in triplicate. These new HPLC-MS runs, unseen by the model during training, were used to test its general performance. From the metabolites present in the data, we located the closest match (see Methods section) for each of the three previously identified best predictors and obtained the class probabilities for each sample. In all cases, the single-feature models were able to correctly classify the genera of the samples with 92.0%–96.5% accuracy. The results are summarized in Supporting Information Table S5.

Discussion

Typically, the similarities between *Photorhabdus* and *Xenorhabdus* are highlighted, particularly with respect to their life cycles. While these similarities hold true, several recent efforts have sought to decipher their differences and what makes these genera unique (Chaston *et al*., 2011; Tobias *et al*., 2017). Our recent work approached this from more of a genomic perspective, while here we attempt to answer this same question using metabolomics as a guide.

It is known that *Photorhabdus* and *Xenorhabdus* are capable of infecting different insect species leading to profoundly different experimental outcomes. This is probably because of the number of compounds which, generally speaking, suppresses the innate insect immune response (Tobias *et al*., 2018a). What we do not know, however, is the degree of dependence that the bacteria have upon their repertoire of metabolites to adapt to the abiotic environment. Interestingly, these bacteria have not yet been isolated as free-living organisms; only in conjunction with their cognate nematode symbionts. We wanted to explore the hypothesis that strains collected in geographically different and abiotically diverse environments (pH, soil type, soil temperature, soil moisture, elevation above sea level) produce different metabolites, specific to that environment, thereby maintaining some form of localized niche despite the mobility afforded by nematode hosts.

A large collection of *Xenorhabdus* and *Photorhabdus* strains was acquired from Thailand, including a number of samples collected from the same geographic locations (Fig. 1). Once isolated, we hypothesized that, by growing the strains under different conditions and collating the data, we would have a data set that represented the metabolic potential of each of the 114 strains. For that reason, we grew the strains in a rich media (LB) in order to provide an environment whereby it would not be disadvantageous (from an energy perspective) to produce compounds and also in SF900, an insect culture medium that reflects the environment these strains may encounter within an insect. A network analysis of the 58 *Photorhabdus* and 56 *Xenorhabdus* was performed using the GNPS platform, which examines mass differences and fragmentation patterns between metabolites in order to determine whether they are likely to be related from a chemical perspective. Despite the overrepresentation of some species in this collection, a combined network analysis of the 114 strains in both media highlights the chemical diversity present in Thailand by entomopathogenic bacteria, regardless of species (6890 nodes, Fig. 2). Our previous work annotated a number of metabolites from both *Photorhabdus* and *Xenorhabdus* and using this library, we identified 14 networks containing known clusters of metabolites (Fig. 2). It is also clear from these analyses that there are a number of major metabolite families that we have yet to identify. Furthermore, it is known that both *Photorhabdus* and *Xenorhabdus* have several different mechanisms at their disposal to help generate natural product diversity from a single gene cluster (Cai *et al*., 2016; Tobias *et al*., 2018b). In fact, the rhabdopeptides are known to be virulence factors towards insects and have an unusual mechanism of generating SM variation by altering the stoichiometry of each module (Cai *et al*., 2016). This variation may actually contribute to the ability of these bacteria to infect different insects, adapting to different insects primarily by altering protein expression levels. In this analysis, we see a large number of features (330) in the network containing known rhabdopeptides (Fig. 2). If this is a major factor conferring virulence to the bacteria, this might be indicative of an insect-specific adaptation.

These bacteria are of general interest due to their SM producing abilities. A recent rarefaction analysis of all sequenced *Xenorhabdus* and *Photorhabdus* genomes suggests that sequencing of a new species would yield, on average, one additional biosynthetic gene cluster per species sequenced. Notably, a recent study in *Myxobacteria* highlights the fact that strain collections with a threshold of taxonomic diversity and coverage is required in order to rapidly identify compounds with a high likelihood of containing structural novelty (Hoffmann *et al*., 2018). In this analysis, there was a large overrepresentation of *X. stockiae* species, but several new derivatives of known compounds. While we do not dispute that structural novelty is important, we do observe that natural structural diversity present in bacteria that make compound libraries may also be important for structure–function studies. To that effect, the generation of new derivatives of known SM from these bacteria, through *in vitro* combinatorial biosynthesis, is ongoing with a view to identifying compounds with higher bioactivities (Bozhüyük *et al*., 2017). What our analysis suggests is that there is a strong possibility that many of these derivatives may also exist ‘naturally’ in the environment as evidenced by the extensive molecular networks containing ‘known’ compounds. Despite the apparent abundance of new derivatives, this also suggests that our prediction of one new SM per species is a significant underestimation if we consider unknown derivatives.

Recently, it was found that genes in strains isolated from similar environments, which are also the same species, contain a number of differences at the genetic level (Murfin *et al*., 2015). We envisaged that we may therefore be able to differentiate between different metadata based upon each strain’s unique metabolome. We used the compiled metabolomic data, together with the metadata, to train a machine learning model; in particular, we chose to make use of GBDTs. Models of this type enjoy a high level of popularity due to their high efficiency and the state-of-the-art performance, as well as the availability of fast, ready-to-use implementations. In addition to this, they tend to perform well, even in very-high-dimensional scenarios, especially in cases when the features outnumber the samples or observations, a phenomenon commonly referred to as the ‘curse of dimensionality’(Mayr *et al*., 2014; Nielsen, 2016). As such, GBDT models are ideally suited for the type of data we are dealing with – and metabolomics data in general – having tens of thousands of metabolites for a few hundred samples.

In addition to the above, GBDT models are also robust to multicollinearity between features. As seen from the results, the model does not suffer a performance drop when highly correlated metabolites are present. Nevertheless, we decided to cluster the metabolites, and drop correlated variables, for interpretability reasons: faced with two or more highly correlated features that are very good predictors, the model will greedily choose to split on one of them in detriment of the others. In other words, features that are otherwise highly discriminatory will have their impact underestimated in the ranking of importance.

One weakness in studies such as this is the use of artificial *in vitro* culture conditions to explore the metabolic diversity. Underpinning this is the range of soil pH (5.2–7) and temperatures (18–32°C) associated with the strains in the present study. Since we were interested in metabolic potential, we made the decision to culture all strains under standard laboratory conditions. However, one future analysis could include the growth and subsequent metabolite extraction under more refined conditions. In comparative genetic studies, we typically compare whole genomes to draw inferences on the data, thus basing future hypotheses on the genetic potential, rather than gene expression. In the same principle, we base our conclusions here on metabolic potential and work towards overcoming the limitations associated with the non-natural environment by using different conditions and collating the data. Given that no evidence was seen for metadata influencing metabolite production, we used a machine learning model to investigate the differences between *Photorhabdus* and *Xenorhabdus.* During training of the model, SHAP values were obtained in order to assess and rank the impact of the feature values on model output. Our reasoning behind this was that we could then prioritize metabolites for purification and chemical structure elucidation. We chose SHAP values as our measure of importance, because they provide per-sample explanations that are proven to be both consistent and locally accurate, as opposed to GBDTs built-in measures (Lundberg and Lee, 2017; Lundberg *et al*., 2018), in addition to being a model-agnostic feature attribution approach that does not require the model to be tree-based.

From the SHAP results, we observe that while only a few metabolites – exactly one, for the most heavily clustered data – have a very large impact on model output in comparison with the rest, many more seem to be strong discriminators between classes, as evidenced by the colouring of their values and the direction of their impact, despite the latter being relatively low. Indeed, removal of the most important cluster from the data set still resulted in very high classification performance when taking all other metabolites in consideration (Supporting Information Fig. S10). Single-feature predictions, however, do suffer from a steeper performance drop than the metabolites we have identified as the best predictors. Therefore, we emphasize that we have not attempted to find the ‘only’ metabolites that set these two genera apart, but to prioritize the ones that appear to be the strongest in doing so. The relevance of this and the usefulness of single-feature models become apparent when dealing with new, unseen data: in the case presented here, the test data set contains 15,098 metabolite columns, which renders futile any attempt at full data set peak matching.

A recent study in Australia examined the differences between the biosynthetic domain compositions in soil across the continent. One key finding from this was that the composition of natural product domains, specifically ketosynthase domains (from PKS) or adenylation domains (from NRPS), changed with latitude and longitude and was often grouped in accordance with the vegetation type (Lemetre *et al*., 2017). This supports our original premise that natural product composition from the *Xenorhabdus* and *Photorhabdus* may change within the country. However, in our analysis, we saw no clear clustering of strains based on any of the abiotic factors measured. Considering that the bacteria have never been isolated independent of the nematode, several explanations exist for the lack of obvious metabolite clustering in different environments. One explanation is that the nematodes, and the insects that they infect, are all motile and may help spread the bacteria in the environment, thus confounding any underlying association with geography. One further explanation is that the nematode hosts provide the greater support in these environments. In turn, the SMs produced by the bacteria then provide specificity for the host and the invertebrate prey. This would actually point towards a dependence of the bacteria upon the nematode in the environment, an area that has not been widely investigated due to the relative simplicity to investigate the bacteria independently in a lab environment.

Purification of compound **1** resulted in elucidation of a cyclic tetramer of hydroxybutyrate (Fig. 4), a compound related to crown ethers. Crown ethers typically demonstrate a high affinity to cations and are often cytotoxic, but may also show characteristics of ionophores. Ionophores in natural biological systems help to transport ions across cell membranes by forming lipid-soluble complexes with polar cations (Bakker *et al*., 1997). Given the probable influence of nematode host on metabolite production, one explanation for the specific presence of these compounds in *Xenorhabdus* could be that they are required during the symbiosis with *Steinernema.* While this is probably not a ubiquitous requirement since the compound was not detected in all species of *Xenorhabdus* (Supporting Information Fig. S13), it is interesting that the majority of the *Xenorhabdus*, with the exception of *X. szentirmaii*, were originally isolated in South East Asia. One interesting note is that the nematode hosts of *X. szentirmaii* (*Steinernema rarum*) and *X. stockiae* (*Steinernema siamkayai*) are close evolutionary relatives (Stock, 1998), supporting a possible role of this metabolite in symbiosis.

One major challenge in large-scale metabolomic studies is how to prioritize research efforts. Here, we set out an analysis pipeline that is capable of using strain-specific metadata, coupled with high-throughput MS experiments. Whether it is determining compounds important for an ecological niche or identifying as yet undiscovered compounds in large high-throughput screening experiments. By incorporating machine learning models such as this into current analysis pipelines, the relative importance of compounds can be determined in order to streamline purification and/or structure elucidation pipelines in a time-efficient manner, yielding low probabilities of rediscovery.

Materials and methods

Soil collection

Samples were taken from diverse habitats including natural grassland, roadside verges, woodlands and banks of ponds and rivers. For each site, five soil samples were randomly taken in an area of approximately 100 m^2 at a depth of 10–20 cm using a hand shovel. Approximately 500 g of each soil sample was placed into a plastic bag. The longitude, latitude and altitude of each sampling site were recorded using a GPSMAP 60CSx (Garmin, Taiwan). The temperature, pH and moisture of each sample were recorded using a Soil pH & Moisture Tester (Model: DM-15, Takemura electric works, Ltd., Japan).

Isolation of Xenorhabdus *and* Photorhabdus *bacteria from entomopathogenic nematodes*

Dead *Galleria mellonella* larvae were surface-sterilized by dipping into absolute ethanol for 1 min and placed in a sterile petri dish to dry. Sterile forceps were used to nip the third ring from the head of *G. mellonella*, thereby removing the cuticle. A sterile loop was used to touch haemolymph of *G. mellonella* and streaked onto a nutrient bromothymol blue agar supplemented with 0.004% (w/v) triphenyltetrazolium chloride (TTC, Sigma, St. Louis, KS) and 0.0025% (w/v) bromothymol blue (Akhurst, 1980). TTC was added to inhibit the growth of Gram-positive, acid-fast bacteria and actinomycetes. Cultured plates were incubated in the dark at room temperature for 4 days. *Xenorhabdus* and *Photorhabdus* strains were characterized based on colony morphology as described by Boemare and Akhurst (Boemare and Akhurst, 1988). Single colonies were then subcultured on the same medium and kept in Luria-Bertani (LB) containing 20% glycerol at –80°C for further identification.

Bacterial identification

DNA was extracted using a Genomic DNA Mini Kit (blood/Cultured Cell) (Geneaid Biotech Ltd., Taiwan). Polymerase chain reaction (PCR) targeting *recA* was performed in 50 μl volumes using 10 μl of 5× buffer (Promega, Madison, WI), 7 μl of 25 mM $MgCl_2$ (Promega, Madison, WI), 1 μl of 200 mM dNTPs (New England Biolabs Inc., Ipswich, MA), 2 μl of 5 μM of each Primer, 0.5 μl of 5 unit Taq Polymerase (Promega, Madison, WI) and 2.5 μl of DNA template. The *recA* primer sequences were recA1_F (5'-GCTATTGA TGAAAATAAACA-3′) and recA2_R (5'-RATTTTRTCWCC RTTRTAGCT-3′) (Tailliez *et al*., 2010).

PCR cycling parameters for *recA* of *Xenorhabdus* included an initial denaturing step of 94°C for 5 min, followed by 30 cycles of denaturation at 94°C for 1 min, annealing temperature of 50°C for 1 min and extension of 72°C for 2 min and a final extension of 72°C for 7 min. Parameters for *Photorhabdus* included an initial denaturing step at 94°C for 5 min, followed by 30 cycles of 94°C for 1 min, 50°C for 45 s and 72°C for 1.5 min, with a final extension of 72°C for 7 min. The PCR products of *recA* of both genera (890 bp) were examined on 1.5% agarose gel electrophoresis. Fifty microlitres of PCR products were purified using Gel/PCR DNA Fragments Extraction Kit (Geneaid Biotech Ltd., Taiwan). *recA* sequencing was performed on the ABI PRISM 3100 Genetic Analyzer (Amersham Bioscience, UK) using the PCR primers for PCR. Chromatograms, sequence ambiguity resolution were visually checked using the SeqManII software (DNASTAR Inc., Wisconsin). Species identification was performed using a nucleotide Blast search of *recA* against the NCBI nucleotide database and the match with the highest similarity score was selected (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Multiple nucleotide sequences representing all of the known species and subspecies of *Photorhabdus* and *Xenorhabdus* spp. were downloaded from the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi), aligned with sequences from the study isolates and trimmed to a 646 bp region using ClustalW (Thompson *et al*., 1994) in MEGA version 5.0 (Tamura *et al*., 2011). Maximum likelihood trees were reconstructed using Nearest-Neighbour-Interchange (NNI) and Tamura–Nei model (Tamura and Nei, 1993) using MEGA version 5.05 (Tamura *et al*., 2011). Bootstrap analysis was carried out with 1000 data sets.

Metabolite extraction

Bacterial cultures were grown in either SF900 media or LB for 72 h at 30°C. A 1 ml sample was taken from each culture and extracted with an equal volume of methanol, mixed briefly by vortexing and centrifuged for 30 min. The resulting supernatant was dried under a constant stream of nitrogen gas, to completion. Prior to measurement, samples were resuspended in 500 μl of methanol and centrifuged for 30 min.

Ultraperformance liquid chromatography high-resolution mass spectrometry measurements

UPLC-ESI-HRMS/MS analyses were performed using an UltiMate 3000 system linked to a Bruker Impact II qTof mass spectrometer. Runs were performed using a flow rate of 0.4 ml min^{-1} and gradient of MeCN/0.1% formic acid in H_2O (5:95% to 95:5% over 15 min). Data acquisition was performed as previously described (Tobias *et al*., 2017).

Molecular networking analysis

The raw MS data of 114 environmental isolates, *E. coli* (all in LB and SF900), LB, SF900 and acetonitrile blanks were converted to the .mzXML format using DataAnalysis v4.3 (Bruker). Molecular networks were created using the online workflow at GNPS (Wang *et al*., 2016). The data were then clustered with MS-Cluster with a parent mass tolerance of 0.05 Da and a MS/MS fragment ion tolerance of 0.01 Da to create consensus spectra. Further, consensus spectra that contained less than two spectra were discarded. A network was then created where edges were filtered to have a cosine score above 0.7 and more than 6 matched peaks. Further edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top seven most similar nodes. The spectra in the network were then searched against GNPS' spectral libraries. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least six matched peaks. Analogue search was enabled against the library with a maximum mass shift of 100.0 Da. The self-loop networks were imported into Cytoscape (v3.4.0) for visualization.

Feature identification

Mass spectrometry files were imported into DataAnalysis (v4.3) and converted from the Bruker .m format to the open mzXML format for processing with MZMine2 (Pluskal *et al*., 2010). After import, mass detection was performed with the mass detector set to centroid, noise level to 1000, at MS level 1 and with a retention time of 0–16.05 min. Chromatograms were then built with the retention time between 0 and 16.05 min, MS level 1, a minimum time span of 0.02, a minimum height of 1000 and an *m*/*z* tolerance of 0.005 m/z or 5.0 ppm. Peak deconvolution was performed with the noise amplitude algorithm, a minimum peak height of 1000, peak duration in the range 0–0.8 min and an amplitude of noise set to 5000.

The peak aligner was then set with an *m*/*z* tolerance of 0.005 m/z or 5.0 ppm, the weight of *m*/*z* at 20, retention time tolerance at 3% relative, weight for retention time of 10, with peaks requiring the same charge state and 'compare isotope pattern' set to yes with the setting for isotope *m*/*z* tolerance 0.005 m/z or 5.0 ppm, a minimum absolute intensity of 1000, and a minimum score of 65%. Gap filling was then used using the 'same RT and m/z range gap filler' with *m*/*z* tolerance set to 0.005 m/z or 5.0 ppm. The aligned, filled mass list was then exported as a .csv file.

Machine learning data preprocessing

In order to determine the importance of compounds, we decided to employ a machine learning model. In conjunction with a recently developed feature attribution method, this serves the twofold purpose of achieving a very high performance in discriminating between the two genera, yielding a model that can be subsequently used to classify new data, while at the same time allowing for a direct visualization of the features that have the largest impact on the model's predictions for each of the samples.

The intensity and AUC data obtained from the MZmine2 peak picking algorithm were used. As a first step, we generated an additional data set by setting to zero all AUC entries for which the corresponding peak intensity was zero. Samples were further processed by removing all columns corresponding to metabolites that were absent in all of the samples after deletion of *E. coli*, media only and acetonitrile blanks, since they would not contribute to the classification. In addition to this, we removed all columns with less than *c.* 10% of non-zero values. The data were further cleaned up by clustering the metabolite columns according to their correlation across samples and discarding all but one of the members of any one cluster; the correlation thresholds used were 0.9, 0.95 and 0.99. Numerical metadata was scaled between 0 and 1 for pH, temperature and moisture, while the elevation, spanning three orders of magnitude, was converted into logarithmic scale. Location data, in turn, was kept to the level of province and one-hot-encoded; soil type and medium data were also one-hot-encoded. The smallest resulting data set consisted of 20,650 and 21,634 metabolite columns, out of a total of 44,836, for the intensity and zeroed AUC data, respectively, plus 20 metadata columns: 2 media conditions, 4 soil types, 10 provinces, pH, temperature, moisture and elevation.

Generating a model

The pruned data sets from the previous section were used to train a GBDT model. Here, we used the Python implementation of LightGBM (Ke *et al*., 2017) to train a classifier on the pruned intensity and AUC data sets. We used 250 iterations, with 50 iterations as the threshold for early stopping, defined as the number of steps the model can take without improvements on the evaluation metric. The latter is calculated from the predictions of the model for a pre-defined validation set. To this end, we performed 100 rounds of fivefold cross-validation on the data sets, and report the resulting mean and standard deviation of the mean accuracy and ROC-AUC across folds.

Determining feature importance

In order to interpret the predictions from the GBDT model and determine the most important features driving its output, we computed the SHAP values for each feature and averaged them over all the training rounds. The values are individualized per sample and correspond to the change in log-odds of the sample being classified as corresponding to one or the other genus – in this case, a positive value indicates a larger probability of being *Xenorhabdus* – relative to the mean prediction upon addition of a given feature, effectively measuring the impact that every feature value has on every sample. This was carried out using the tree ensemble implementation of the SHAP Python package (Lundberg *et al*., 2018).

All code used for this paper is available at https://github.com/systemsmedicine/geographical-chemotypes as Jupyter notebooks, providing a step-by-step walkthrough.

Compound isolation and purification

For the isolation and purification of (4*R*,8*R*,12*R*,16*R*)-4,8,12,16-tetramethyl-1,5,9,13-tetraoxacyclohexadecane-2,6,10,14-tetrone, the XAD-16 resin from a 4 l M63 medium culture of *X. szentirmaii*_P1 (phenazine gene cluster knockout) mutant (Shi *et al*., 2019) were harvested after 72 h of incubation at 30°C with shaking at 120 r.p.m., washed with water and extracted with methanol (3 l × 1 l) to yield the crude extract (1.1 g) after evaporation. The extract was dissolved in methanol and was subjected to preparative HPLC-MS with C-18 column (21.2 mm × 250 mm, 7.0 μm, Agilent) using an acetonitrile/water gradient (0.1% formic acid) in 30 min, 5%–95% to afford a sub-fraction mainly containing 8.3 mg. The sub-fraction was further purified by semipreparative HPLC with C-18 column (9.4 mm × 250 mm, 5.0 μm, Agilent) using an acetonitrile/water gradient (0.1% formic acid) 0–30 min, 30%–45% to afford (4*R*,8*R*,12*R*,16*R*)-4,8,12,16-tetramethyl-1,5,9,13-tetraoxacyclohexadecane-2,6,10,14-tetrone (2.1 mg). ^{1}H and ^{13}C NMR, ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC), ^{1}H-^{13}C Heteronuclear Multiple Bond Correlation (HMBC), and ^{1}H-^{1}H Correlation Spectroscopy (COSY) were measured. Chemical shifts (δ) were reported in parts per million (ppm) and referenced to the solvent signals. Data are reported as follows: chemical shift, multiplicity (d = doublet, dd = doublet of doublet, and m = multiplet), and coupling constants in Hertz (Hz).

Acknowledgements

The authors would like to thank Dr. Lothar Fink from Goethe University for conducting the X-ray crystallography structure determination. Financial support was provided by Naresuan University (Grant Number R2560B073). CPR and EAHV were supported by the Alfons und Gertrud Kassel-Stiftung. YMS is the recipient of a Humboldt Postdoctoral Fellowship. Work in the Bode lab was supported by the LOEWE-TBG initiative.

References

Akhurst, R.J. (1980) Morphological and functional dimorphism in Xenorhabdus spp., bacteria symbiotically associated with the insect pathogenic nematodes Neoaplectana and Heterorhabditis. *Microbiology* **121**: 303–309.

Bakker, E., Bühlmann, P., and Pretsch, E. (1997) Carrier-based ion-selective electrodes and bulk Optodes. 1. General characteristics. *Chem Rev* **97**: 3083–3132.

Böcker, S., Letzel, M.C., Lipták, Z., and Pervukhin, A. (2009) SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**: 218–224.

Boemare, N.E., and Akhurst, R.J. (1988) Biochemical and physiological characterization of Colony form variants in Xenorhabdus spp. (Enterobacteriaceae). *Microbiology* **134**: 751–761.

Bozhüyük, K.A.J., *et al*. (2017) De novo design and engineering of non-ribosomal peptide synthetases. *Nat Chem* **10**: 275–281.

Cai, X., *et al*. (2016) Entomopathogenic bacteria use multiple mechanisms for bioactive peptide library design. *Nat Chem* **9**: 379–386.

Chaston, J.M., Suen, G., Tucker, S.L., Andersen, A.W., Bhasin, A., Bode, E., *et al*. (2011) The entomopathogenic bacterial endosymbionts Xenorhabdus and Photorhabdus: convergent lifestyles from divergent genomes. *PLoS ONE* **6**: e27909.

Forst, S., Dowds, B., Boemare, N., and Stackebrandt, E. (1997) Xenorhabdus and Photorhabdus spp.: bugs that kill bugs. *Annu Rev Microbiol* **51**: 47–72.

Han, R., and Ehlers, R.U. (2000) Pathogenicity, development, and reproduction of Heterorhabditis bacteriophora and Steinernema carpocapsae under axenic in vivo conditions. *J Invertebr Pathol* **75**: 55–58.

Hoffmann, T., Krug, D., Bozkurt, N., Duddela, S., Jansen, R., Garcia, R., *et al*. (2018) Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat Commun* **9**: 803.

Katajamaa, M., Miettinen, J., and Oresic, M. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22**: 634–636.

Ke, G. *et al*. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. 3146–3154 (2017).

Lemetre, C., Maniko, J., Charlop-Powers, Z., Sparrow, B., Lowe, A.J., and Brady, S.F. (2017) Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. *Proc Natl Acad Sci U S A* **114**: 11615–11620.

Lundberg, S.M., and Lee, S.-I. (2017) A unified approach to interpreting model predictions. *Proceedings of the 31st Neural Information Processing Systems (NIPS-17)*. arXiv: 1705.07874v2.

Lundberg, S. M., Erion, G. G. & Lee, S.-I. (2018) Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888v3.

Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014) The evolution of boosting algorithms - from machine learning to statistical Modelling. *Methods Inf Med* **53**: 419–427.

Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., *et al*. (2018) Dereplication of microbial metabolites through database search of mass spectra. *Nat Commun* **9**: 4035.

Murfin, K.E., Whooley, A.C., Klassen, J.L., and Goodrich-Blair, H. (2015) Comparison of Xenorhabdus bovienii bacterial strain genomes reveals diversity in symbiotic functions. *BMC Genomics* **16**: 889.

Nielsen, D. (2016) Tree Boosting With XGBoost-Why Does XGBoost Win' Every' Machine Learning Competition? (Master's thesis). Trondheim, Norway: Norwegian University of Science and Technology.

Plattner, D.A., *et al*. (2004) Cyclische Oligomere von (R)-3-Hydroxybuttersäure: Herstellung und strukturelle Aspekte. *Helv Chim Acta* **76**: 2004–2033.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**: 395.

Riddell, F.G., Seebach, D., and Müller, H.-M. (2004) Solid-state CP/MAS 13C-NMR spectra of Oligolides derived from 3-hydroxybutanoic acid. *Helv Chim Acta* **76**: 915–923.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Shi, Y.-M., and Bode, H.B. (2018) Chemical language and warfare of bacterial natural products in bacteria–nematode–insect interactions. *Nat Prod Rep* **92**: fiw007.

Shi, Y.-M., Brachmann, A.O., Westphalen, M.A., Neubacher, N., Tobias, N.J., and Bode, H.B. (2019), under revision) Dual phenazine gene clusters enable diversification during biosynthesis. *Nat Chem Biol* **15**: 331–339.

Stock, S.P. (1998) Steinernema siamkayai n. sp. (Rhabditida: Steinernematidae), an entomopathogenic nematode from Thailand. *Syst Parasitol* **41**: 105–113.

Stock, S.P., Campbell, J.F., and Nadler, S.A. (2001) Phylogeny of Steinernema travassos, 1927 (Cephalobina: Steinernematidae) inferred from ribosomal DNA sequences and morphological characters. *J Parasitol* **87**: 877–889.

Tailliez, P., Laroui, C., Ginibre, N., Paule, A., Pagès, S., and Boemare, N. (2010) Phylogeny of Photorhabdus and Xenorhabdus based on universally conserved protein-coding sequences and implications for the taxonomy of these two genera. Proposal of new taxa: X. vietnamensis sp. nov., P. luminescens subsp. caribbeanensis subsp. nov., P. luminescens subsp. hainanensis subsp. nov., *P. temperata* subsp. khanii subsp. nov., *P. temperata* subsp. tasmaniensis subsp. nov., and the reclassification of P. luminescens subsp. thracensis as *P. temperata* subsp. thracensis comb. nov. *Int J Syst Evol Microbiol* **60**: 1921–1937.

Tamura, K., and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512–526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Tobias, N.J., Mishra, B., Gupta, D.K., Sharma, R., Thines, M., Stinear, T.P., and Bode, H.B. (2016) Genome comparisons provide insights into the role of secondary metabolites in the pathogenic phase of the Photorhabdus life cycle. *BMC Genomics* **17**: 537.

Tobias, N.J., *et al*. (2017) Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. *Nat Microbiol* **1354**: 82–1685.

Tobias, N.J., Shi, Y.-M., and Bode, H.B. (2018a) Refining the natural product repertoire in Entomopathogenic bacteria. *Trends Microbiol* **26**: 833–840.

Tobias, N.J., Linck, A., and Bode, H.B. (2018b) Natural product diversification mediated by alternative transcriptional starting. *Angew Chem Int Ed Engl* **57**: 5699–5702.

Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., *et al*. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828–837.

Wilkinson, P., Waterfield, N.R., Crossman, L., Corton, C., Sanchez-Contreras, M., Vlisidou, I., *et al*. (2009) Comparative genomics of the emerging human pathogen Photorhabdus asymbiotica with the insect pathogen Photorhabdus luminescens. *BMC Genomics* **10**: 302.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Supplementary Fig. S1. Clustering of features of either intensity or AUC was performed using three different correlation cutoffs; 0.9, 0.95 and 0.99. These clusters can be explored interactively at cparrarojas.github.io/blog/2019/02/geographical-chemotypes.

Supplementary Fig. S2. Representative extracted ion chromatograms of the top 3 ranking features as determined by the machine learning model.

Supplementary Fig. S3. Network containing **1** showing parent masses inside nodes and edges representing mass differences between metabolites.

Supplementary Fig. S4. SHAP values were generated on models following the clustering of intensity data based upon a correlation cut-off of 0.9.

Supplementary Fig. S5. SHAP values were generated on models following the clustering of intensity data based upon a correlation cut-off of 0.95.

Supplementary Fig. S6. SHAP values were generated on models following the clustering of intensity data based upon a correlation cut-off of 0.99.

Supplementary Fig. S7. SHAP values were generated on models following the clustering of AUC data based upon a correlation cut-off of 0.9.

Supplementary Fig. S8. SHAP values were generated on models following the clustering of AUC data based upon a correlation cut-off of 0.95.

Supplementary Fig. S9. SHAP values were generated on models following the clustering of AUC data based upon a correlation cut-off of 0.99.

Supplementary Fig. S10. All features associated with the top-ranking cluster were removed and the model was recalculated. The top 10 highest ranking features are shown based on the output of SHAP. Performance metrics can be seen in **Supplementary Table 4.**

Supplementary Fig. S11. All features associated with the top two ranking clusters were removed and the model was recalculated. The top 10 highest ranking features are shown based on the output of SHAP. Performance metrics can be seen in **Supplementary Table 4.**

Supplementary Fig. S12a. Base peak chromatogram of a representative *Photorhabdus* strain (number 448), **(b)** with an extracted ion chromatogram of the signal with the highest negative correlation to **1** and *m/z* of 487.186. **(c)** The fragmentation pattern of this compound, **2**, is also shown.

Supplementary Fig. S13. Extracted ion chromatograms of other *Xenorhabdus* species containing the (cyclo) tetrahydroxybutyrate (**1**).

Supplementary Fig. S14. ^{1}H NMR spectrum of **1** in DMSO-d_6.

Supplementary Fig. S15. ^{13}C NMR spectrum of **1** in DMSO-d_6.

Supplementary Fig. S16. HSQC spectrum of **1** in DMSO-d_6.

Supplementary Fig. S17. HMBC spectrum of **1** in DMSO-d_6.

Supplementary Fig. S18. ^{1}H-^{1}H COSY spectrum of **1** in DMSO-d_6.

Supplementary Table S1. All metadata associated with isolates.

Supplementary Table S2. Performance of gradient boosting decision tree model on full data set compared to that of the pruned and clustered data. Clustered data is shown with different cutoff thresholds. All models were calculated using both intensity data and area under the curve (AUC). The degree of clustering can be seen in **Supplementary Fig. S1.**

Supplementary Table S3. ^{1}H (500 MHz) and ^{13}C (125 MHz) NMR data assignments for **1** in DMSO-d_6 (for NMR spectra see Fig. S14-S18).

Supplementary Table S4. Performance of model when using previously identified highly-ranked signals as a single predictor.

Supplementary Table S5. Model performance of single features as sole predictors on unseen data. A total of 15 *Xenorhabdus* (X) and 14 *Photorhabdus* (P) were grown in triplicate and their metabolites extracted as described in the Methods. Listed are percentages representing how often the correct genus was called. Probabilities for calling each sample can be seen in **Supplementary Table S6**.

Supplementary Table S6. Probabilities of calling each sample using data unseen during model creation. In each case the value represents the probability of the sample being *Xenorhabdus*.