Computational and Structural Biotechnology Journal 19 (2021) 3051–3057

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Metabolites as predictive biomarkers for *Trypanosoma cruzi* exposure in triatomine bugs

Fanny E. Eberhard [a], Sven Klimpel [a,b,c], Alessandra A. Guarneri [d], Nicholas J. Tobias [b,c,*]

[a] Institute for Ecology, Evolution and Diversity, Goethe University Frankfurt, Frankfurt/Main, Germany
[b] LOEWE Centre for Translational Biodiversity Genomics (LOEWE TBG), Frankfurt/Main, Germany
[c] Senckenberg Gesellschaft für Naturforschung, Senckenberg Biodiversity and Climate Research Centre, Frankfurt/Main, Germany
[d] Vector Behaviour and Pathogen Interaction Group, Instituto René Rachou, Avenida Augusto de Lima,1715, Belo Horizonte, MG CEP 30190-009, Brazil

ARTICLE INFO

Article history:
Received 18 February 2021
Received in revised form 10 May 2021
Accepted 19 May 2021
Available online 21 May 2021

Keywords:
Trypanosoma cruzi
Metabolomics
Chagas disease
Host-parasite interaction
Rhodnius prolixus
Supervised machine learning

ABSTRACT

Trypanosoma cruzi, the causative agent of Chagas disease (American trypanosomiasis), colonizes the intestinal tract of triatomines. Triatomine bugs act as vectors in the life cycle of the parasite and transmit infective parasite stages to animals and humans. Contact of the vector with *T. cruzi* alters its intestinal microbial composition, which may also affect the associated metabolic patterns of the insect. Earlier studies suggest that the complexity of the triatomine fecal metabolome may play a role in vector competence for different *T. cruzi* strains. Using high-resolution mass spectrometry and supervised machine learning, we aimed to detect differences in the intestinal metabolome of the triatomine *Rhodnius prolixus* and predict whether the insect had been exposed to *T. cruzi* or not based solely upon their metabolic profile. We were able to predict the exposure status of *R. prolixus* to *T. cruzi* with accuracies of 93.6%, 94.2% and 91.8% using logistic regression, a random forest classifier and a gradient boosting machine model, respectively. We extracted the most important features in producing the models and identified the major metabolites which assist in positive classification. This work highlights the complex interactions between triatomine vector and parasite including effects on the metabolic signature of the insect.

1. Introduction

Chagas disease (American trypanosomiasis) is considered to be one of the major neglected tropical disease affecting more than 6 million people worldwide [1]. It is caused by the flagellated protozoan parasite *Trypanosoma cruzi* and occurs predominantly in Central and South America. The parasite is transmitted by haematophagous triatomine vectors of the Reduviidae family, subfamily Triatominae. The insect ingests *T. cruzi* during blood feeding on an infected mammal and, after a period of development in the intestinal tract, releases infective trypomastigote forms of *T. cruzi* within the faeces during a subsequent blood meal. The parasites are then accidentally rubbed into the bite wound or enter the host bloodstream through the mucosa. Further transmission routes include the ingestion of contaminated food, transmission by infected blood products and congenital transmission [1,2]. People who develop chronic Chagas disease face severe medical, social and also economic challenges as they are often vulnerable to superinfections and struggle with a limited ability to work [3,4].

Infection with *T. cruzi* has an impact on the development, fitness and fecundity of their insect vectors. Triatomine bugs infected by the parasite seem to have a retarded developmental time and a reduced survival rate compared to uninfected insects, which applies both for hatching as well as moulting [5,6]. In addition, the reproductive performance is impaired by the presence of *T. cruzi* leading to a decreased fertility and number of eggs [7,7]. These effects are particularly influenced by the insect's gender and the surrounding temperature causing a delay in moulting and allowing the parasite to develop and reach the insect's hindgut [5,7,8,9]. Moreover, the parasite-induced modifications and the loss of fitness are evident in the ecological niche space of the infected vector species by narrowing their niche breadth [10]. Interestingly, a direct influence of the infection status on the behaviour of the insect vector can also be observed. While feeding, infected individuals tend to ingest more blood and defecate earlier and in greater quantity than uninfected controls potentially increasing the probability of mammalian host infection [15,12]. The contact of the triatomine vector with *T. cruzi* also influences

* Corresponding author at: LOEWE Centre for Translational Biodiversity Genomics (LOEWE TBG), Frankfurt/Main, Germany.
E-mail address: Nicholas.tobias@senckenberg.de (N.J. Tobias).

https://doi.org/10.1016/j.csbj.2021.05.027

and alters the prevailing microbial status in the insect intestine. In particular, the microbial species diversity in *T. cruzi*- infected compared to uninfected vectors differs considerably. Higher bacterial species richness is associated with positive infection status and an overrepresentation of distinct bacterial taxonomic groups depending on the vector species. Also, the discrete typing unit (DTU) of *T. cruzi* (TcI, TcIV, TcII/TcV) seems to have an effect on the abundance of different bacterial groups in its host [17–17]. The synergy between *T. cruzi*-infection and an altered microbiota composition suggests a consequentially adapted gastrointestinal metabolome. It has been shown that, additional to a uniform core metabolome which represents 80% of all detected metabolites in the triatomine intestinal tract, it further consists of a highly variable composition of chemical compounds contingent on the triatomine species [18]. Some of the microbiota producing these metabolites serve as symbiotic suppliers of essential nutrients, e.g. vitamin B complexes from *Rhodococcus*, *Dickeya* and other bacterial genera in *Rhodnius prolixus*, and induce vector antiparasitic activity and humoral immune defence factors [23–22]. However, it has yet not been investigated whether the infection of the triatomine vector with *T. cruzi* has an actual impact on its metabolome, and if so, whether the metabolic signature is indicative of an infection with the parasite.

To investigate the chemical ecology inside the triatomine insect, we challenged 5th instar *R. prolixus* with *T. cruzi* and analysed the changes that occur in their intestinal metabolome at different time points after contact, by using ultra high-performance liquid chromatography and mass spectrometry. To our knowledge, this is the first approach to describe the metabolic changes in a triatomine vector following trypanosomal exposure and to integrate our findings into the existing knowledge on vector-host interactions.

2. Methods

2.1. Ethics statement

All experiments using live animals were performed in accordance with FIOCRUZ guidelines on animal experimentation and were approved by the Ethics Committee in Animal Experimentation (CEUA/FIOCRUZ) under the protocol number LW-8/17.

2.2. Insect rearing and parasite cultivation

The *Rhodnius prolixus* colony used in this study was established with insects collected in Honduras in the 1990s and is maintained by the Vector Behavior and Pathogen Interaction group at FIOCRUZ. The colony is maintained at 25 ± 1 °C, 60 ± 10% RH and a natural illumination cycle. Insects were fed monthly on citrated rabbit blood obtained from CECAL (Fiocruz, Rio de Janeiro, Brazil) offered through an artificial feeder at 37 °C, or chicken previously anesthetized with intraperitoneal injections containing mixtures of ketamine (20 mg/kg; Cristália, Brazil) and detomidine (0.3 mg/kg; Syntec, Brazil).

Trypanosoma cruzi (Dm28c strain, TcI) isolated from naturally infected *Didelphis marsupialis* [23] was used to infect the triatomines. Parasites were cultured *in vitro* by twice a week passages in LIT (liver-infusion tryptose) medium supplemented with 15% fetal bovine serum (FBS), 100 mg/ml streptomycin and 100 units/ml penicillin. Strain infectivity was maintained by continuous full cycle infections on triatomine and mouse hosts every six months [24].

2.3. Insect infection

Groups of 5th instar *R. prolixus* nymphs were infected with either *T. cruzi* epimastigotes or trypomastigotes. For epimastigote exposure, insects were fed through a latex membrane with an artificial feeder containing citrated (10% v/v), heat-inactivated (56 °C, 30 min) rabbit blood heated to 37 °C containing a suspension of epimastigotes (10^6 epimastigotes/ml). For infection with *T. cruzi* trypomastigotes, mice were infected by the injection of metacyclic trypomastigotes intraperitoneally and used to feed the insects at day 9 post-infection. Parasitemia in the mice was 7–15 trypomastigotes/μl.

2.4. Sample preparation

Triatomine guts were separated into anterior midgut, posterior midgut and hindgut and transferred into Eppendorf tubes containing 40 μl of PBS in pools of five insects per tube. Three glass beads (Sigma-Aldrich) and 1 ml of methanol (99%) were added to each tube and samples were homogenized for 3 min at a frequency of 30/s (Retsch MM 400) to break up both insect tissue and cells. Separation of remaining cell fragments was achieved through centrifugation (Eppendorf 5424 R) for 5 min at 10.000 rpm at a temperature of 10 °C. Cell pellets were then extracted twice more, with the supernatants from each extraction combined (3 ml total) and dried in a SpeedVac (Eppendorf Concentrator Plus, Labconco CentriVap Concentrator) at 30 °C. An unexposed control group was also investigated, using insects fed blood free of parasite (either uninfected mice for trypimastigotes or uninfected blood for epimastigote experiments) as well as a group of starved, unfed insects. In addition, extraction blanks were conducted using PBS. All experiments were performed in triplicate. Samples were taken immediately after feeding and at 24 h, 48 h and 72 h post-infection for both epimastigote and trypomastigote experiments. In order to reduce the amount of fatty acids and other lipids, all samples were dissolved in 1 ml of methanol and 1 ml of hexane and mixed thoroughly. Subsequently, the upper fatty acid-containing hexane phase was eliminated using a separating funnel. Metadata for all samples can be found in Supplementary Table 1.

2.5. HPLC-MS/MS measurements

Dried samples were redissolved in 1 ml of methanol and centrifuged at maximum speed of 13.000 rpm for 20 min. Chromatography was performed with 5 μl per sample on a Thermo Scientific UltiMate 3000 System using a C18 column (ACQUITY UPLC BEH C18 Column, 1.7 μm, 2.1 mm X 50 mm, Waters). Acetonitrile was used as a control for blank measurements. Mass spectrometry measurements were performed on a Bruker Impact II System (Bruker Daltonik GmbH). The measuring range was 50 to 1800 *m*/*z* and the run time 22 min. All measurements were conducted in positive ionization mode.

2.6. Data analysis and feature based molecular networking

Data obtained from mass spectrometry was first converted to the open mzXML format and uploaded to MZmine (v2.53) to filter the raw data [25]. It was processed as follows: mass detection - MS level = 1, mass detector = centroid, noise level = 1000; mass detection MS level = 2, mass detector = centroid, noise level = 100; chromatogram builder (ADAP) - min group size = 5, group intensity threshold = 500, min highest intensity = 3000, *m*/*z* tolerance = 0.01 *m*/*z* or 20 ppm; chromatogram deconvolution - baseline-cutoff with min peak height = 2000, peak duration range 0.01–3.00, baseline level 1000, *m*/*z* range for MS2 0.02, RT range for MS2 0.1, *m*/*z* center calculation = MEDIAN; isotopic peak grouper

- m/z tolerance = 0.01 m/z or 20 ppm, RT tolerance = 0.1 absolute, max charge = 3, representative isotope = most intense; join aligner - m/z tolerance = 0.01 m/z or 20 ppm, weight for m/z = 75, RT tolerance = 0.1 min absolute, weight for RT 25; feature list row filter = min peaks in a row 3, minimum peaks in an isotope pattern = 2, keep only peaks with MS2 scan = yes. During sample preparation, we also performed PBS extractions as well as extraction blanks (methanol only) as controls. Prior to running the samples on the mass spectrometer, we also ran acetonitrile through the column to determine any metabolites from previous runs. The acetonitrile values were subtracted from all samples, followed by the removal of an average of the PBS blanks and an average of the extraction blanks. All values subsequently below zero were reset to zero, and any metabolite that no longer contained non-zero values was removed from the list. Filtered rows were then exported to GNPS/FBMN as MGF files.

Data were imported into the Global Natural Product Social Molecular Networking (GNPS) site [26], with networks created using the Feature Based Molecular Networking (FBMN) workflow (release 26) [27]. Data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor m/z, with further filtering to select only the top 6 fragment ions in a +/- 50 Da window of each spectrum. A precursor ion mass tolerance and MS/MS fragment ion tolerance of 0.2 Da was used. A molecular network was created with edges filtered to have a cosine score above 0.7 and more than 6 matched peaks. Edges between two nodes in the network were kept only if each of the nodes appeared in the respective others top 10 most similar nodes. A maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The analogue search mode was used by searching against MS/MS spectra with a maximum difference of 100.0 in the precursor ion value. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks. Networks were further annotated using DEREPLICATOR+ (v1.0.0), MS2LDA (release 23.1) [28] and the Network Annotation Propagation (v1.2.5) [29] as a part of the GNPS site. These results were then combined into a final annotated network using MOLNETENHANCER (release 22) [30] and visualized in Cytoscape v3.8.0.

2.7. Discriminant analysis of principal components (DAPC)

DAPC was performed using adegenet [31] (v2.1.3) in R (v4.0.3) on a data frame containing the presence or absence of each metabolite. To determine the optimal number of principal components we used 30-fold cross-validation with the *xvalDapc* function in adegenet. The aim was to optimize the trade-off between retaining too many and too few PCs by splitting the data set into training (90%) and validation (10%) sets. DAPC was then carried out on the training set with varying numbers of PCs, and the degree to which validation set members were accurately assigned to the exposed/unexposed group was measured. The optimal number of PCs was then retained for use with the *dapc.data.frame* method together with the presence/absence matrix of metabolites.

2.8. Predicting infection status

The quantified output from MZmine2 was also used to predict the infection status of triatomines. Since the quantified peaks spanned several magnitudes, we converted the values to 1 (compound present) or 0 (compound absent). We then integrated the output from the feature-based network analysis as a unique feature in each sample. Therefore, if a metabolite from a given network was present in any given sample then that network was also present in that sample. We then dropped all features that were present in less than 10% of samples as they were unlikely to meaningfully contribute to a given model. Finally, we ended up with a total of 931 features that could be used for predicting the infection status.

For logistic regression, a correlation matrix was created to determine the correlation of all features to the infection status. We then took the top correlated features (correlation greater than 0.3) for training the model. Cross-validation was performed with 5 folds, using the top 10 correlated features, top 25 correlated features or all strongly correlated features (greater than 0.3). The feature importance was then analyzed using the shap package [32]. Shapley Additive exPlanations (SHAP) is an approach to describe the output of any machine learning model based on game theory [33]. We utilized the kernel explainer function, which makes no assumptions about the model used.

In addition to logistic regression, we also trained a random forest classifier and a light gradient boosting machine (LightGBM) model using the scikit-learn [33] (v0.22.1) and lightgbm [34] (v2.3.0) packages, respectively. For random forest, hyperparameters were tuned using a grid search with 5-fold cross validation. Both the LightGBM and random forest classifier were created using 5-fold cross validation over 100 runs, with the mean accuracy reported.

2.9. Mass spectrometry search tool (MASST)

We performed single spectrum searches of important features using MASST [35], through the online workflow on the GNPS website (http://gnps.ucsd.edu). MASST searches are analogous to BLAST for protein or nucleotide sequences. With MASST, metabolite peaks are matched to databases within set thresholds. The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 2.0 Da and a MS/MS fragment ion tolerance of 0.5 Da. The library spectra were filtered in the same manner as the input data. All matches kept between input spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks.

2.10. Data availability

All code used for processing data are available at https://github.com/ntobias-85/Rprolixus_metabolites. MS data sets used for network analysis are available from the public MassIVE database with ID MSV000086832.

3. Results & discussion

3.1. Feature based molecular networking

Antunes *et al.* provided the first hints that the metabolic fingerprint of triatomines may affect vector competence for specific *T. cruzi* strains. Furthermore, it has already been shown that infections with *T. cruzi* in mammalian hosts are accompanied by distinct biochemical changes [36]. Our aim here was to further explore these tantalizing findings and determine whether it would be possible to predict the exposure status of triatomines based solely upon their metabolic signature.

The use of a high-resolution mass spectrometer generates a significant amount of information and so we aimed to summarize this using feature based molecular networking. Essentially, this process involves examining the fragmentation patterns of individual peaks within the mass spectra to determine structural relatedness and

create a synopsis of compounds detected in all samples. An added benefit to molecular networking is that we can assess the similarity to known compounds and infer compound classes from the data. The molecular networking of our mass spectrometry data resulted in 1,436 nodes, each representing a different compound and were joined by a total of 2,409 edges. Edges in this context represent a mass difference between two given metabolites (nodes) that have some structural similarity (see methods for settings used). These nodes then assembled into 177 networks consisting of two or more nodes (Fig. 1A,B). As a part of our annotation pipeline, we used MolNetEnhancer, which combined the outputs of the Feature Based Molecular Networking, MS2LDA, DEREPLICATOR+, Network Annotation Propagation tool as well as an automated chemical classification with CLASSYFIRE, and were thereby able to annotate 544 nodes with structurally similar compounds (Supplementary Table 2). Based on these annotations, we see a majority of networks are associated with lipids and lipid-like molecules, with several networks also represented by benzenoids and phenylpropanoids & polyketides (Fig. 1A,B).

The large proportion of lipids and lipid-like molecules is unsurprising given their important role in regulating fundamental metabolic process. Lipids are important as a source of energy for triatomines, but are also involved in the detoxification of heme from blood, in developmental regulation and other fundamental metabolic processes. One of these processes is the maintenance of immunity and the production of antimicrobial peptides (AMP), such as defensins, lysozymes and prolixicin. The production of prolixicin in the fat body and the midgut of *R. prolixus* results in the

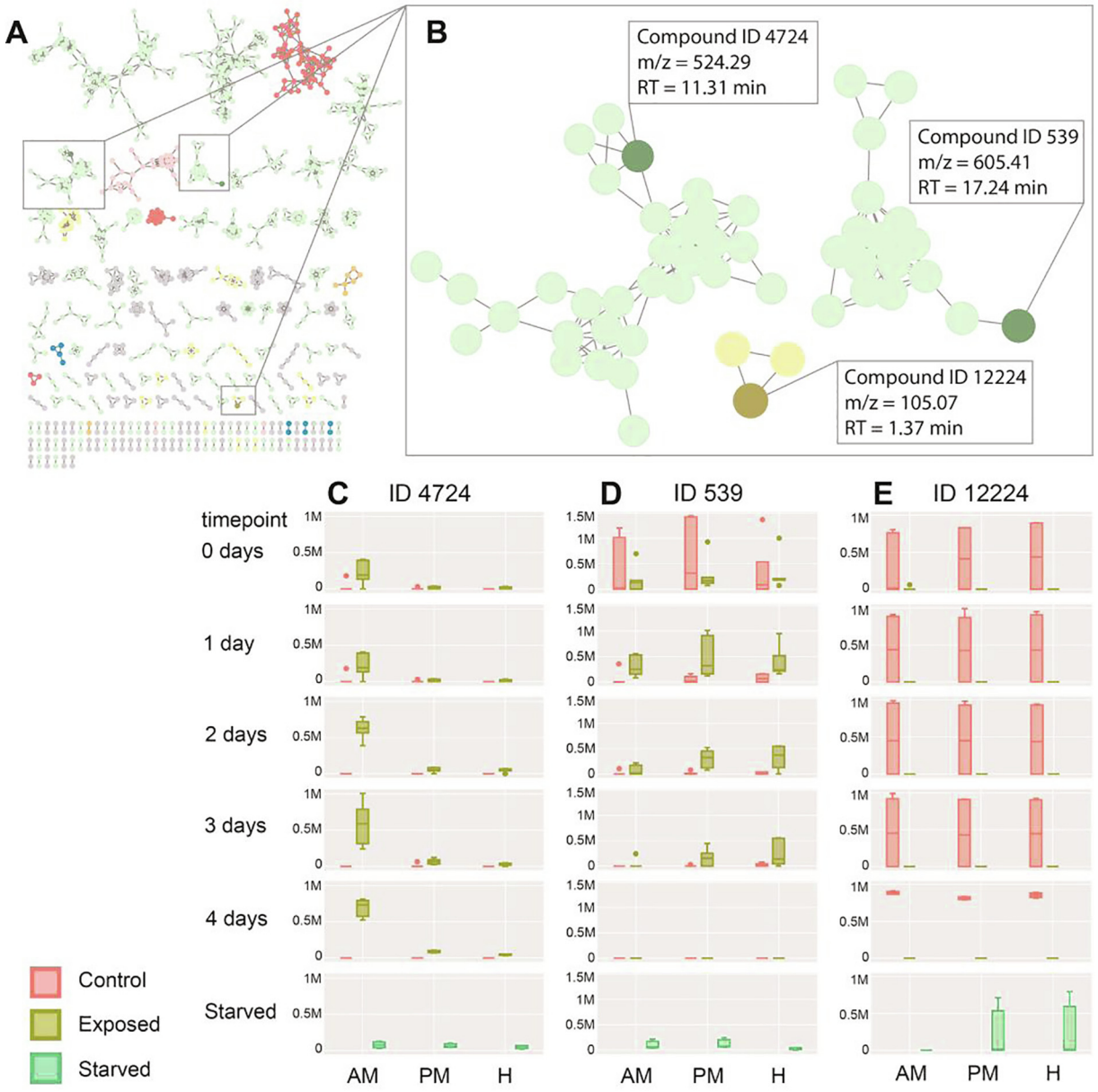

Fig. 1. A. Condensed feature based molecular network of all metabolites from all samples that assemble into a network of two or more nodes. Colors indicate the compound class of the network: lipids and lipid-like molecules (green), phenylpropanoids and polyketides (red), benzenoids (yellow), alkaloids and derivatives (orange), organoheterocyclic compounds (pink), organic acids and derivatives (blue) and not matches (grey). B Also indicated are the top three key compounds identified by supervised machine learning, C-E with metabolite abundance of these top three compounds (feature IDs 4724, 539 and 1224) in exposed group, control group and starved insects, broken down by gut compartment (AM: anterior midgut, PM: posterior midgut, H: hindgut) and time point (in days). Box plots show interquartile range, with median, minimum, maximum and outliers shown. All plots were generated with the GNPS Feature Based Molecular Networking dashboard (v0.1). Further details of these metabolites can be found in Supplementary Table 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

modulation of bacterial microbiota and is altered by the infection status and the trypanosomal pathogen (e.g. *T. cruzi* Dm 28c, *T. cruzi* Y strain, *T. rangeli*) [41–39]. Another group of immunologically relevant compounds are the eicosanoids, which are fatty acid derivatives and mainly synthesized from arachidonic acid. These compounds act as lipid mediators in insects driving specific cell reactions to pathogen invasion, including phagocytosis, microaggregation, nodulation and hemocytes activity [22,37,40].

3.2. Exposure-dependent differences in metabolic profiles

Since insects were not individually tested for colonization by *T. cruzi*, we will refer to insects as being exposed, as opposed to colonized or infected. In order to determine general changes in metabolic profiles, initial investigation of differences between exposed and non-exposed insects was carried out using the discriminant analysis of principal components (DAPC). This method for analysis of metabolic studies is limited. However, its ability to derive biologically meaningful insights with respect to different chemotype classes is well documented [40–38]. We determined the optimal number of principal components (PCs) to use in the DAPC analysis by using cross-validation implemented by the *xvalDapc* function of the adegenet package (Supplementary Fig. 1). We used 20 PCs for the DAPC analysis, which corresponded with the lowest root mean squared error in the cross-validation data (Supplementary Fig. 1). Twenty PCs explained 59.2% of variance (Supplementary Fig. 2) and resulted in a mean successful assignment of individuals into the exposed or unexposed groups 80.7% of the time, using cross-validation. We observed clear separation of exposed and unexposed insects based on all detected features (Fig. 2), suggesting that feeding on blood infected by *T. cruzi* is sufficient to alter the metabolic profile of insects in a detectable way, irrespective of whether the insect is ultimately colonized by the parasite. This could be a consequence of the microbiota changed by the presence of *T. cruzi*, while it could also be a reaction of the insect itself. If these results are a direct immunological reaction of the insect to *T. cruzi*, it would be important to rule out cross-reactions in further studies. For example, *R. prolixus* can also be infected by other trypanosomatid parasites such as *Trypanosoma rangeli,* which has adverse effects to its insect vector and might trigger divergent reactions [41]. In order to prevent false positive results and to establish the effects of *T. rangeli* on the metabolome, triatomines infected with *T. rangeli* should be included in future studies.

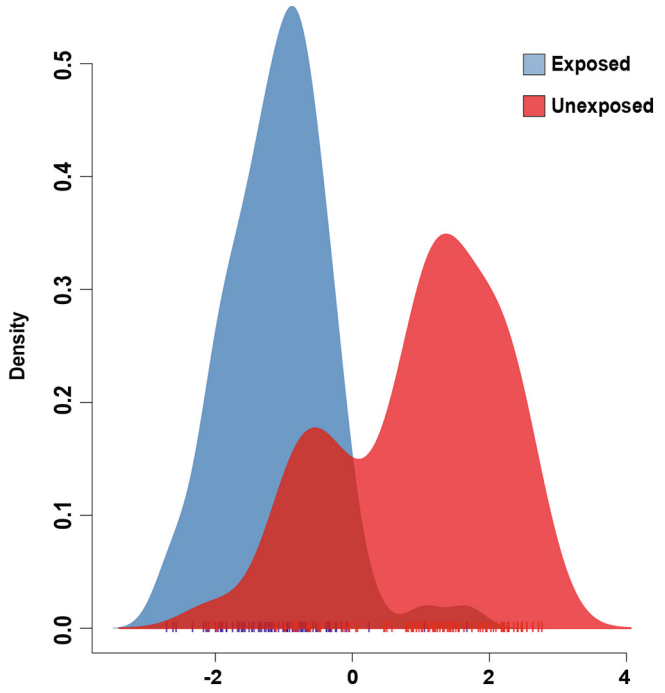

Fig. 2. Discriminant analysis of principal components for exposed and non-exposed samples. Taking into account all detected molecules from all samples, a clear separation between both groups (*T. cruzi*-exposed and unexposed) is apparent, indicating a change in the metabolic profile. Vertical lines beneath the plot represent individual samples.

3.3. Predictive models of exposure status

We wondered then if the differences are the result of the presence or absence of a specific combination of metabolites and whether we might be able to accurately predict exposure status using machine learning. To investigate, we took the output from MZmine2, which detailed the mass to charge ratio (m/z), retention time and intensity for each sample. The details from the network analysis from GNPS were also added to each table, where the presence of a metabolite in a network, also meant that network was present in a given sample. 908 metabolites were identified as being part of a network (two or more metabolites), while 528 metabolites were not identified as being structurally related to anything. To reduce computational processing time and the likelihood of overfitting the model, we also removed any metabolite that was present in less than 10% of the 171 samples, since they were unlikely to contribute meaningfully to developing a model. Finally, metadata regarding gut section, time point after feeding and parasite type (trypomastigote or epimastigote) was added to our feature table. After data cleaning, we were left with 932 features across 171 samples. Based upon these cleaned data, we utilized three different supervised machine learning algorithms to predict the exposure status of insects.

Of the 171 samples, 90 contained results from unexposed insects, while 81 were exposed. We began by developing a logistic regression model using highly-correlated features to exposure status. These were initially limited to the top 10, top 25 or any features with a correlation score higher than 0.3 (top 51). We used 5-fold cross-validation on our dataset while developing a logistic regression model, which returned average accuracies of 88.9% (top 10), 93.6% (top 25) and 93.0% (top 51). We also employed a gradient boosting machine model and random forest, similarly using 5-fold cross validation, which returned average accuracies of 91.8% and 94.2%, respectively. Confusion matrices were produced for each of the models detailing false positive, false negative, true positive and true negative values (Supplementary Figure 3). Next, we investigated the top 25 features from each of the models by extracting them using the shap package, which is a model agnostic method for determining feature importance. This step was important to us as we wanted to know whether independent models were performing well due to the presence (or absence) of an individual feature. All features that each of the three models agreed were important are listed in Table 1.

In our current study, we utilized a total of 171 samples (81 infected and 90 uninfected, including 9 starved) from *R. prolixus.* Our models predicting the exposure status of insects performed with average accuracies of between 91.8 and 94.2%. Interestingly, the most commonly misclassified samples were those from our unexposed, starved samples, which tended to result in false positives. This is perhaps unsurprising since only nine samples were included for starved samples. Generally, having more observations should improve the performance of models. In the future we will place more emphasis on increased numbers of observations, over longer time frames as the complete digestion of a blood meal takes 12–14 days in adult *R. prolixus*. Afterwards, the insects are able to withstand prolonged periods of starvation resulting in the loss of the blood-induced increased microbial diversity in the intestinal

Table 1
Summary of important group discriminating compounds (feature IDs) agreed upon by each of our three models, including the mass to charge ratio (*m*/*z*), average retention time and network. Also indicated are the top MASST compound annotation matches. Further details of MASST matches can be found in Supplementary Table 3.

Feature ID	*m*/*z*	Retention time (min)	Network index	Top MASST Result		
				MSV ID	Source	Cosine score
4724	524.29	11.31	7	MSV000083446	Mice tissue	0.93
12,224	105.07	1.37	–	–	–	–
539	605.41	17.24	–	–	–	–
4000	548.29	12.87	2	MSV000080655	Sputum samples	0.97
184	607.39	16.71	11	MSV000086109	*Moorena bouillonii*	0.85
12,240	462.28	9.38	–	–	–	–
6800	595.38	11.28	–	–	–	–
7453	459.30	15.94	11	MSV000080050	Human and rat stool sample	0.84
6828	559.36	14.77	7	–	–	–

tract [46–44]. By using longer time frames, we might be able to see metabolic changes in exposed (and infected) triatomines without the interfering effects of blood uptake. Also, the immune reaction of the insect only reaches its peak after 5 to 9 days after exposure, which may also influence results [21,45,46].

3.4. Feature investigation and origins of key compounds

By exploring the reasoning behind exposure status classification for a given sample, we were able to elucidate the contribution that each compound makes (see Supplementary Fig. 4 for an example). Looking at feature contributions allowed us to develop an understanding of how large an impact any one metabolite plays in the prediction. To break this down further, we explored the abundance of these top features in different gut compartments (Fig. 1B-E). Interestingly, we see clear differences in gut compartment, infection status and time point. In part, this helps to explain why there was no one single metabolite that lead to an accurate prediction of exposure, but a series of metabolites. Intuitively this also makes sense, since the different gut compartments undertake different metabolic functions and we expect them to contain different metabolites [47]. For example, metabolite with compound ID 4724 was much more abundant in the anterior midgut of exposed insects (Fig. 1B,C). On the other hand, compound ID 539 was occasionally present at time point zero of unexposed insects but rapidly decreased from day 1 in unexposed insects, in all gut compartments (Fig. 1B,D). There are also clear differences in the abundance of compounds between epimastigote and trypomastigote *T. cruzi* exposed triatomines. For example, the molecule with feature ID 4000 is considerably more present in epimastigote exposed insects. Taking a higher level look at just the gut compartment and infection status, we see similar differences, although not always as pronounced (Supplementary Fig. 5). This demonstrates that the occurring changes can be traced back to several different compounds. Each metabolite can be explored in detail through http://dorresteintesthub.ucsd.edu:6549/?task=5a6231ed03724b088fa0055e33397038 using the feature IDs present in Supplementary Table 2.

Recently, the development of the Mass Search Tool (MASST) has opened up the possibility to search unknown metabolites in compound databases in an analogous way to nucleotide or protein sequences using BLAST. We explored our top-ranking features for hits in compound databases (Table 1, Supplementary Table 3). Only four of the top hits had similarity in the database demonstrating the difficulties of annotating untargeted metabolomics data. Three of these matched to studies involving human or mouse/rat, while the fourth matched a sugarcane-microbe interaction study. This may indicate that the metabolic signature detected in our study is a result of ingesting exposed blood, rather than an insect specific finding. However, since the *T. cruzi* epimastigotes were added to the blood in the artificial feeder and do not originate from an infected mouse, this blood should not contain compounds produced by the mammal host in reaction to a Chagas infestation. In addition, comparing exposed and control samples, the differences in abundance of the most highly rated machine learning features were caused by trypomastigote as well as epimastigote *T. cruzi*. Nevertheless, to determine whether the ingested blood has a substantial influence on the results, studies on other triatomine species would be extremely beneficial as this may indicate a more widespread trend that could be detected. Also, the metabolomic screening of triatomines with a verified persistent *T. cruzi* infection and without a previous blood meal might be advantageous as this would minimize the effects of metabolites originating from ingested blood.

4. Concluding remarks

Our results show that the exposure of triatomine bugs with the parasite *T. cruzi* leads to a change in the composition of metabolites in the insects' intestinal tract. These differences in the metabolic signature can be used to determine whether or not an insect has been exposed to the parasite with up to 94.2% accuracy. Furthermore, it is possible to narrow down the differences to specific compounds shaping the metabolome of infected triatomines. These key compounds may prove to be robust biomarkers of *T. cruzi* exposure in *R. prolixus*, but also in other triatomine species.

CRediT authorship contribution statement

Fanny E. Eberhard: Investigation, Writing - review & editing, Funding acquisition. **Sven Klimpel:** Resources, Writing - review & editing. **Alessandra A. Guarneri:** Conceptualization, Methodology, Investigation, Resources, Writing - review & editing, Funding acquisition. **Nicholas J. Tobias:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing - original draft, Visualization, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Helge Bode for helpful discussions regarding methodology and for providing access to mass spectrometry facilities. This work was funded by the LOEWE-Centre TBG supported by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, CRA-APQ-00569-15 and

CRA-PPM-00162-17), Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular (INCTEM/CNPq, 465678/2014-9) and the Vereinigung von Freunden und Förderern der Johann Wolfgang Goethe-Universität Frankfurt am Main, AAG was supported by CNPq productivity grants.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.05.027.

References

[1] Lidani KCF et al. Chagas Disease: From Discovery to a Worldwide Health Problem. Front Public Health 2019;7:166.
[2] Vianna Martins A et al. Biology of *Trypanosoma cruzi*: An update. Infectio 2012;16:45–58.
[3] Olivera MJ, Buitrago G. Economic costs of Chagas disease in Colombia in 2017: A social perspective. Int J Infect Dis 2020;91:196–201.
[4] Ozaki Y, Guariento ME, de Almeida EA. Quality of life and depressive symptoms in Chagas disease patients. Qual Life Res 2011;20:133–8.
[5] Botto-Mahan D. *Trypanosoma cruzi* induces life-history trait changes in the wild kissing bug *Mepraia spinolai*: implications for parasite transmission. Vector Borne Zoonotic Disease 2009;5:505–10.
[6] Cordero-Montoya G, Flores-Villegas AL, Salazar-Schettino PM, Vences-Blanco MO, Rocha-Ortega M, Gutiérrez-Cabrera AE, et al. The cost of being a killer's accomplice: *Trypanosoma cruzi* impairs the fitness of kissing bugs. Parasitol Res 2019;118:2523–9.
[7] Fellet MR, Lorenzo MG, Elliot SL, Carrasco D, Guarneri AA. Effects of Infection by *Trypanosoma cruzi* and *Trypanosoma rangeli* on the Reproductive Performance of the Vector *Rhodnius prolixus*. PLoS ONE 2014;9:e105255.
[8] Botto-Mahan C, Campos V, Medel R. Sex-dependent infection causes nonadditive effects on kissing bug fecundity. Ecol Evol 2017;7:3552–7.
[9] Elliot SL, Rodrigues,, J.d.O.,, Lorenzo,, M.G.,, Martins-Filho,, O.A.,, Guarneri,, A.A.. *Trypanosoma cruzi*, etiological agent of Chagas Disease, is virulent to its triatomine vector *Rhodnius prolixus* in a temperature-dependent manner. PLoS NeglTrop Dis 2015;9:e0003646.
[10] Villalobos G, Nava-Bolaños A, De Fuentes-Vicente JA, et al. A reduction in ecological niche for *Trypanosoma cruzi*-infected triatomine bugs. Parasites Vectors 2019;12:240.
[11] Pereyra N, Lobbia P, Mougabure-Cueto G. Effects of the infection with *Trypanosoma cruzi* on the feeding and excretion/defecation patterns of *Triatoma infestans*. Bull Entomol Res 2020;110:169–76.
[12] Verly T, Costa S, Lima N, Mallet J, Odêncio F, et al. Vector competence and feeding-excretion behavior of *Triatoma rubrovaria* (Blanchard, 1843) (Hemiptera: Reduviidae) infected with *Trypanosoma cruzi* TcVI. PLoS NeglTrop Dis 2020;14:e0008712.
[13] Dumonteil E et al. Interactions among *Triatoma sanguisuga* blood feeding sources, gut microbiota and *Trypanosoma cruzi* diversity in southern Louisiana. Mol Ecol 2020;29:3747–61.
[14] Waltmann A et al. Hindgut microbiota in laboratory-reared and wild *Triatoma infestans*. PLoS Negl Trop Dis 2019;13:e0007383.
[15] Orantes LC et al. Uncovering vector, parasite, blood meal and microbiome patterns from mixed-DNA specimens of the Chagas disease vector *Triatoma dimidiata*. PLoS Negl Trop Dis 2018;12:e0006730.
[16] Díaz S, Villavicencio B, Correia N, Costa J, Haag KL. Triatomine bugs, their microbiota and *Trypanosoma cruzi*: asymmetric responses of bacteria to an infected blood meal. Parasit Vectors 2016;9:45.
[17] Castro DP et al. *Trypanosoma cruzi* immune response modulation decreases microbiota in *Rhodnius prolixus* gut and is crucial for parasite survival and development. PLoS ONE 2012;7:e36591.
[18] Antunes LCM et al. Metabolic Signatures of Triatomine Vectors of *Trypanosoma cruzi* Unveiled by Metabolomics. PLoS ONE 2013;8:e77283.
[19] Tobias NJ, Eberhard FE, Guarneri AA. Enzymatic biosynthesis of B-complex vitamins is supplied by diverse microbiota in the *Rhodnius prolixus* anterior midgut following *Trypanosoma cruzi* infection. Comput Struct Biotechnol J 2020;18:3395–401.
[20] Salcedo-Porras N, Umaña-Diaz C, de O. B. Bitencourt R, Lowenberger C. The Role of Bacterial Symbionts in Triatomines: An Evolutionary Perspective. *Microorganisms* 2020;8:1438.
[21] Garcia ES, Genta FA, de Azambuja P, Schaub GA. Interactions between intestinal compounds of triatomines and *Trypanosoma cruzi*. Trends Parasitol 2010;26:499–505.
[22] Azambuja P, Garcia ES, Ratcliffe NA. Gut microbiota and parasite transmission by insect vectors. Trends Parasitol 2005;21:568–72.
[23] Contreras VT et al. Biological aspects of the Dm 28c clone of *Trypanosoma cruzi* after metacyclogenesis in chemically defined media. Memórias do Instituto Oswaldo Cruz 1988;83:123–33.
[24] Guarneri AA. in *Trypanosomatids* **2116,** 69–79. New York, NY: Humana; 2020.
[25] Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf 2010;11:395–411.
[26] Wang M et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. 2016;34:828–37.
[27] Nothias L-F et al. Feature-based molecular networking in the GNPS analysis environment. Nat. Methods 2020;17:905–8.
[28] van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. Proc. Natl. Acad. Sci. U.S.A. 2016;113:13738–43.
[29] Da Silva RR et al. Propagating annotations of molecular networks using in silico fragmentation. PLoS Comput Biol 2018;14:e1006089.
[30] Ernst M et al. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites 2019;9:144.
[31] Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 2008;24:1403–5.
[32] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. 4765–4774 (2017).
[33] Pedregosa F et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.
[34] Ke, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in (eds. Guyon, I. et al.) 30, 3146–3154 (Curran Associates, Inc., 2017).
[35] Wang M et al. Mass spectrometry searches using MASST. Nat. Biotechnol. 2020;38:23–6.
[36] Gironès N et al. Global metabolomic profiling of acute myocarditis caused by *Trypanosoma cruzi* infection. PLoS Negl Trop Dis 2014;8:e3337.
[37] Salcedo-Porras N, Guarneri A, Oliveira PL, Lowenberger C. *Rhodnius prolixus*: Identification of missing components of the IMD immune signaling pathway and functional characterization of its role in eliminating bacteria. PLoS ONE 2019;14:e0214794.
[38] Vieira CS et al. Impact of *Trypanosoma cruzi* on antimicrobial peptide gene expression and activity in the fat body and midgut of *Rhodnius prolixus*. Parasit Vectors 2016;9:119–212.
[39] Ursic-Bedoya R, Buchhop J, Joy JB, Durvasula R, Lowenberger C. Prolixicin: a novel antimicrobial peptide isolated from *Rhodnius prolixus* with differential activity against bacteria and *Trypanosoma cruzi*. Insect Mol Biol 2011;20:775–86.
[40] Stanley D, Miller J, Tunaz H. Eicosanoid actions in insect immunity. J Innate Immun 2009;1:282–90.
[41] Guarneri AA, Lorenzo MG. Triatomine physiology in the context of trypanosome infection. J Insect Physiol 2017;97:66–76.
[42] Grillo LAM, Majerowicz D, Gondim KC. Lipid metabolism in *Rhodnius prolixus* (Hemiptera: Reduviidae): role of a midgut triacylglycerol-lipase. Insect Biochem Mol Biol 2007;37:579–88.
[43] Almeida CE, Francischetti CN, Pacheco RS, Costa J. *Triatoma rubrovaria* (Blanchard, 1843) (Hemiptera-Reduviidae-Triatominae) III: patterns of feeding, defecation and resistance to starvation. Memórias do Instituto Oswaldo Cruz 2003;98:367–71.
[44] Cortéz MG, Gonçalves TC. Resistance to starvation of *Triatoma rubrofasciata* (De Geer, 1773) under laboratory conditions (Hemiptera: Reduviidae: Triatominae). Memórias do Instituto Oswaldo Cruz 1998;93:549–54.
[45] Batista KKDS et al. Nitric oxide effects on *Rhodnius prolixus*'s immune responses, gut microbiota and *Trypanosoma cruzi* development. J Insect Physiol 2020;126:104100.
[46] Genta FA, Souza RS, Garcia ES, Azambuja P. Phenol oxidases from *Rhodnius prolixus*: temporal and tissue expression pattern and regulation by ecdysone. J Insect Physiol 2010;56:1253–9.
[47] Terra WR, Ferreira C. Insect digestive enzymes: properties, compartmentalization and function. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 1994;109:1–62.