pharmaceuticals

MDPI

Article

Relevant Features of Polypharmacologic Human-Target Antimicrobials Discovered by Machine-Learning Techniques

Rodrigo A. Nava Lara [1], Jesús A. Beltrán [2], Carlos A. Brizuela [2] and Gabriel Del Rio [1,*]

[1] Department of Biochemistry and Structural Biology, Instituto de Fisiologia Celular, UNAM, Mexico City 04510, Mexico; Rodrigo_Andres1993@hotmail.com
[2] Department of Computer Science, CICESE Research Center, Ensenada 22860, Mexico; armando.3eltran@gmail.com (J.A.B.); cbrizuel@cicese.edu.mx (C.A.B.)
* Correspondence: gdelrio@ifc.unam.mx

Received: 11 July 2020; Accepted: 7 August 2020; Published: 21 August 2020

Abstract: Polypharmacologic human-targeted antimicrobials (polyHAM) are potentially useful in the treatment of complex human diseases where the microbiome is important (e.g., diabetes, hypertension). We previously reported a machine-learning approach to identify polyHAM from FDA-approved human targeted drugs using a heterologous approach (training with peptides and non-peptide compounds). Here we discover that polyHAM are more likely to be found among antimicrobials displaying a broad-spectrum antibiotic activity and that topological, but not chemical features, are most informative to classify this activity. A heterologous machine-learning approach was trained with broad-spectrum antimicrobials and tested with human metabolites; these metabolites were labeled as antimicrobials or non-antimicrobials based on a naïve text-mining approach. Human metabolites are not commonly recognized as antimicrobials yet circulate in the human body where microbes are found and our heterologous model was able to classify those with antimicrobial activity. These results provide the basis to develop applications aimed to design human diets that purposely alter metabolic compounds proportions as a way to control human microbiome.

Keywords: polypharmacological compounds; heterologous machine learning; broad-spectrum antibiotics; human metabolites

1. Introduction

Drug discovery nowadays involves high-throughput screenings of compounds with or without knowledge of the molecular target to treat a condition or disease [1–3]; after conducting in vitro and in vivo experiments (e.g., toxicity and efficiency tests), drugs are ready to be tested on humans. It has been noted that although the mechanism of action for some drugs is well studied (e.g., inhibitors of GPCRs [4], protein kinases [5] and penicillin-binding proteins [6]) for many others this is not the case [7]. In fact, it has been recognized that the pharmacological activity of many FDA-approved drugs depends on having multiple targets [8]. This situation has led to drug re-purposing (e.g., viagra [9], aspirin [10]).

In parallel with drug discovery efforts, researchers in biomedicine have discovered the prevalent role of the gut microbiome in human health [11,12]. For instance, the unsupervised antibiotic consumption can induce dysbiosis in gut microbiome that has been associated with celiac disease [13,14] or inflammatory bowel disease [15,16], among others [17].

Considering these two scenarios, polypharmacologic drugs and human microbiome, it has been argued that some FDA-approved human-targeted drugs (FHD) may act through a secondary

antimicrobial activity [18]; that is, since the microbiome control different aspects of human health, some drugs that are acting as their primary target on a human protein may also have an antimicrobial activity as a secondary activity. To support these observations, a high-throughput drug screening on human gut microbes was performed using FHD by Maiers and collaborators [19]; 24% of FHD had a secondary antimicrobial activity. The authors noted that antimetabolites and antipsychotics were enriched in 24% of antimicrobial FHD, and found few previous reports about the antibacterial activity on a particular class of antipsychotics. While Maiers and collaborators were concerned about resistance gained by infectious agents exposed to these polypharmacological compounds, here we focus on a different aspect not explored before, the use of antimicrobials as a source to identify polypharmacological compounds. We will focus in this study to polypharmacological compounds that act on human and on microbial molecular targets, which, from now on, we will refer to as polypharmacologic Human-targeted AntiMicrobials or polyHAMWe have previously reported that using peptide and non-peptide antimicrobial compounds (heterologous training set) was an effective method to identify polyHAM from FDA human-targeted drugs [20]. Here we show that polyHAM are more likely found among antimicrobials presenting antibiotic activity against multiple bacterial strains than from antimicrobials acting against a single microbe. Since metabolism is controlled by and controls the microbiome, we tested this model in identifying human metabolites with antimicrobial activity with reliable results. Thus, we report features relevant for the activity of polyHAM compounds that are also found among human metabolites. The implications about these findings in the diagnosis and/or possible treatment of complex diseases are discussed.

2. Results

We built three different training datasets: anti-infective, anti-gut1 and anti-gut4 (see Methods). All these sets include the 1181 FDA-approved compounds previously tested by Maier and collaborators for gut antimicrobial activity; consequently, some compounds will change their classification as antimicrobial or non-antimicrobials in these groups as shown in Figure 1. Figure 1A shows the 1096 non-antimicrobials annotated in the three groups, while Figure 1B shows 436 antimicrobials annotated in those same three groups, indicating there are 351 compounds (1532–1181) that are interchanged between antimicrobials and non-antimicrobials in these three groups. The anti-infective set includes 137 compounds annotated as antimicrobials (see Supplementary File S1) by the Anatomical Therapeutic Chemical (ATC) classification [21]; the anti-gut1 set includes 398 compounds (see Supplementary File S2) found to have antimicrobial activity against one or more gut bacterial strains; the anti-gut4 set includes 255 compounds (see Supplementary File S3) with antimicrobial activity against four or more gut bacterial strains; these 255 compounds are also antimicrobials of the anti-gut1 set. In Supplementary Files S1–S3, antimicrobials are instances of class 1 and non-antimicrobials of class 0. The anti-infective is the control group by not including antimicrobial compounds with a known human target so far, while anti-gut1 and anti-gut4 include multifunctional compounds, that is, antimicrobials with a human target. We expect that the number of antimicrobials with human targets would increase from anti-gut1 to anti-gut4, if broad antimicrobial activity would imply acting on multiple targets. Anti-gut1 or anti-infective, on the other hand, are sets with less broad antimicrobial activity than those found in anti-gut4 set. These three sets were used along this study to either: (i) identify all publications on PubMed for those compounds with antimicrobial ontological annotations and, (ii) characterize and classify antimicrobials from non-antimicrobials using molecular features (see Methods).

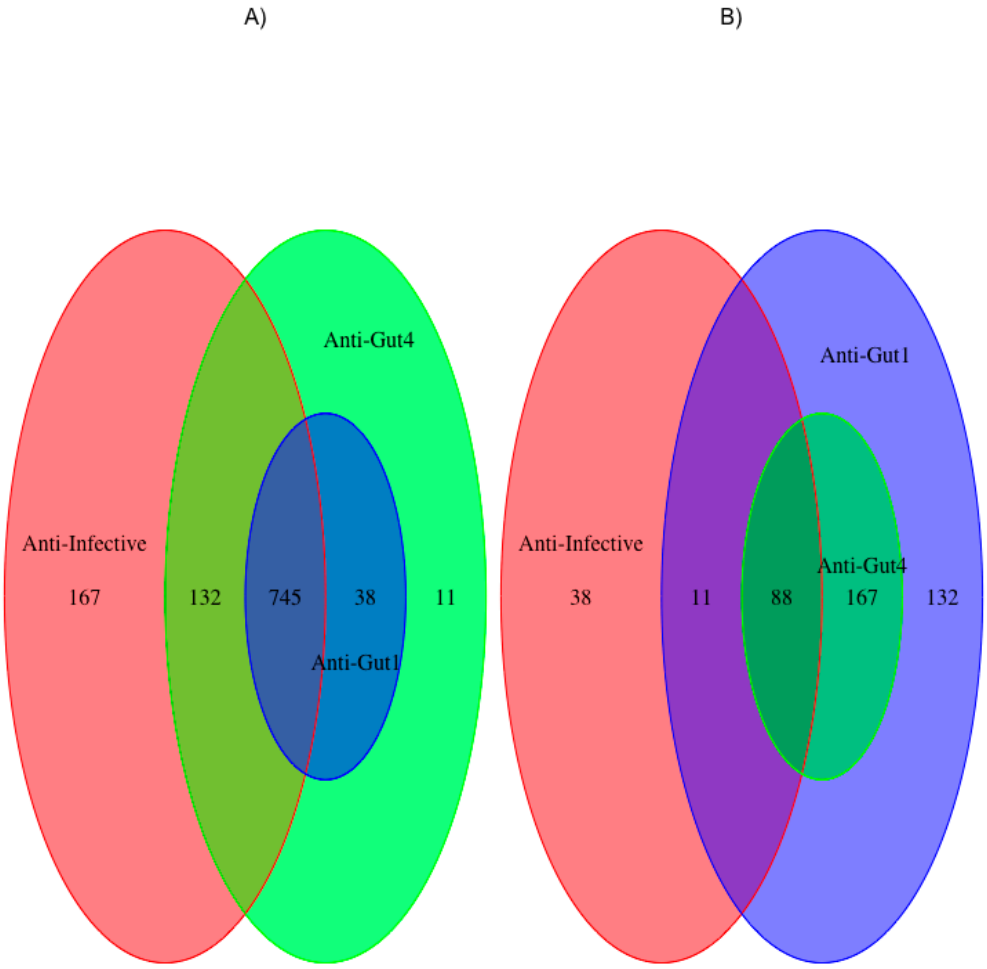

Figure 1. Datasets. The three datasets used throughout this work, including 1181 FDA-approved drugs. Anti-infective set (red ovals) includes 137 compounds that had been classified by the ATC as anti-infective, anti-gut1 set (blue ovals) includes 398 compounds with antimicrobial activity against 1 or more gut bacterial strains, anti-gut4 set (green ovals) includes 255 compounds with antimicrobial activity against four or more gut bacterial strains; the actual compound names and chemical formulas are reported in Supplementary Files S1–S3, respectively. These 3 sets are derived from the same set of compounds hence share some overlaps that are presented in Venn diagrams for (**A**) non-antimicrobials and (**B**) antimicrobials.

2.1. Publications about Antimicrobial Activity for polyHAM

To identify how many of the compounds with a human target studied by Maier and collaborators had previous publications as antimicrobials, we designed a naive text-mining procedure to identify publications reporting antimicrobial activity for a chemical compound (see Methods). This identification was based on ontological terms associated with broad antimicrobial activity (see Table 1); these ontological terms were identified from each PubMed entry in MedLine format (see Methods). The proportion of publications including antimicrobial terms is presented in Figure 2A–C; this proportion is derived by dividing the number of publications with antimicrobial-related ontological terms by the total number of publications reported (see Methods) for every FHD studied by Maier and collaborators. These results indicate that the broader the antimicrobial activity the more frequent antimicrobial publications: the proportion of antimicrobial publications for dataset anti-gut4 is larger than for anti-gut1 as expected (compare the mean values presented by the black line crossing the boxplots in Figure 2B,C). Although the antimicrobials in the anti-infective set have more publications with ontological terms related to antimicrobial activity than the other training sets, several compounds

in the anti-gut4 have more publications with ontological terms related to antimicrobial activity than those found in the anti-infective set, confirming that anti-gut4 set includes compounds with more studied antimicrobial activity than the ones included in the anti-infective or anti-gut1 sets, as expected.

Table 1. Antimicrobial ontological terms.

Antifungal agents pharmacology	Antifungal agents administration and dosage	Antifungal agents administration and dosage therapeutic use	Antifungal agents chemical synthesis chemistry pharmacology
Antifungal agents therapeutic use	Drug resistance fungal	Fungi drug effects	Virus replication drug effects
Antiviral agents therapeutic use	Anti-bacterial agents analysis	Anti-bacterial agents pharmacology	Anti-bacterial agents therapeutic use
Drug resistance bacterial	Drug-resistance multiple bacterial	*Mycobacterium tuberculosis* drug effects	Anti-bacterial agents
Anti-bacterial agents administration and dosage adverse effects	Anti-bacterial agents administration and dosage pharmacology	Anti-bacterial agents administration and dosage therapeutic-use	Anti-bacterial agents adverse-effects
Anti-bacterial agents chemistry	Anti-bacterial agents pharmacology therapeutic-use	Anti-bacterial agents toxicity	Bacterial infections drug-therapy
DNA-bacterial genetics	Drug-resistance bacterial-genetics	Gram-negative-bacteria drug-effects	Gram-positive-bacteria drug-effects
Helicobacter infection drug-therapy microbiology	*Helicobacter pylori*	*Helicobacter pylori* drug-effects	*Mycobacterium avium* complex-drug-effects

The actual data summarized in Figure 2 are in supplementary Table S1. Table S2 includes the 16 compounds (out of 203 reported by Maier and collaborators as FHD) that are known to act through a human target having two or more publications with ontological terms associated to broad antimicrobial activity. Our results indicate that the use of ontological terms related to antimicrobials did not identify every polyHAM analyzed. While it is possible to use other terms to identify antimicrobial compounds, our aim here was to explore the broad-spectrum activity of the antimicrobial activity.

2.2. Classifying Antimicrobials Using Physicochemical Features

We have previously described a procedure to increase the training set size for antimicrobial compound classification and consequently increase the reliability of predictions; we referred to this procedure as heterologous machine learning, because sets of peptides and non-peptide chemical compounds are used to train models that efficiently classified antimicrobial from non-antimicrobial compounds [20]. Here we further explored this approach to classify antimicrobial compounds in the anti-infective, anti-gut1 and anti-gut4 sets; for that end, we added 7999 antimicrobial peptides and 3546 non-antimicrobial peptides to each of these sets for the heterologous training set construction (see Methods). Our aim is to compare the classification efficiency of antimicrobial compounds using molecular features relevant for heterologous machine learning with that achieved using ontological terms.

To achieve this goal, we characterized the applicability domain of any model derived from these datasets using two general aspects of peptides and non-peptidic chemical compounds. Figure 3 shows the comparison in molecular weight observed between these two sets; as expected, peptides tend to be larger than non-peptidic chemical compounds, ranging from 100 up to 6600 Daltons, with peaks at 300–400, 3000, and 4500 Daltons.

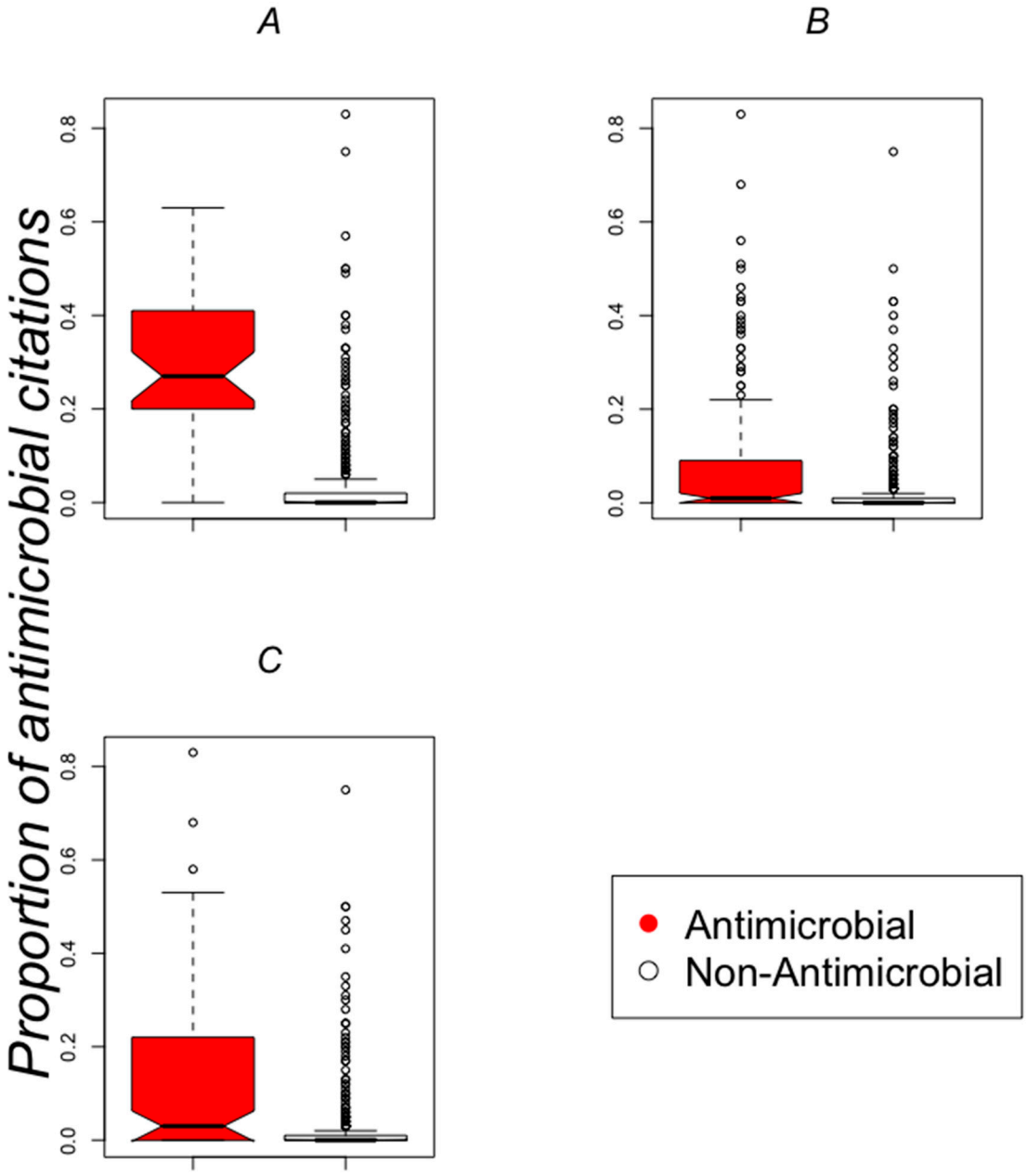

Figure 2. PubMed citations with ontological terms associated with antimicrobial activity. (**A**) Anti-infective, (**B**) anti-gut1 and (**C**) anti-gut4 data sets. The images present the distribution of the proportion of publications that included an ontological term related to antimicrobial activity with respect to the total number of publications reported for every compound tested by Maier and collaborators; to see the actual ontological terms, please refer to Table 1. The data are presented in box plots, where the first and third quartile are represented below and above the black line that corresponds with the mean value of the distribution. Every box plot presents the data for each of the 572 compounds that presented at least one publication with ontological terms associated with antimicrobial activity, out of the 1181 compounds tested by Maier and collaborators. Red boxes represent antimicrobials, white otherwise.

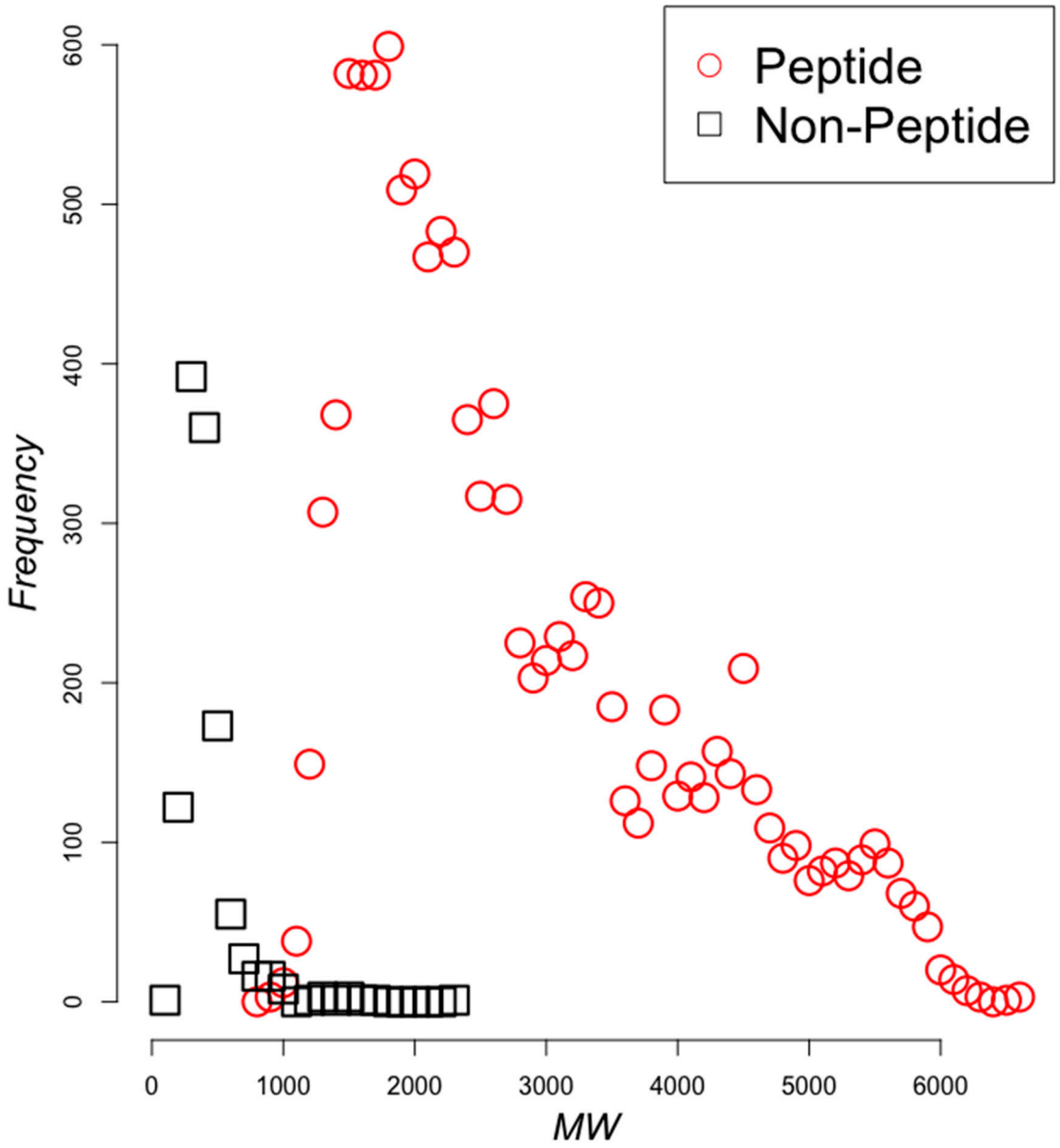

Figure 3. Molecular weights comparison between peptides and non-peptides in the training set. Red squares represent peptides and non-peptides are black squares. The y-axis presents the observed frequency of molecules within the specified range of molecular weights (x-axis). Molecular weights were accumulated in bins of size 100 Da (0–100, 101–200, . . . , 6401–6500, 6501–6600). Thus, every symbols correspond with the number of molecules observed every range of 100 Da in molecular weight.

Identifying chemical functional groups from a chemical formula is not a standard or a trivial matter [22]. Indeed, it has been recently noted that no automatic procedure to accomplish this goal leading to the development of a machine-learning-based approach for that goal (see Methods); this machine learning implementation, however, only detects chemical groups that are too broad. For instance, we identified 274 chemical groups in both peptides and non-peptides (see Methods) and observed some similarities in the frequency observed of some of these chemical groups (oxygen, nitrogen, nitrogen aromatic, sulfur, acid, amide; see Figure 4). These chemical groups are different from the features calculated by PaDel-descriptor [23]. To compare the frequency of occurrence of the detected functional group, percentages reported in Figure 5 are normalized per total number of peptides or total number of non-peptidic chemical compounds.

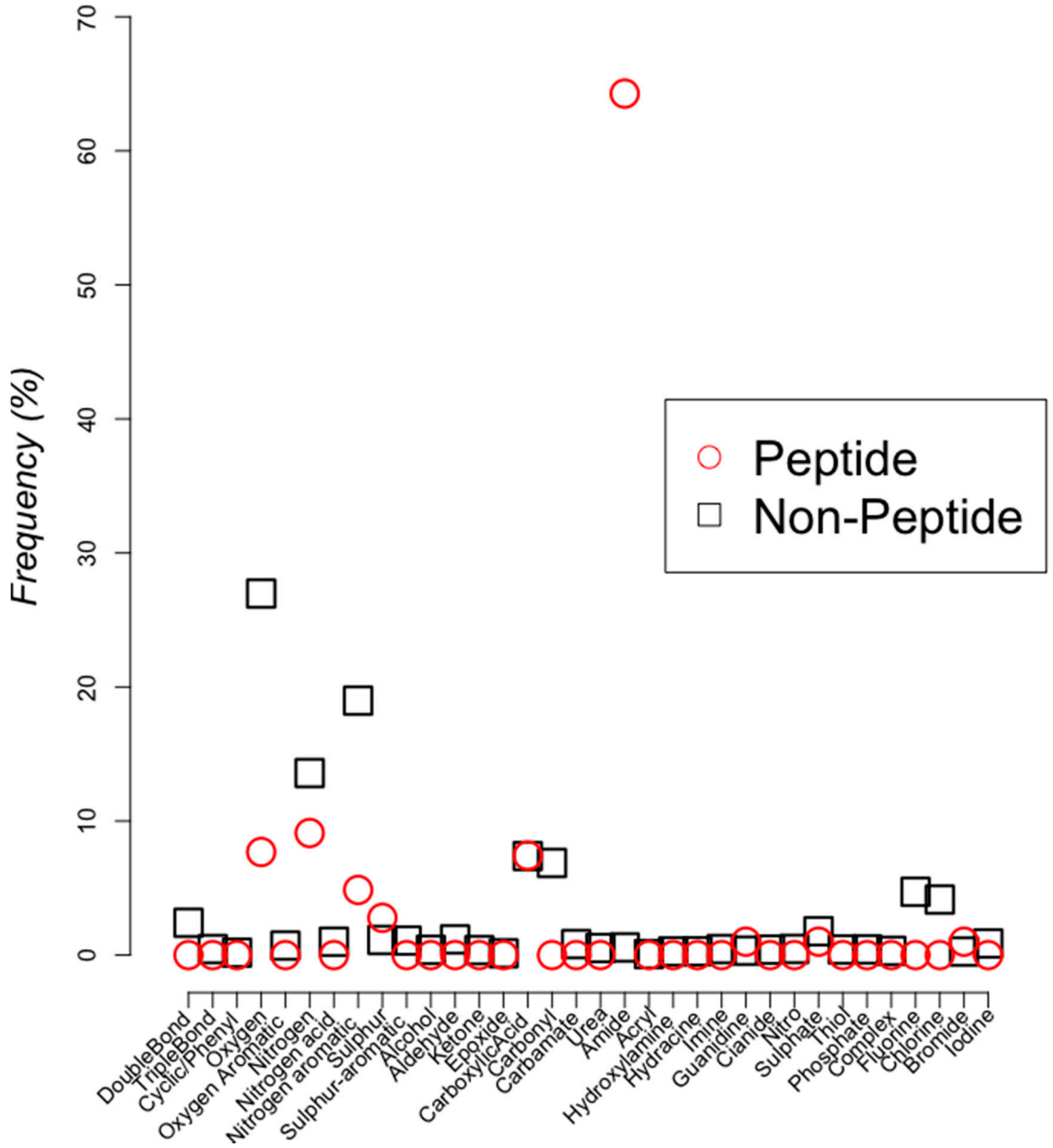

Figure 4. Chemical groups comparison. Functional group frequency (%) is compared between peptides (red circles) with non-peptidic chemical compounds (black squares). The names of the functional groups found in both sets are indicated on the x-axis (see Methods).

In our previous work, training for drugs that acted against a single gut bacterium identified broad-spectrum antibiotics among FDA approved antibiotics [20]. Our heterologous model in that previous work was not trained to identify broad-spectrum antibiotics, but we were able to identify them because FDA-approved antibiotics are targeted against pathogens while our model would predict these to also act on the healthy gut microbiome. In our current work, we trained our models to predict compounds acting against multiple microbes; hence these models were trained to identify broad-spectrum antibiotics. The best features, model and corresponding parameters were identified using an automatic optimization procedure (see Methods); then the models were tested using a 10-fold cross-validation (see Methods). The use of cross-validation is useful in cases where there is no evidence of noise or incorrect labeling of instances; hence it is a convenient option for our approach.

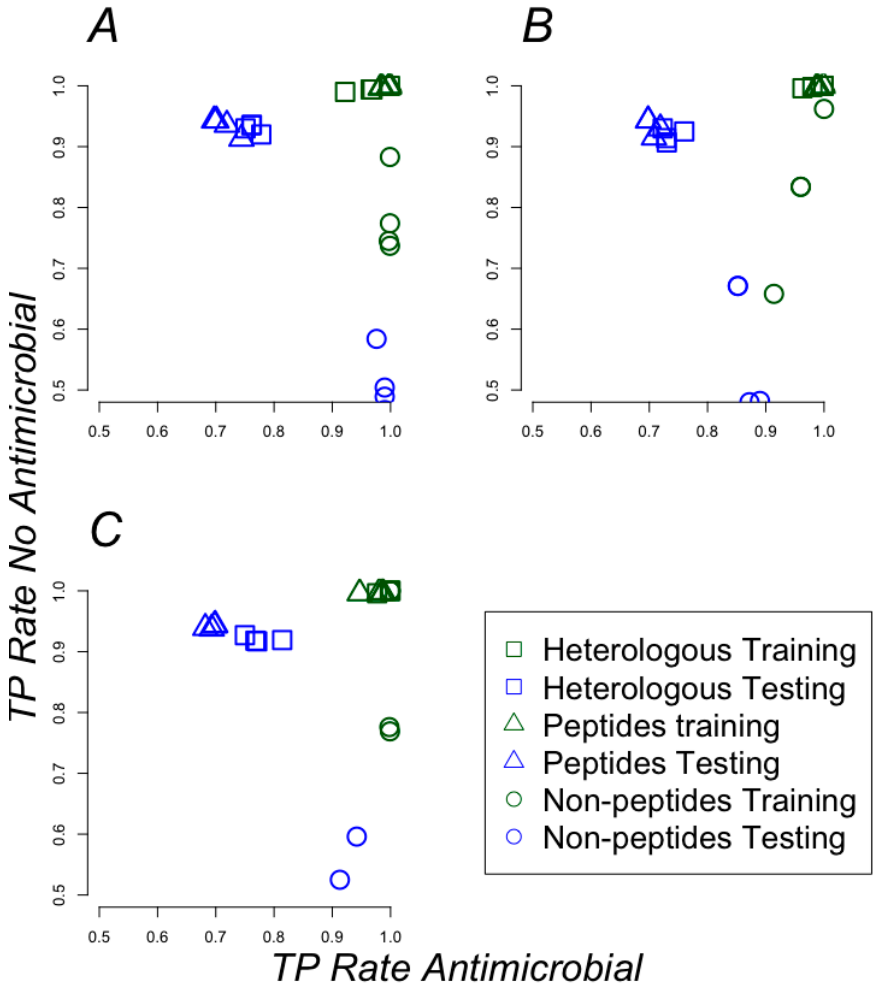

Figure 5. Learning efficiency based on physicochemical features. (**A**) Anti-infective, (**B**) anti-gut1, and (**C**) anti-gut4 training sets. True positive (TP) rates (number of correct predictions divided by the number of positive instances) for classifying antimicrobials and non-antimicrobials are shown for models trained with each training sets (green symbols) and for a 10-fold cross-validation test (blue symbols). Squared symbols correspond with models trained with heterologous data (peptides and non-peptide compounds); triangle symbols represent models trained with peptides; circular symbols represent models trained with non-peptidic compounds.

Twelve different models were generated for each training set: four sets for heterologous training, four sets using only non-peptidic chemical compounds and other four sets using only peptidic chemical compounds (see Methods and Supplementary Files S4–S39). For instance, a set of compounds (e.g., anti-gut1) is complemented with antimicrobial and non-antimicrobial peptides to create a heterologous training set; the original set includes only non-peptide compounds hence are considered as training set with non-peptidic compounds; for the only peptidic chemical compounds, we used the set of antimicrobial and non-antimicrobial peptides used to create the heterologous set. For each of these sets (heterologous, only peptides, non-peptidic chemical compounds), four different representations of the PadelDescriptor features were generated: (a) nominal (classes –antimicrobial and non-antimicrobials, were labeled as nominal), (b) nominal and normalized (nominal plus PadelDescriptor features were normalized and nominal), (c) nominal, normalized and attribute selection (as b plus PadelDescriptor features were selected by the CfsSubsetEval Weka filter), and (d) nominal and attribute selection (see Methods). Figure 5A–C show the true positive rates (see Methods) for antimicrobials and non-antimicrobials achieved by these models on each training set. Noticeable, models trained with anti-gut4 achieved the best classification rates and heterologous datasets (square symbols in Figure 5)

rendered some of the best models followed by models trained with only peptides (triangles in Figure 5); these results indicate that polyHAM are to be found among antibiotics that act on multiple species, as expected. The confusion matrices and MCC score for every model are reported in Supplementary Files S40–S42; the best models achieved MCC scores above 0.95.

The names of the models found for all training sets are presented in Table S3. The best model (dataset anti-gut4 using heterologous and nominal data using the random forest model) was able to classify correctly 89% of every antimicrobial and non-antimicrobial compound in the anti-gut4 set as reported by AutoWeka. The top 10 features used by the best model, out of the 507 features used by the model, are shown in Table 2. Please note that all these features are related to the information or graph theory parameters of chemical groups, rather than chemical attributes; this is in agreement with our previous observation that peptides and non-peptide compounds shared few and too general chemical groups, hence chemical features were not useful for classification purposes. TP rates were used to evaluate the possible bias in classification induced by the biased composition in the training sets. For instance, the anti-infective data set has 137 positive and 1044 negative instances (see Figure 2); yet, we observed that the best models (heterologous sets that added 7999 antimicrobial peptides and 3546 non-antimicrobial peptides, rendering a total of 8136 positives and 4590 negatives) did not favor the classification of antimicrobials or non-antimicrobials as it can be observed in Figure 5A, where the square blue symbols show around 0.9 of TP rate for non-antimicrobial and around 0.7 of TP rate for antimicrobial, a strong influence of the class imbalance should have produced a larger than the observed gap (0.9 vs 0.7) between these two cases (8136 vs 4590).

Table 2. Top 10 features for the best model (random forest) during training.

Attribute Name [1]	Description
BCUTw-1h	Eigenvalue based descriptor noted for its utility in chemical diversity
AATS5m	Average Broto—Moreau autocorrelation—lag 5/weighted by mass
MIC5a	Modified information content index (neighborhood symmetry of 5-order)/ weighted by atoms
AATS6m	Average Broto—Moreau autocorrelation—lag 6/weighted by mass
AATS7m	Average Broto—Moreau autocorrelation—lag 7/weighted by mass
IC5	Information content index (neighborhood symmetry of 5-order)
MIC4	Modified information content index (neighborhood symmetry of 4-order)
AATSC0m	Average centered Broto—Moreau autocorrelation—lag 0/weighted by mass
topoDiameter	Topological diameter (maximum atom eccentricity)
AATS0m	Average Broto—Moreau autocorrelation—lag 0/weighted by mass

[1] Attribute ranking was based on the information Gain Ranking filter implemented in Weka (see Methods).

Based on this data, we performed a test of the 12 anti-gut4 models (four models with heterologous training set, four models with only peptides, and four models with non-peptidic compounds; the four models correspond to four different ways to process the data, see Methods) with a set of 17 metabolites (see Supplementary Table S4) that are part of the Human Metabolome Database, which includes metabolites found in the human body. Please note that none of the compounds in the testing set were part of the original training set. While plants and microbes are known to produce secondary metabolites with antimicrobial activity, human metabolites are not commonly recognized to harbor this antimicrobial activity. Applying our naïve text-mining approach, six metabolites on this test set were identified as antimicrobials based on publications that included the ontological terms associated to antimicrobial activity (see Section 2.1) and for the other 11 there was no publication about their antimicrobial activity. According to the ZINC database, the 17 human metabolites have a human target and six have antimicrobial activity corresponding with true polyHAM compounds; four out of these six true polyHAM are broad-spectrum antibiotics (see Supplementary Table S4). We observed that in this test set, the best model is derived from non-peptidic compounds as training set using nominal representation of the data (see red circle on Figure 6). We also noted that overall heterologous models

classified better the non-antimicrobials than the models trained with only peptides or non-peptidic chemical compounds (see Figure 6): three out of the four models trained with only peptides predicted none of the non-antimicrobials (observation derived from data presented in Figure 6; these three models are observed as a single remarked triangle at the upper left corner of the plot); three out of the four models trained with non-peptidic chemical compounds predicted every non-antimicrobial, but no antimicrobial (observation derived from data presented in Figure 6; these three models are observed as a single remarked circle at the bottom right corner of the plot). Hence in general, the models obtained with heterologous sets rendered more balanced predictions. Yet, it is clear that this test set was not an easy task for any of the models.

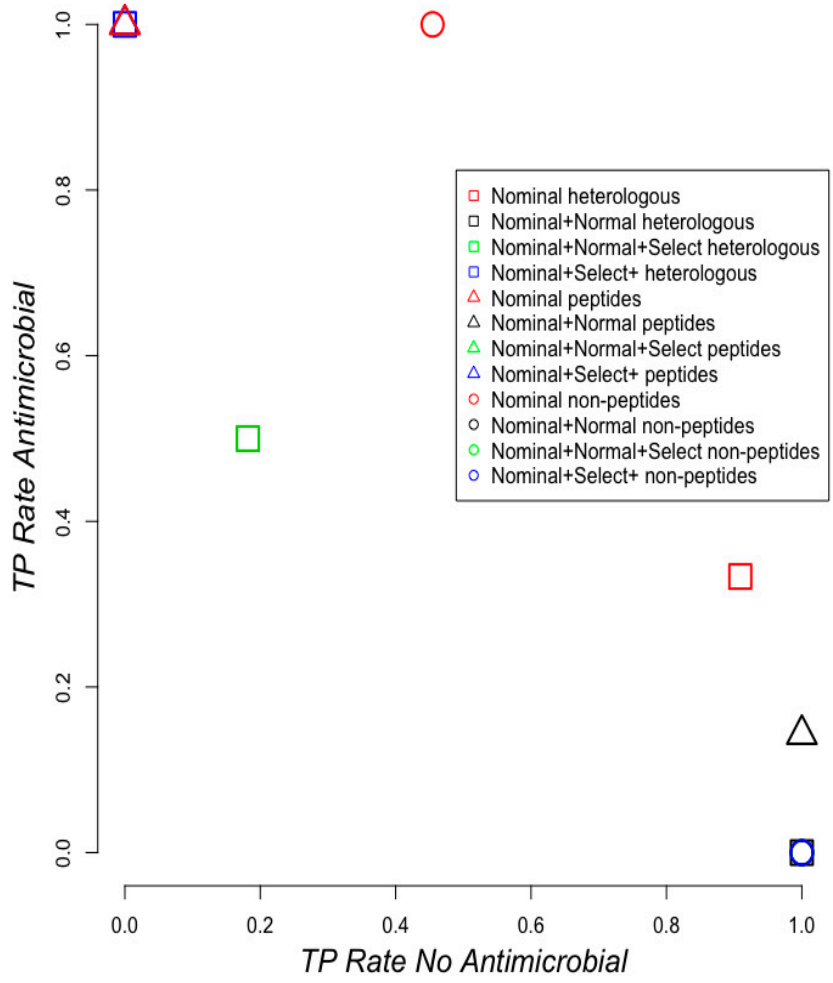

Figure 6. Testing models on human metabolites. True positive (TP) rates (number of correct predictions divided by the number of true positives) for classifying antimicrobials and non-antimicrobials are shown for models trained in the anti-gut4 dataset using heterologous data (peptides and non-peptide compounds; squares), only peptides (triangles) and non-peptidic compounds (circles). There are four instances of each symbol corresponding with the four data representation used: (i) nominal (red); (ii) nominal and normalized (black); (iii) nominal normalized and selected attributes (green); and (iv) nominal and selected attributes (blue) (see Methods).

Considering the relatively low performance to classify polyHAM in human metabolites, we evaluated how different were the testing set compounds from those found in the training sets. The dataset rendering the best model (anti-gut4 using heterologous and nominal modeling of data) was used to compute the shortest Manhattan distance between the polyHAM in the training and the testing sets (D1); as a reference, we also computed the shortest distance between the non-polyHAM

and polyHAM in the same training set (D2); the Manhattan distance was calculated from the vector representation of every compound, in which the vector values were the features calculated for each compound. If D1 > D2 then, the compounds in the testing set were more different to the polyHAM than the non-polyHAM compounds, yet D1 = D2 = 0; as noted in Figures 3 and 4, the diversity in mass and chemical groups in the training set was large, hence this result is in agreement with these previous observations. We then identified the shortest distance between the polyHAM in the testing sets with those of the non-polyHAM in the training set (D3 = 6708088) and compared it with the shortest distance between polyHAM and non-polyHAM in the training set (D2 = 0) and observed that the testing set was more distant from the non-polyHAM compounds in the training sets than the polyHAM compounds in the training set. These results indicate that the testing set was more distant from the non-polyHAM compounds yet, close to the polyHAM compounds in the training set. This distinct distribution of the testing set may provide an explanation for the difficulty in classification for the best models, that is, the best models may have traced frontiers between the positive and negative antimicrobial compounds that excluded several of the compounds in the test set that were too distant from the true positive polyHAM in the training set.

3. Discussion

The story of penicillin set a hallmark for antibiotic discovery: it supported the magic-bullet idea proposed by Paul Erlich in which a single drug would traverse the body and only act on a specific target [24]. However, the discovery of antimicrobial peptides as part of the defense mechanism of every cell brought a new concept into the antibiotics field: natural antimicrobial peptides are less likely to evoke resistance in microbes because they act on multiple targets [25]. Thus, while it is possible to kill microbes targeting essential cellular functions coded into a single protein (*e.g.*, penicillin-binding proteins), it is also possible to kill microbes by targeting multiple targets. This last view is important for many complex diseases humans are facing nowadays that are related to gut microbiome, such as obesity [26], hypertension [27], among others [28,29], where the antimicrobial activity in drugs is relevant to either treat the disease (by killing microbes associated with the disease) or prevent it (by not killing microbes that prevent the development of the disease).

Drug repositioning or repurposing frequently identifies antibiotics to treat diseases other than infections [30]; such is the case of minocycline that is a semisynthetic tetracycline-derived antibiotic that has shown to have neuroprotective activity and it is currently being tested in treating Parkinson's disease [31]. Hence, it seems that polypharmacologic antimicrobials that act on human targets, polyHAM, represent a resource to identify effective drugs to treat different conditions beyond infections. In our previous work, we were interested in testing the reliability of heterologous training sets in identifying polyHAM [20]. In this study we characterize several features of these compounds using different computational approaches. We show that there are very few chemical similarities between peptides and non-peptide human-targeted drugs, and consistent with these findings we observed that topological features of chemical structures are more informative than chemical descriptors of molecules for the classification of polyHAM. Thus, for polyHAM to act on multiple targets it is relevant to display specific topological features rather than particular chemical groups. This suggests that the structural organization of chemical groups rather than the chemical groups, per se, are relevant for acting on multiple targets. These results indicate that antimicrobial classifications based on chemical descriptors (see, for instance, [32–34]) may not work properly to classify polyHAM.

We used three different training sets for identifying the best model to classify polyHAM: anti-infective, anti-gut1, and anti-gut4; these last two sets differ in the number of bacterial strains that these compounds act against to, increasing from anti-gut1 to anti-gut4, and expecting that anti-gut4 had the most broad ability to act as antimicrobials than the other compounds in anti-infective and anti-gut1 sets. Indeed, we verified by an automatic bibliographic search that anti-gut4 had a larger proportion of known broad antimicrobial compounds than anti-gut1. As noted above, anti-gut4 also rendered the best model to classify polyHAM from non-polyHAM. This result indicates that broad-spectrum

antimicrobials make better polyHAM than specific ones and that broad-spectrum antimicrobials are more likely to act on multiple targets.

Finally, human metabolites were used to test our trained model to classify polyHAM; these compounds are not commonly regarded as a source of antimicrobial compounds. Here we show that there is a group of these human metabolites with previous reports about their broad-spectrum antimicrobial activity, hence, these may represent a natural way for humans to control microbiome composition. For instance, the presence of 3-phenylpropionic acid has been shown to be affected by antibiotics altering the healthy microbiome composition [35], and at the same time has been shown to prevent the growth of the pathogenic *Listeria monocytogenes* in combination with a natural antimicrobial peptide [36]. In this case, the quantification of this metabolite may indicate the susceptibility of a healthy individual to be infected by pathogens such as *L. monocytogenes*. Thus, predictions of polyHAM among human metabolites may promote the development of tools to design human diets aimed to alter the specific composition of human metabolites. It is expected that certain diets would promote the accumulation of metabolites with broad antimicrobial activity that, in turn, promote gut microbiome dysbiosis. Further studies are required to validate this idea, yet our findings represent an important advance in that direction.

4. Materials and Methods

4.1. Dataset Preparation

Three datasets were constructed for the purpose of training machine-learning models, including anti-infective (Supplementary File S1), anti-gut1 (Supplementary File S2), and anti-gut4 (Supplementary File S3); all these datasets were derived from the work reported by Maier and collaborators that included 1181 FDA-approved compounds [19]. These sets differ in the labels identifying compounds as antimicrobials or non-antimicrobials; for instance, in the anti-infective set the compounds were labeled as antimicrobials if they belong to class J in the Anatomical Therapeutic Chemical (ATC) Classification System, in the anti-gut1 set, compounds were labeled as antimicrobials when Maier and collaborators identified antimicrobial activity against at least one gut bacteria and anti-gut4 those hitting at least four gut bacteria. The features were normalized and/or performed a selection of features using the weka.attributeSelection.CfsSubsetEval filter [37].

For each of these three training sets (anti-infective, anti-gut1 and anti-gut4), 12 different sets were generated: four for heterologous data, four for non-peptidic compounds, and four for only peptidic compounds; each of these four sets corresponds with nominal, nominal-normalized, nominal-normalized-selected attributes, and nominal selected attributes, in a similar fashion as described previously [20]. For the non-peptidic compounds set, 1181 non-peptidic compounds were used (the number of antimicrobials and non-antimicrobials depended on the training set, see Figure 7); the only peptidic compounds set included 11,545 peptides (7999 antimicrobial and 3546 non-antimicrobials) that were obtained from a non-redundant compilation of multiple antimicrobial peptide databases [38]; the heterologous set included all non-peptidic (1181 compounds) and peptidic compounds (11,545 peptides). In order to obtain the features describing each compound the PaDelDescriptor software was used [23]; to do that, non-peptidic compounds were represented in SMILES format while for peptides, since the FASTA format is not readable by PaDelDescriptor, they were converted to MOL2 format (version V200) by the program Seq2Mol.jar (see supplementary material to get the code and instructions to execute it). Every feature with values equals to 0 or null in 50% or more of all instances was removed. Finally, 12 different models were generated for each training set: four sets for heterologous training, four sets using only non-peptidic chemical compounds and other four sets using only peptidic chemical compounds (see Figure 7). These training sets were converted into ARFF format (see supplementary files S4–S39). The reliability of every Weka classifier trained with any of these training sets was tested using a 10-fold cross-validation. The script to run the

cross-validation test that includes the model name and its corresponding parameters are available at the supplementary Files S43 (anti-infective), S44 (anti-gut1), and S45 (anti-gut4).

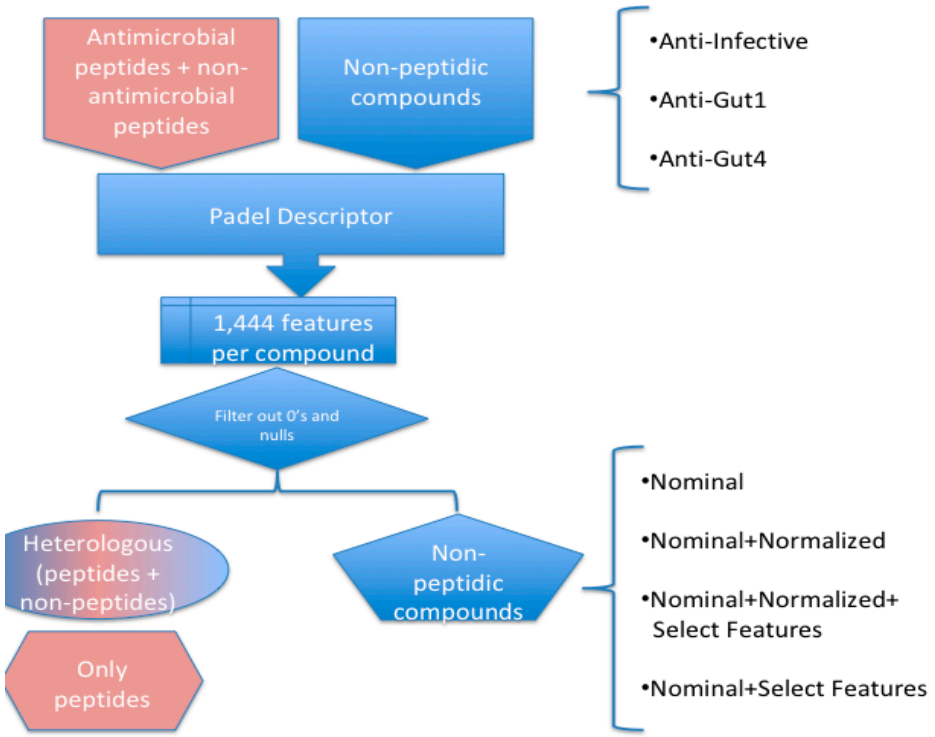

Figure 7. Dataset construction. The steps described to build the datasets (anti-infective, anti-gu1, and anti-gut4) used throughout this work are described. From 1,181 non-peptidic compounds and 11,545 peptides, three sets were derived: heterologous, only peptides and non-peptidic compounds. Each of these sets was converted into ARFF format using the specified combination of the following procedures: nominal representation of the class (antimicrobial or non-antimicrobial), features normalized, and/or feature selection.

4.2. Naïve Text Mining Approach

The bibliographic data was obtained from PubChem [39]. The PubChem ID from every compound used in the training sets was obtained from its SMILES formula; from this PubChem ID, every associated PMID was retrieved using PubChem REST server. The PMID was used to retrieve the MedLine format of each publication from PubMed server and to identify the corresponding ontologic terms (entries identified by the headers MH or OT in the MedLine registry). The code used for this purpose is available as a supplementary file named GetPubMedEntriesFromSMILES.java. Figure 3 displays the proportion of references that included an antimicrobial-related ontological term according to the formula:

$$\text{Proportion} = \text{NCAMP}/\text{TotalNC} \quad (1)$$

where Proportion is the proportion of antimicrobial citations; NCAMP is the number of publications of the compound of interest including an antimicrobial-related ontological term; TotalNC is the total number of publications for the compound of interest.

The test data set was obtained from the ZINC database (version 15) [40] and corresponds with metabolites identified from The Human Metabolome Database [41]. We chose human metabolites for the test set because of the known relationship between microbiome and human metabolism relevant for human health and disease [42]. Every feature describing a compound was calculated as in the training sets, using the PaDelDescriptor software [23]. Briefly, the chemical features correspond with any of the 1444 two-dimensional features calculable by PaDelDescriptor that include chemical

features (e.g., acidic groups count, longest aliphatic chain, Hbond donor count) and topological features (e.g., atom type electrotopological state, path counts, topological distance matrix); these features were obtained directly from the SMILES representation of each of the 17 metabolites in the test set. These 17 metabolites were those found with at least one publication including the ontological terms associated with antimicrobial activity identified in the training set. The same features obtained for every model during the training (12 models for anti-gut4) were included for the test set. Antimicrobials were considered those compounds that contained two or more publications with ontological terms related to antimicrobial activity (see Supplementary Table S4); the annotation as antimicrobial was done after a human reviewed the literature to confirm the antimicrobial activity.

4.3. Machine-Learning Approach

Weka version 3.8 [43] and the AutoWeka [44] plugin were used to train and randomly cross-validate the models. It is worth to mention that AutoWeka includes the state-of-the-art machine-learning algorithms, like SVM, random forest, and logistic regression, among others, so the resulting learning model can be considered as the most suitable for the classification task at hand. For identifying chemical groups (these are different than the features calculated by PaDel descriptor) in peptides and non-peptidic chemical compounds, we used a Python implementation developed for that purpose [22]. Briefly, molecules in SMILES format were integrated into the python code, and the program named rdkitpy was installed as instructed by the developers of Python program; the code searches for 3080 known chemical groups in molecules. True positive rates (TP rates) were used to estimate the proportion of correctly classified instances for antimicrobials and non-antimicrobials; this would allow evaluating for any possible bias in the classification. True positives refer to those instances that were predicted correctly within a class; e.g., an antimicrobial compound that was predicted as antimicrobial is a true positive. Attribute ranking for the best model identified in the training and cross-validation test was performed using the information gain for attribute evaluation filter implemented in Weka; briefly, the information gain for each attribute is derived from the information entropy for each attribute for each class.

All data required to reproduce and analyze the results presented in this work is available through GitHub: https://gdelrioifc.github.io/PolyHAM/.

5. Conclusions

In summary, we characterize human-targeted antimicrobials using heterologous training sets and machine learning approaches. PolyHAM display broad-spectrum antibiotic activity and are found circulating the human body where microbes are found.

Supplementary Materials: The following are available online at https://gdelrioifc.github.io/PolyHAM/, File S1. Anti-Infective set of compounds; File S2. Anti-Gut1 set of compounds; File S3. Anti-Gut4 set of compounds; File S4. Anti-Infective Hetero Nominal training set in ARFF format; File S5. Anti-Infective Hetero Nominal Normalized training set in ARFF format; File S6. Anti-Infective Hetero Nominal Normalized Selected Attributes training set in ARFF format; File S7. Anti-Infective Hetero Nominal Selected Attributes training set in ARFF format; File S8. Anti-Infective Non-peptidic Chemical Compounds Nominal training set in ARFF format; File S9. Anti-Infective Non-peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S10. Anti-Infective Non-peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S11. Anti-Infective Non-peptidic Chemical Compounds Nominal Selected Attributes training set in ARFF format; File S12. Anti-Infective Peptidic Chemical Compounds Nominal training set in ARFF format; File S13. Anti-Infective Peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S14. Anti-Infective Peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S15. Anti-Infective Peptidic Chemical Compounds Nominal Selected Attributes training set in ARFF format; File S16. Anti-Gut1 Hetero Nominal training set in ARFF format; File S17. Anti-Gut1 Hetero Nominal Normalized training set in ARFF format; File S18. Anti-Gut1 Hetero Nominal Normalized Selected Attributes training set in ARFF format; File S19. Anti-Gut1 Hetero Nominal Selected Attributes training set in ARFF format; File S20. Anti-Gut1 Non-peptidic Chemical Compounds Nominal training set in ARFF format; File S21. Anti-Gut1 Non-peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S22. Anti-Gut1 Non-peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S23. Anti-Gut1 Non-peptidic Chemical Compounds Nominal Selected Attributes training set in

ARFF format; File S24. Anti-Gut1 Peptidic Chemical Compounds Nominal training set in ARFF format; File S25. Anti-Gut1 Peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S26. Anti-Gut1 Peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S27. Anti-Gut1 Peptidic Chemical Compounds Nominal Selected Attributes training set in ARFF format; File S28. Anti-Gut4 Hetero Nominal training set in ARFF format; File S29. Anti-Gut4 Hetero Nominal Normalized training set in ARFF format; File S30. Anti-Gut4 Hetero Nominal Normalized Selected Attributes training set in ARFF format; File S31. Anti-Gut4 Hetero Nominal Selected Attributes training set in ARFF format; File S32. Anti-Gut4 Non-peptidic Chemical Compounds Nominal training set in ARFF format; File S33. Anti-Gut4 Non-peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S34. Anti-Gut4 Non-peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S35. Anti-Gut4 Non-peptidic Chemical Compounds Nominal Selected Attributes training set in ARFF format; File S36. Anti-Gut4 Peptidic Chemical Compounds Nominal training set in ARFF format; File S37. Anti-Gut4 Peptidic Chemical Compounds Nominal Normalized training set in ARFF format; File S38. Anti-Gut4 Peptidic Chemical Compounds Nominal Normalized Selected Attributes training set in ARFF format; File S39. Anti-Gut4 Peptidic Chemical Compounds Nominal Selected Attributes training set in ARFF format; File S40. Confusion matrix parameters and MCC score for anti-infective set; File S41. Confusion matrix parameters and MCC score for anti-gut1 set; File S42. Confusion matrix parameters and MCC score for anti-gut4 set; File S43. Weka Script for cross-validation of the anti-infective set; File S44. Weka Script for cross-validation of the anti-gut1 set; File S45. Weka Script for cross-validation of the anti-gut4 set; Table S1. PubMed Citations with ontological terms associated with antimicrobial activity; Table S2. PolyHAM in training set; Table S3. Best models hyperparameters; Table S4. Testing set. Three codes (GetPubMedEntriesFromSMILESTraining.java, GetPubMedEntriesFromSMILESTesting.java and Seq2Mol.jar) and two files (TableS1_Prestwick_STRINGWithNames.tsv, hmdbmetab-metabolites.csv) for testing these codes are also available at the github web site: https://github.com/gdelrioifc/PolyHAM.

Author Contributions: Conceptualization: G.D.R., R.A.N.L., J.A.B., and C.A.B.; methodology: G.D.R., J.A.B., and R.A.N.L.; software: G.D.R., J.A.B., and R.A.N.L.; validation: G.D.R., J.A.B., and R.A.N.L.; formal analysis: G.D.R. and R.A.N.L.; resources: G.D.R. and C.A.B.; data curation: G.D.R., R.A.N.L., and C.A.B.; writing—original draft preparation: G.D.R.; writing—review and editing: R.A.N.L., J.A.B., and C.A.B.; funding acquisition: G.D.R. and C.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CONACyT (CB252316), PAPIIT (IN208817) and Instituto de fisiologia celular-UNAM.

Acknowledgments: To Dra. Maria Teresa Lara Ortiz at Instituto de fisiologia celular UNAM, for her technical assistance relevant for the development of this work and Camila Del Rio (https://camidelrica.wixsite.com/neurocosa-que-dibuja?lang=en) for the graphical abstract image.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Fang, Y. Label-free drug discovery. *Front. Pharmacol.* **2014**, *5*, 52. [CrossRef] [PubMed]
2. Schneider, P.; Walters, W.P.; Plowright, A.T.; Sieroka, N.; Listgarten, J.; Goodnow, R.A., Jr.; Fisher, J.; Jansen, J.M.; Duca, J.S.; Rush, T.S.; et al. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* **2019**, *19*, 353–364.
3. Sun, X.; Vilar, S.; Tatonetti, N.P. High-Throughput Methods for Combinatorial Drug Discovery. *Sci. Transl. Med.* **2013**, *5*, 205rv1.
4. Oh, D.Y.; Kim, K.; Kwon, H.B.; Seong, J.Y. Cellular and Molecular Biology of Orphan G Protein-Coupled Receptors. *Int. Rev. Cytol.* **2006**, *252*, 163–218. Available online: http://www.ncbi.nlm.nih.gov/pubmed/16984818 (accessed on 12 February 2020).
5. Vieth, M.; Sutherland, J.J.; Robertson, D.H.; Campbell, R.M. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discov. Today* **2005**, *10*, 839–846. Available online: http://www.ncbi.nlm.nih.gov/pubmed/15970266 (accessed on 12 February 2020). [CrossRef]
6. Drawz, S.M.; Bonomo, R.A. Three Decades of β-Lactamase Inhibitors. *Clin. Microbiol. Rev.* **2010**, *23*, 160–201. [CrossRef]
7. Amelio, I.; Lisitsa, A.; Knight, R.A.; Melino, G.; Antonov, A.V. Polypharmacology of approved anticancer drugs. *Curr. Drug Targets* **2016**, *18*, 534–543. Available online: http://www.ncbi.nlm.nih.gov/pubmed/26926468 (accessed on 12 February 2020). [CrossRef]
8. Li, Y.; Wang, P.P.; Li, X.X.; Yu, C.Y.; Yang, H.; Zhou, J.; Xue, W.; Tan, J.; Zhu, F. The Human Kinome Targeted by FDA Approved Multi-Target Drugs and Combination Products: A Comparative Study from the Drug-Target Interaction Network Perspective. *PLoS ONE* **2016**, *11*, e0165737. [CrossRef]

9. Osterloh, I.H. The discovery and development of Viagra® (sildenafil citrate). *Sildenafil* **2004**, 1–13. [CrossRef]
10. Dutta, S.; Kumar, S.; Hyett, J.; Salomon, C. Molecular Targets of Aspirin and Prevention of Preeclampsia and Their Potential Association with Circulating Extracellular Vesicles during Pregnancy. *Int. J. Mol. Sci.* **2019**, *20*, 4370. [CrossRef]
11. Mohajeri, M.H.; Brummer, R.J.M.; Rastall, R.A.; Weersma, R.K.; Harmsen, H.J.M.; Faas, M.; Eggersdorfer, M. The role of the microbiome for human health: From basic science to clinical applications. *Eur. J. Nutr.* **2018**, *57*, 1–14. [CrossRef]
12. Kho, Z.Y.; Lal, S.K. The Human Gut Microbiome—A Potential Controller of Wellness and Disease. *Front. Microbiol.* **2018**, *9*, 1835. [CrossRef] [PubMed]
13. Nadal, I.; Donant, E.; Ribes-Koninckx, C.; Calabuig, M.; Sanz, Y. Imbalance in the composition of the duodenal microbiota of children with coeliac disease. *J. Med. Microbiol.* **2007**, *56*, 1669–1674. Available online: http://www.ncbi.nlm.nih.gov/pubmed/18033837 (accessed on 12 February 2020). [CrossRef]
14. Marasco, G.; Di Biase, A.R.; Schiumerini, R.; Eusebi, L.H.; Iughetti, L.; Ravaioli, F.; Scaioli, E.; Colecchia, A.; Festi, D. Gut Microbiota and Celiac Disease. *Dig. Dis. Sci.* **2016**, *61*, 1461–1472. [CrossRef] [PubMed]
15. Nagao-Kitamoto, H.; Shreiner, A.B.; Gillilland, M.G.; Kitamoto, S.; Ishii, C.; Hirayama, A.; Kuffa, P.; El-Zaatari, M.; Grasberger, H.; Seekatz, A.M.; et al. Functional Characterization of Inflammatory Bowel Disease-Associated Gut Dysbiosis in Gnotobiotic Mice. *Cell. Mol. Gastroenterol. Hepatol.* **2016**, *2*, 468–481. Available online: http://www.ncbi.nlm.nih.gov/pubmed/27795980 (accessed on 12 February 2020). [CrossRef]
16. Ochoa-Cortes, F.; Turco, F.; Linan-Rico, A.; Soghomonyan, S.; Whitaker, E.; Wehner, S.; Cuomo, R.; Christofi, F.L. Enteric Glial Cells: A New Frontier in Neurogastroenterology and Clinical Target for Inflammatory Bowel Diseases. *Inflamm. Bowel. Dis.* **2016**, *22*, 433–449. Available online: http://www.ncbi.nlm.nih.gov/pubmed/26689598 (accessed on 12 February 2020). [CrossRef]
17. The Integrative HMP (iHMP) Research Network Consortium; Integrative HMP (iHMP) Research Network Consortium; Buck, G. The Integrative Human Microbiome Project. *Nature* **2019**, *569*, 641–648. [CrossRef]
18. Forslund, K.; MetaHIT Consortium; Hildebrand, F.; Nielsen, T.G.; Falony, G.; Le Chatelier, E.; Sunagawa, S.; Prifti, E.; Vieira-Silva, S.; Gudmundsdottir, V.; et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **2015**, *528*, 262–266. [CrossRef] [PubMed]
19. Maier, L.; Pruteanu, M.; Kuhn, M.; Zeller, G.; Telzerow, A.; Anderson, E.E.; Brochado, A.R.; Fernandez, K.C.; Dose, H.; Mori, H.; et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **2018**, *555*, 623–628. [CrossRef]
20. Lara, R.A.N.; Aguilera-Mendoza, L.; Brizuela, C.A.; Peña, A.; Del Rio, G. Heterologous Machine Learning for the Identification of Antimicrobial Activity in Human-Targeted Drugs. *Molecules* **2019**, *24*, 1258. Available online: http://www.ncbi.nlm.nih.gov/pubmed/30935109 (accessed on 12 February 2020). [CrossRef]
21. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH.: Weinheim, Germany, 2000; p. 667.
22. Ertl, P. An algorithm to identify functional groups in organic molecules. *J. Chemin.* **2017**, *9*, 36. Available online: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0225-z (accessed on 18 February 2020). [CrossRef] [PubMed]
23. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2010**, *32*, 1466–1474. Available online: http://www.ncbi.nlm.nih.gov/pubmed/21425294 (accessed on 12 February 2020). [CrossRef] [PubMed]
24. Zaffiri, L.; Gardner, J.; Toledo-Pereyra, L.H. History of Antibiotics. From Salvarsan to Cephalosporins. *J. Investig. Surg.* **2012**, *25*, 67–77. Available online: http://www.ncbi.nlm.nih.gov/pubmed/22439833 (accessed on 12 February 2020). [CrossRef]
25. Nuti, R.; Goud, N.S.; Saraswati, A.P.; Alvala, R.; Alvala, M.; Nuti, N.S.G.R. Antimicrobial Peptides: A Promising Therapeutic Strategy in Tackling Antimicrobial Resistance. *Curr. Med. Chem.* **2017**, *24*, 4303–4314. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28814242 (accessed on 12 February 2020). [CrossRef] [PubMed]
26. Maruvada, P.; Leone, V.; Kaplan, L.M.; Chang, E.B. The Human Microbiome and Obesity: Moving beyond Associations. *Cell Host Microbe* **2017**, *22*, 589–599. Available online: http://www.ncbi.nlm.nih.gov/pubmed/29120742 (accessed on 12 February 2020). [CrossRef]

27. Tang, W.H.W.; Kitai, T.; Hazen, S.L. Gut Microbiota in Cardiovascular Health and Disease. *Circ. Res.* **2017**, *120*, 1183–1196. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28360349 (accessed on 12 February 2020). [CrossRef]
28. Wang, J.; Kurilshikov, A.; Radjabzadeh, D.; Turpin, W.; Croitoru, K.; Bonder, M.J.; Jackson, M.A.; Medina-Gomez, C.; Frost, F.; Homuth, G.; et al. Meta-analysis of human genome-microbiome association studies: The MiBioGen consortium initiative. *Microbiome* **2018**, *6*, 101. Available online: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0479-3 (accessed on 17 February 2020). [CrossRef]
29. Goodrich, J.K.; Davenport, E.R.; Clark, A.G.; Ley, R.E. The Relationship Between the Human Genome and Microbiome Comes into View. *Annu. Rev. Genet.* **2017**, *51*, 413–433. [CrossRef]
30. Gns, H.S.; Gr, S.; Murahari, M.; Krishnamurthy, M. An update on Drug Repurposing: Re-written saga of the drug's fate. *Biomed. Pharmacother.* **2019**, *110*, 700–716. [CrossRef]
31. Cankaya, S.; Cankaya, B.; Kilic, U.; Kilic, E.; Yulug, B. The therapeutic role of minocycline in Parkinson's disease. *Drugs Context* **2019**, *8*, 1–14. [CrossRef]
32. Fingerhut, L.C.H.W.; Miller, D.J.; Strugnell, J.M.; Daly, N.L.; Cooke, I.R. OUP accepted manuscript. *Bioinformatics* **2020**, btaa653. [CrossRef] [PubMed]
33. Khatun, S.; Hasan, M.; Kurata, H.; Khatun, M.S. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. *FEBS Lett.* **2019**, *593*, 3029–3039. [CrossRef] [PubMed]
34. Lin, Y.; Cai, Y.; Liu, J.; Lin, C.; Liu, X. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinform.* **2019**, *20*, 291. [CrossRef] [PubMed]
35. Swann, J.R.; Tuohy, K.M.; Lindfors, P.; Brown, D.A.; Gibson, G.R.; Wilson, I.D.; Sidaway, J.; Nicholson, J.K.; Holmes, E. Variation in Antibiotic-Induced Microbial Recolonization Impacts on the Host Metabolic Phenotypes of Rats. *J. Proteome Res.* **2011**, *10*, 3590–3603. [CrossRef] [PubMed]
36. Molinos, A.C.; Abriouel, H.; Ben Omar, N.; Lucas, R.; Valdivia, E.; Galvez, A. Inactivation of Listeria monocytogenes in Raw Fruits by Enterocin AS-48. *J. Food Prot.* **2008**, *71*, 2460–2467. [CrossRef] [PubMed]
37. Hall, M.A. Correlation-based Feature Selection for Machine Learning The University of Waikato, New Zeland, 1999. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.4643 (accessed on 12 February 2020).
38. Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J.A.; Ibarra, R.T.; Guillen-Ramirez, H.; Brizuela, C.A. Graph-based data integration from bioactive peptide databases of pharmaceutical interest: Toward an organized collection enabling visual network analysis. *Bioinformatics* **2019**, *35*, 4739–4747. [CrossRef]
39. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **2018**, *47*, D1102–D1109. [CrossRef]
40. Sterling, T.; Irwin, J.J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. Available online: https://pubs.acs.org/doi/10.1021/acs.jcim.5b00559 (accessed on 12 February 2020). [CrossRef]
41. Wishart, D.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef]
42. Barton, W.; O'Sullivan, O.; Cotter, P.D. Metabolic phenotyping of the human microbiome. *F1000Research* **2019**, *8*, 1956. [CrossRef]
43. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. Available online: https://www.kdd.org/exploration_files/p2V11n1.pdf (accessed on 12 February 2020). [CrossRef]
44. Kotthoff, L.; Thornton, C.; Hoos, H.H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn Res.* **2017**, *18*, 1–5. Available online: http://automl.org/autoweka (accessed on 12 February 2020).

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).