

Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics

Nathaniel J. Brittin, Josephine M. Anderson, Doug R. Braun, Scott R. Rajski, Cameron R. Currie, and Tim S. Bugni*



Cite This: *J. Nat. Prod.* 2025, 88, 361–372



Read Online

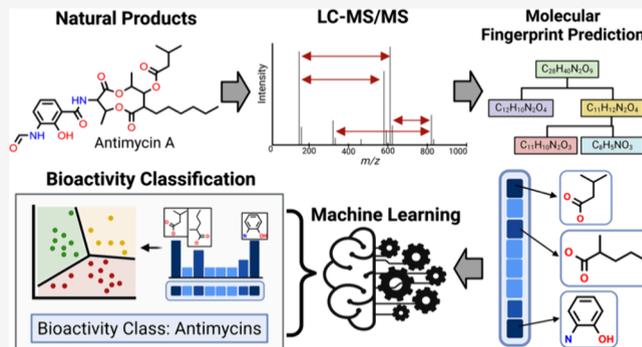
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: The rediscovery of known drug classes represents a major challenge in natural products drug discovery. Compound rediscovery inhibits the ability of researchers to explore novel natural products and wastes significant amounts of time and resources. This study introduces a novel machine learning framework that can effectively characterize the bioactivity of natural products by leveraging liquid chromatography tandem mass spectrometry and untargeted metabolomics analysis. This accelerates natural product drug discovery by addressing the challenge of dereplicating previously discovered bioactive compounds. Utilizing the SIRIUS 5 metabolomics software suite and *in-silico*-generated fragmentation spectra, we have trained a ML model capable of predicting a compound's drug class. This approach enables the rapid identification of bioactive scaffolds from LC-MS/MS data, even without reference experimental spectra. The model was trained on a diverse set of molecular fingerprints generated by SIRIUS 5 to effectively classify compounds based on their core pharmacophores. Our model robustly classified 21 diverse bioactive drug classes, achieving accuracies greater than 93% on experimental spectra. This study underscores the potential of ML combined with MFPs to dereplicate bioactive natural products based on pharmacophore, streamlining the discovery process and expediting improved methods of isolating novel antibacterial and antifungal agents.



The global surge in multidrug-resistant pathogens has intensified the need for new antimicrobial therapies.^{1–3} Natural products (NP) have historically been a rich source of bioactive compounds, with unique organisms from diverse ecosystems providing novel mechanisms and scaffolds to explore.^{4–6} The contribution of NPs, especially to infectious disease treatment where 73% of approved small molecule antibiotics are NPs or use a NP scaffold, has been immense.⁷ However, navigating the vast chemical space of NPs presents significant challenges, primarily the issue of rediscovery.⁸ With hundreds of thousands of NPs already reported, the high probability of reisolating known compounds leads to duplicated efforts and wasted resources, stalling novel discoveries.⁸ Dereplication, the process of distinguishing known compounds from novel ones in complex mixtures, has become crucial in NP discovery.⁹ Although various mass spectrometry (MS) and machine learning (ML) methods have been developed for dereplicating bacterial metabolites, limitations persist, necessitating improved strategies to facilitate dereplication and accelerate the prioritization of novel bioactive scaffolds.^{8,10}

Improved access to comprehensive MS data has aided dereplication efforts. Platforms like the Global Natural Products Social Molecular Networking (GNPS), MassBank,

and METLIN facilitate sharing of MS/MS spectra globally.^{11–13} Complementing these data repositories, computational tools like SIRIUS, DEREPLICATOR+, MS2LDA, and MSNovelist enable database-independent annotation of MS/MS data to identify compounds, tapping into extensive chemical libraries documenting over 200 million unique structures.^{14–19}

Machine learning has emerged as a powerful tool in the analysis of complex NP metabolomics data and the acceleration of NP-based drug discovery campaigns. ML algorithms can extract meaningful insights from the vast chemical space of NPs, surpassing traditional analytical methods in speed and accuracy.²⁰ By integrating diverse data types, including MS/MS spectra, molecular structures, and biological activities, ML approaches create predictive models that guide the prioritization of promising lead compounds.²¹

Received: October 1, 2024

Revised: January 30, 2025

Accepted: January 31, 2025

Published: February 7, 2025



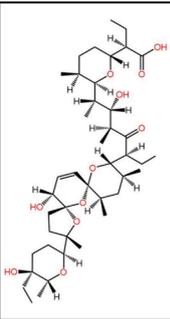
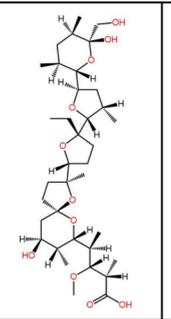
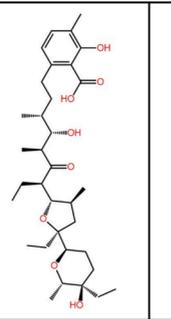
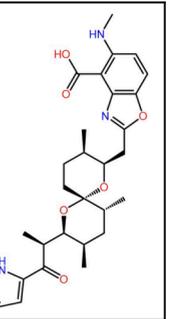
Structure				
Name	Salinomycin	Monensin	Lasalocid	Calcimycin
GNPS ID	CCMSLIB00010111110	CCMSLIB00005723246	CCMSLIB00005489645	CCMSLIB00000085752
Classyfire	Diterpene Glycosides	Heterocyclic Fatty Acids	Diterpene Glycosides	Amino acids and derivatives
NP Class	Open-chain Polyketides	Polyether Ionophores	Open-chain Polyketides	Depsipeptides

Figure 1. CANOPUS classifications of GNPS spectra of compounds within the polyether ionophore antifungal class.

Significant advances have been made in predicting molecular structures from spectral data, automating compound family classification, and enhancing screening of large chemical libraries for potential drug candidates.^{22,23} To ensure a comprehensive evaluation of machine learning approaches, we selected 10 diverse algorithms for this study, encompassing linear models (e.g., Logistic Regression), the historically important perceptron model, tree-based models (e.g., Random Forest), support vector machines, neural networks, and ensemble methods.^{24–27} These algorithms were chosen based on their proven success in similar applications, their ability to effectively process high-dimensional data, and their balance between computational efficiency and predictive performance. As the field continues to evolve, the synergy between ML and metabolomics shows great promise in overcoming traditional bottlenecks in NP drug discovery.

Although machine learning models for natural product analysis would ideally be trained on experimentally collected MS data, current spectral libraries are limited in both size and chemical diversity. For instance, GNPS, the largest repository, contains over 1,011,644 MS/MS spectra representing approximately 56,626 compounds as of December 17th, 2024, but this covers only a fraction of the 400,000 known NPs.^{11,28} Relying solely on these data risks poor generalization to experimental data not represented in the training set. To overcome this limitation, we explored the generation of *in-silico* MS2 spectra as an alternative approach to produce sufficient training data while maintaining relevance to real-world applications.

New approaches in metabolomics focus on methods independent of MS/MS databases to overcome the limitations of finite experimental data. Tools like Mass Frontier and CFM-ID4 generate *in-silico* mass spectra and expand the availability of fragmentation spectra where experimental references are not available.^{29,30} Among newer tools, SIRIUS 5 computes molecular formulas, generates molecular fingerprints (MFPs) of likely substructures, and classifies compounds based on MS/MS spectra.^{14,22,31,32} These *in-silico* fragmentation approaches, combined with SIRIUS's ability to generate MFPs, provide a solution to data scarcity, expanding training sets for machine learning models while maintaining relevance to NP analysis and dereplication. By leveraging these computational tools, we generate a vast array of simulated spectral data and corresponding MFPs, significantly expanding the training set for ML models beyond experimentally available spectra.

Molecular fingerprints play a crucial role in the identification and classification of chemical compounds in many new metabolomics tools.^{17,33–37} These MFPs, consisting of predefined chemical features or substructures, act as a molecular “barcode”, effectively summarizing key features of compounds.^{31,33} By capturing this detailed information in a standardized fixed-length vector, predicted MFPs enable precise structural characterization of fragmentation mass spectra, which is essential for these MFPs to be useful in ML. This precision is particularly valuable in NP research, where the diversity of molecular structures is vast and often includes novel entities.³⁸ Predicting MFPs from MS data enhances dereplication by matching the structural information from compound fragmentation to the information on known NPs.^{14,17} The high variability of small molecule fragmentation and spectrometer type make using fragmentation spectra directly a poor representation in training machine learning models to generalize to other spectra (SI, Table S6). Importantly we hypothesize that, because the MFPs are a summation of MS2 structural information, they will be less impacted by experimental variance and noise and will be more well suited to ML in contrast to using MS2 spectra directly (SI, Figure S6).

Current classification methods for untargeted MS data, such as CANOPUS within SIRIUS 5, utilize specific chemical or biosynthetic features to group compounds based on their structural framework or biosynthetic origins. Although valuable, these methods can be limited by their focus on all aspects of the molecular architecture, which may not strongly correlate with biological activity. As illustrated in Figure 1, CANOPUS classifications can vary significantly with structural modifications outside a compound's core structure, changing the predicted chemical and natural product classifications while the bioactive pharmacophore remains the same. This variability is evident in classifying polyether ionophore antifungals, where the bioactive core structure is not always the final basis for classification. The variations in structure outside the polyether core are the features prioritized such as, for instance, the “Diterpene Glycosides” and “Amino acids and derivatives” in the classification of salinomycin and calcimycin, respectively. This stands in contrast to monensin, which retains its classification as a polyether ionophore. We theorize this can be accounted for by utilizing a pharmacophore-based approach for classification, zeroing in on the functional groups and spatial arrangements essential for a compound's biological

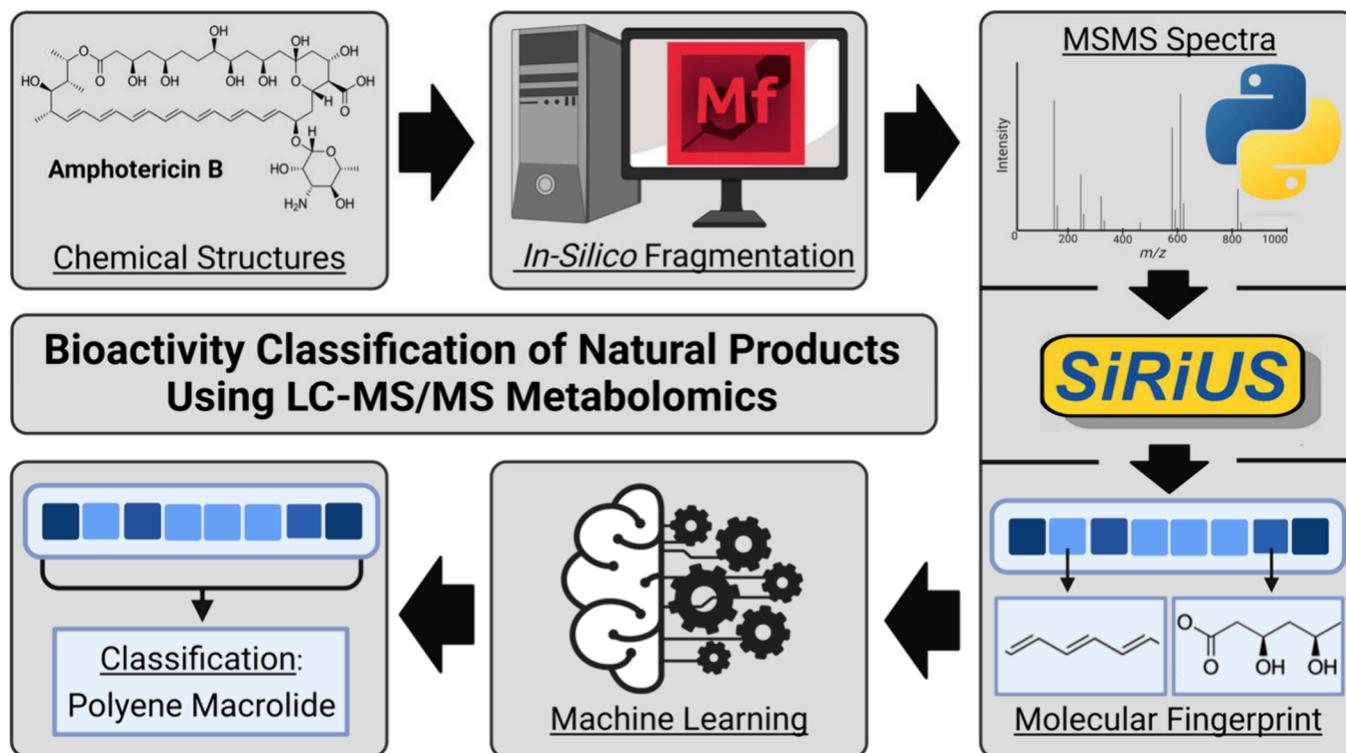


Figure 2. Workflow for generating *in-silico* fragmentation MS from input structures, building their MFs, and using them to train machine learning classifiers of NP bioactivity. Created in BioRender. Brittin, N. (2025) <https://BioRender.com/f78w206>.

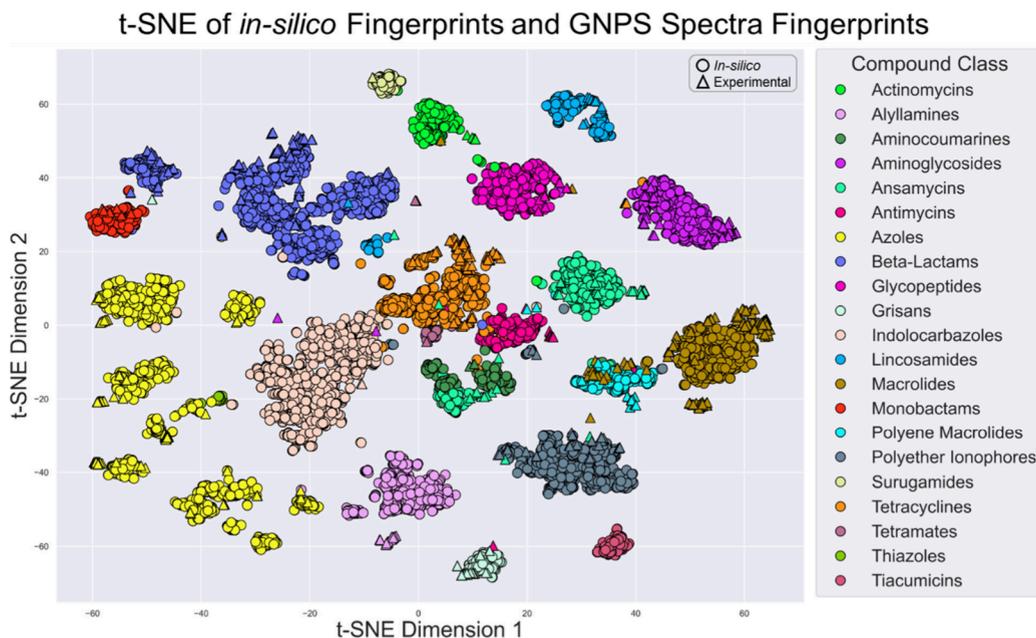


Figure 3. t-Distributed Stochastic Neighbor Embedding (t-SNE) of *in-silico* and GNPS spectra derived molecular fingerprints. The 21 classes of bioactive compounds cluster distinctly, demonstrating that the MFs can provide distinction based on pharmacophore. Additionally, *in-silico* and experimentally based MFs cluster closely indicating similar structural information encoded in them.

activity. With a focus on the pharmacophore, dereplication of known compounds can be expedited, prioritizing the discovery of new drugs with unknown biological effects; such an approach overcomes the limitations of current classification strategies that may overlook crucial bioactive features in favor of broader chemical or biosynthetic considerations. This paper emphasizes features driving bioactivity, providing a more direct

link to therapeutic potential. We hypothesize that pharmacophore classification is less susceptible to variations in noncritical parts of the molecule, making it more robust in predicting bioactivity across structurally diverse compounds.

In this paper, we present a novel ML framework designed to address the most pressing challenges in natural product discovery and dereplication. Our approach explores the

application of *in-silico* generated MS2 spectra and MFPs to overcome/circumvent the limitations of finite experimental data sets and current classification strategies. Specifically, we generated a total of 11,665 *in-silico* MS2 spectra and tested this approach using 9 different ML architectures with a support vector classifier model providing the best performance. Notably, our approach significantly outperformed CANOPUS in classifying compounds within 21 different drug classes, demonstrating accuracy nearly 30% greater than that of current CANOPUS technology. By utilizing pharmacophore-based classification, we aim to provide a more robust and biologically relevant method for identifying and prioritizing natural products.

RESULTS AND DISCUSSION

Comparison of *in-Silico* and Experimentally Derived Fingerprints. To train our ML models, we generated MFPs from *in-silico* MS2 spectra. This process, outlined in Figure 2, involved simulating compound fragmentation using Mass Frontier software, then using SIRIUS 5 to create MFPs from these simulated spectra. The classification in this study is based on pharmacophore groupings rather than hierarchical structure-based methods like ClassyFire or NP Classifier. Each class was built using 5–8 representative compounds and expanded through PubChem similarity searches (Tanimoto score ≥ 98), ensuring a biologically relevant classification focused on pharmacophore features rather than structural taxonomy. The resulting data set was used to train our models. To assess the similarity of the MFPs produced from *in-silico* generated mass spectra and those derived from experimental spectra, MFPs were generated with SIRIUS for 8,521 *in-silico* spectra generated using Mass Frontier and 1,256 GNPS spectra. To properly summarize the information within the high dimensional data set, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to project the high-dimensional fingerprint data onto two dimensions for cluster visualization.³⁹ The t-SNE projection in Figure 3 reveals distinct and well-separated clustering of the fingerprints according to their respective drug classes. The separate, cohesive clusters demonstrate that the MFPs effectively encode the unique structural features and patterns characteristic of each pharmacophore scaffold. Some MFPs, especially experimental fingerprints from GNPS spectra, were outside of their respective cluster.

Importantly, fingerprints derived from the *in-silico* fragmentation data, *in-silico* MFPs, clustered cohesively with those generated from real experimental GNPS spectral data for the same class. This colocalization highlights the high degree of integrity and accuracy with which *in-silico* MFPs effectively summarize key structural features from each of the 21 classes. This t-SNE visualization

provides a powerful confirmation that MFP patterns do, indeed, distinguish between the diverse drug pharmacophores.

Generating Training and Testing Sets of *in-Silico* MFPs for Polyene Macrolide Classification Model. To demonstrate our proof-of-concept on a single bioactive antifungal class, polyene macrolides were selected; the problematic characteristics of polyenes during isolation, characterization and application are well established.^{40–44} Polyenes tend to be highly bioactive at low concentrations and degrade during LC-MS/MS data collection, leading to difficulties in dereplication. Polyene antifungals also present a difficult target for dereplication due to low spectra diversity

and abundance in publicly available databases with only 37 spectra covering 4 unique compounds available in GNPS to date. To compile a diverse training set, eight polyene antifungals known for their antifungal activity (SI, Table S3) were expanded via PubChem similarity searches to retrieve all similar analogues above a Tanimoto similarity score of 0.98, resulting in a set of 366 unique structures (SI, Table S7). *In-silico* fragmentation spectra were generated using Mass Frontier 8.1 and processed with SIRIUS 5 to create molecular fingerprints, forming the positive training set. For the negative data set, 2,778 compounds were selected from the RIKEN Natural Products Depository. Notably, this repository contains both natural and synthetic derivatives, characterized mechanisms of action within the MOSAIC database, and coverage of a large chemical space (SI, Figure S7);^{45,46} special care was taken to avoid any overlap with polyene antifungals. The combined data set was split into an 80–20 ratio for training and testing and ten ML models were evaluated using this data set.

As seen in Table 1, each model displayed high metrics for learning the training data. The models were evaluated using

Table 1. Results of Polyene Macrolide Trained ML Classifier Models Evaluated on Test Data (20% Held-out Set)^a

Model	Accuracy	Precision	Recall	F1	FPR (%)
Passive Aggressive Classifier	0.9968	0.9733	1.0000	0.9865	1.556
Ridge Classifier	0.9968	0.9863	0.9863	0.9863	1.707
MLP Regressor	0.9968	0.9863	0.9863	0.9863	2.572
Random Forest Classifier	0.9968	1.0000	0.9726	0.9861	0.540
Support Vector Classification	0.9952	0.9730	0.9863	0.9796	1.129
Logistic Regression	0.9936	0.9726	0.9726	0.9726	1.669
Perceptron	0.9936	0.9859	0.9589	0.9722	0.508
Decision Tree Classifier	0.9841	0.9565	0.9041	0.9296	2.321
Gaussian Naïve Bayes	0.9825	0.8875	0.9726	0.9281	5.484

^aResults of polyene trained ML models on test set data using accuracy, precision, recall, and F1 score. The false positive rate of each model when tested on 5000 random GNPS spectra.

accuracy, precision, recall, and F1. Accuracy measures the number of correctly identified samples over the total samples. Precision measures true positives over the total true and false positives. Recall measures true positives over the total true positive and false negatives. F1 is the harmonic mean of the precision and recall values and allows for prioritization of a model that has a balance between false positive (FP) and false negatives (FN). The top four models, the Passive Aggressive Classifier, the MLP Regressor, the Ridge Classifier, and Random Forest Classifier, achieved nearly identical accuracies and F1 scores, yet differed by recall and precision. The Passive Aggressive Classifier excelled at recall and predicting with no FN predictions. Conversely the Random Forest Classifier demonstrated high precision with no FP predictions. The Ridge Classifier and MLP Regressor both demonstrated a more balanced profile with a high F1 score despite having some FP and FN. Overall, all models displayed high levels of learning on the training set and high evaluation metrics on the 20% held-out testing data.

To test the specificity of classification and ensure that the models can accurately distinguish between polyene and nonpolyene MFPs, each model was tested against 5000 randomly selected spectra from GNPS. These included no polyene spectra and were processed with SIRIUS 5 to collect a max of 10 formula predictions for each yielding a total of 15,938 fingerprints. Each model was evaluated to determine how likely the models falsely predicted a compound as a polyene; the false positive rate (FPR). The FPR of the models ranged from 0.508% to 5.484%, with the Random Forest and Perceptron models showing around a 0.50% FPR and most models having between 1 and 2.5% FPR (Table 1).

Each model was confirmed to perform well on the same *in-silico* data it was trained on and generated classifications specific to polyenes, but it was most important that each model generalized to experimental spectra. Each model's performance on experimental spectra was then evaluated using 37 polyene antifungal spectra from GNPS. The spectra were processed using SIRIUS 5 with either the known molecular formula, five predicted molecular formulas, or 10 predicted molecular formulas. Evaluation of spectra from known formulas provided an ideal scenario. Use of predicted molecular formulas emulated the results from untargeted metabolomics methods where the compound identity was unknown. The K-Neighbors Regressor performed best, identifying 100% of the GNPS spectra correctly as polyenes and identifying 71% and 67% of predicted fingerprints when multiple formulas were predicted. Most models showed accuracies

between 78 and 95%, dropping to 45–70% on predicted formula MFPs (Table 2). These studies showed that high accuracy in predicted formulas is crucial for applications in untargeted identification since the error in mass accuracy can result in the correct molecular formula not being the one with the lowest error. Therefore, despite only one correct formula among multiple predictions, the models demonstrated flexibility in correctly identifying the pharmacophore class.

Table 2. Results of Polyene Macrolide Trained ML Classifier Models Tested on 37 Polyene Macrolide GNPS Spectra^a

Model	Accuracy		
	Known Formula	5 Predicted Formula	10 Predicted Formula
K-Neighbors Regressor	1.0000	0.7182	0.6779
MLP Regressor	0.9730	0.7072	0.6667
Logistic Regression	0.9459	0.6298	0.6050
Passive Aggressive Classifier	0.9459	0.6243	0.5966
Support Vector Classifier	0.9459	0.6298	0.5938
Gaussian Naïve Bayes	0.8919	0.5138	0.4034
Ridge Classifier	0.8919	0.7072	0.6415
Perceptron	0.7838	0.4972	0.4566
Random Forest Classifier	0.4595	0.3315	0.3109
Decision Tree Classifier	0.2432	0.1934	0.1765
CANOPUS	0.9189	0.5028	0.4678

^aResults of polyene macrolide trained ML classifier models tested on 37 polyene macrolide GNPS spectra using the known molecular formula, five predicted molecular formula, and 10 predicted molecular formula. Additionally, the performance for SIRIUS 5's tool, CANOPUS, on the same data sets.

As shown in Table S5 (SI), the number of fingerprints identified as polyene macrolides increased with more predicted fingerprints, surpassing the 37 fingerprints from the known formula.

Generating a Data Set of *in-Silico* Molecular Fingerprints for Training and Testing Multiclass Classification Model for 21 Bioactive Drug Classes. Given the high performance of the binary classifier for polyenes, we next compiled a multiclass data set for 21 different classes of bioactive NPs. The classes created represented monobactams, actinomycins, beta-lactams, indolocarbazoles, cyclic peptides, azoles, tetracyclines, aminocoumarins, allylamines, tiacumicins, aminoglycosides, polyether ionophores, polyene macrolides, lincosamides, macrolides, grisans,^{47,48} antimycins, ansamycins, surugamides, thiazoles, and tetramates.

To establish each drug class around a bioactive pharmacophore, between five to ten representative structures were selected for each class, which were then expanded with compounds obtained from PubChem similarity searches with a Tanimoto score of 98. Using structural similarity allows for each drug class to be built outward from the core pharmacophore containing compounds. After removing duplicates, a total of 8,521 unique structures were retrieved representing the 21 drug classes (SI, Table S1). For these structures, *in-silico* MS/MS spectra were created for each using Mass Frontier 8.1, processed through SIRIUS 5 to generate molecular fingerprints, and split into 80/20 training/testing sets. Nine different multiclass classifiers were trained and evaluated on the held-out test set (Table 3).

All trained models were evaluated using accuracy, precision, recall, F1, and Matthew's correlation coefficient (MCC). MCC is a measure of the quality of classifications, providing a single value score from -1 to 1 that reflects model performance using all four basic rates (TP, FP, TN, FN).⁴⁹ All models demonstrated strong learning capabilities, with eight of the nine achieving above 94% accuracy, precision, recall, and F1 scores on the test set. The logistic regression model was the optimal performing model; all metrics for the logistic regression model exceeded 98%

indicating the model is easily capable of distinguishing between classes. Most other models also demonstrated metrics over 96%, highlighting their robust multiclass predictive abilities.

As shown in Figure S5, the logistical regression model showed 0 false positives in 11 of the 21 classes. Specifically, the logistic regression model showed <1% of false negatives between the negative class and the larger classes like beta-lactams, azoles, and polyether ionophores, as well as 1–2.7% of false negatives with the antimycin, indolocarbazole, monobactam, tetracycline, and polyene classes. Notably, these classes typically had significantly fewer representative structures within the training set. The logistic regression model's high precision score demonstrates low false positives, with 11 of the 21 classes having 0 FPs and the remaining classes all equal or less than 0.5% excluding the beta-lactam and indolocarbazole classes with 1.3% and 1.1% false positives respectively. The other models were found to perform similarly except for the decision tree model, which performed markedly worse than the other 8 model types.

To determine the specificity of each multiclassification model, a data set of 9,443 random GNPS spectra, excluding spectra from the 21 drug classes, was tested to observe any falsely classified spectra. These MFPs represent a vast diversity

Table 3. Results of Bioactive Drug Class Multiclass Classification ML Models on the 20% Held Out Testing Data^a

Model	Accuracy	Precision	Recall	F1	MCC	Total FPR	Avg. FPR
Logistic Regression	0.9858	0.9861	0.9858	0.9858	0.9842	10.717%	0.5103%
Ridge Classifier	0.9832	0.9834	0.9832	0.9832	0.9812	11.691%	0.5567%
PA Classifier	0.9823	0.9825	0.9823	0.9823	0.9802	10.039%	0.4781%
Perceptron	0.9801	0.9804	0.9801	0.9801	0.9777	12.782%	0.6087%
Support Vector Classifier	0.9788	0.9792	0.9788	0.9787	0.9763	5.528%	0.2632%
SGD Classifier	0.9765	0.9776	0.9765	0.9765	0.9738	8.652%	0.4120%
MLP Classifier	0.9695	0.9703	0.9695	0.9695	0.9660	14.116%	N/A
K-Neighbors Classifier	0.9473	0.9514	0.9473	0.9468	0.9423	18.236%	N/A
Decision Tree Classifier	0.8876	0.8907	0.8876	0.8884	0.8743	22.715%	N/A

^aResults of trained multiclass classification models on the 20% held-out testing data. Total and average FPR of each model on 9,443 random GNPS spectra without any representatives of the drug classes trained on.

of unrelated chemical structures that should be classified as negatives by the models. As shown in Table 3, the total FPRs ranged from 5.5% to 22.7%, indicating there is a small to moderate number of unrelated spectra being falsely classified as belonging to our classes of interest. However, since many of these models are trained in a “One-versus-Rest” or “One-versus-All” training scheme, each class is assigned an individual binary classifier. Each classifier is then trained separately, so the average FPR per classification model gives insight into each decision’s performance. The average FPR, in Table 3, for each class ranged from 0.261% to 1.08%, a significant reduction compared to the binary polyene classifier with its single class. Interestingly, the Support Vector Classifier (SVC), while not having the highest metrics, had the lowest total and average false positive rates (5.5% and 0.261%, respectively); 2.5% lower than the logistic regression model. Overall, these models effectively discriminated against fingerprints lacking the key features associated with the 21 drug classes with average FPRs below 0.5% and performance metrics above 97%.

Multiclass Classification Model Performance on Experimental MS/MS Data. To evaluate the performance of the trained multiclass classification models on real experimental data outside the training set, models were tested using 1,256 spectra from across the 21 drug classes in the GNPS public data repository (Table 4). The molecular fingerprints were generated by SIRIUS 5 using the known molecular formulas of these GNPS spectra.

Table 4. Results of Bioactive Drug Class Multiclassification ML Models Testing on 1,256 GNPS Spectra^a

Model	Accuracy	Precision	Recall	F1	MCC
SVC	0.9358	0.9469	0.9358	0.9389	0.8791
SGD Classifier	0.9032	0.9249	0.9032	0.9102	0.8218
Ridge Classifier	0.8909	0.9229	0.8909	0.9009	0.8101
Passive Aggressive Classifier	0.8879	0.9144	0.8879	0.8965	0.7985
MLP Classifier	0.8780	0.9228	0.8780	0.8919	0.7963
Logistic Regression	0.8852	0.9104	0.8852	0.8935	0.7958
Perceptron	0.8677	0.9104	0.8677	0.8805	0.7694
K-Neighbors Classifier	0.8400	0.9116	0.8400	0.8611	0.7494
Decision Tree Classifier	0.7388	0.8350	0.7388	0.7751	0.5783
CANOPUS	0.6301	N/A	N/A	N/A	N/A

^aResults of trained models on the fingerprints of 1,256 GNPS spectra using the known molecular formula.

The SVC model achieved 93.58% accuracy, precision of 94.69% (which indicates a low rate of false positives), and an MCC score of 0.879 indicating high performance for all basic rates. The Ridge Classifier and SGD Classifier also exhibited strong performance, with accuracies over 89%, F1 scores around 90%, and MCC scores above 0.8. In contrast, the Decision Tree Classifier struggled considerably on this diverse GNPS data set, achieving only 73.8% accuracy; this reduced MCC of 0.6543 indicated relatively poor predictive performance compared to the top models. Relative to CANOPUS’s ability to distinguish the correct drug class, the SVC model exhibited a 30.5% increase in accuracy. Because CANOPUS evaluates compound class using the hierarchical classifications of ClassyFire and NP Classifier, we needed to map its classifications to the pharmacophore groupings used in this study. To determine what constitutes a correct classification by CANOPUS, we included all relevant hierarchical classes that could align with the assigned bioactivity class, even though some of these classes do not specifically describe a bioactive pharmacophore (SI, Table S8).

Evaluating Multiclassification Models on Complex Bacterial Extracts. To evaluate the accuracy of models on complex LC-MS/MS samples, 25 bacterial extract fractions displaying activity against *C. albicans* and suspected to contain polyenes, were selected. After fermentation, a two-step chromatographic approach was employed to array molecules into 96-well plates over a total of 80 fractions. The moderately purified fractions demonstrating strong activity against *C. albicans* were prepared, and then analyzed by LC-MS/MS. MFPs were compiled from a SIRIUS 5 analysis for predictions as well as the polyene classifications from SIRIUS 5’s CANOPUS tool. The SVC model was chosen to evaluate all fingerprints within the LC-MS/MS data due to the highest metrics on GNPS spectra and lowest FPR. Overall, the model identified polyenes in 19 of the 25 samples (Figure 4). These active fractions were simultaneously analyzed by UV–vis spectroscopy to confirm the presence of polyenes using their unique UV absorbance pattern (Figure S2). The SVC model revealed a significant number of MFPs classified as polyenes; this approach revealed nearly 4 times as many “polyene samples” as the CANOPUS classifier with nearly twice as many MFPs classified on average per sample.

Discussion of Polyene Antifungal Binary Classifier Results. The results derived from the polyene classifier demonstrate high metrics in distinguishing polyene antifungals from a diverse array of compounds. Specifically, the MLP Regressor model displayed exceptional performance in generalizing to experimental spectra, correctly identifying 97.3% of

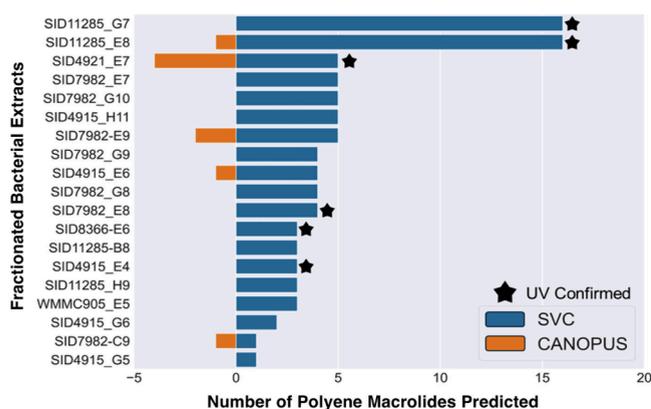


Figure 4. Results of the multiclassification SVC model predictions for complex bacterial extract LC-MS/MS data of bioactive antifungal samples. Samples labeled with a star are confirmed to contain polyenes based on their unique UV absorbance patterns (Figure S2).

polyene MFPs from known molecular formulas. This level of accuracy underscores the model's robust capability to recognize polyene molecular fingerprints. Notably, the model maintained high classification rates despite predicting from multiple predicted molecular formulas of which only one for each spectrum was correct, achieving 70% and 66% accuracy for 5 and 10 predicted formulas, respectively. The ability to correctly identify fingerprints, despite the underlying molecular formula being slightly incorrect, is an important feature; this enables models to function with MFPs from spectra with lower mass accuracies if necessary. This level of performance, especially in scenarios mimicking real-world untargeted metabolomics workflows, highlights the model's adaptability and its potential to streamline the identification of polyenes in complex biological matrices.

The significance of these predictions lies not only in the high accuracy of the polyene classifier but in its comparison with current classification standards. For instance, the SVC model's ability to outperform CANOPUS—a tool within SIRIUS that is trained on GNPS database spectra—by 7% on polyene fingerprints and nearly 20% in predicted polyene fingerprints (Table 2) highlights this model's predictive power. Additionally, the fact that most of the models can achieve superior classification rates without having been trained on the GNPS spectra directly speaks to the effectiveness of the *in-silico* fragmentation spectra and molecular fingerprinting approach.

Furthermore, the polyene classifier's low FPR reinforces its specificity and reliability in identifying polyene antifungals amidst a plethora of nonpolyene compounds. This precision is crucial for reducing the likelihood of misidentification when screening complex natural extracts for potential antifungal agents and thereby focusing efforts on the exploration of novel bioactive entities.

Multiclassification Model Considerations. The support vector classifier, designed to classify compounds into 21 distinct bioactive classes based on molecular fingerprints derived from *in-silico* fragmentation spectra, demonstrated exceptional capability in navigating the complex chemical space of NPs. The high accuracies achieved across 21 diverse bioactive families underscore the model's proficiency in capturing the unique structural characteristics indicative of each class's pharmacophore.

The success of the multiclassification approach is largely credited to the comprehensive data set of molecular finger-

prints employed during its training phase. This data set has enabled the model to distinguish between subtle differences among classes effectively as reflected by low rates of misclassification and minimized false positives. Specifically, the analysis of confusion matrices highlights the fact that rarer classes with fewer training examples—such as polyenes, monobactams, and aminocoumarins—were more susceptible to misclassification, indicating the impact of training diversity on model performance. The SVC model maintained high precision, recall, and F1 scores, suggesting that, even with a 0–1.3% occurrence of false negatives and a similar range for false positives, the model's overall classification accuracy is robust.

The instances of misclassification, particularly among classes with fewer representative structures in the training set, underscore the potential for enhancing model accuracy by expanding the data set to include a more diverse array of structural examples. This expansion could particularly benefit underrepresented classes, potentially reducing classification ambiguities. Despite these challenges, the consistently high accuracies and balanced metrics across all models underscore the algorithms' power in discerning spectral fingerprint patterns specific to each bioactive class's pharmacophore.

Remarkably, each model generalized well to an external test set comprising GNPS experimental spectra, thereby affirming their ability to learn structural patterns within the *in-silico* molecular fingerprints comparable to real experimental data. This validation emphasizes the SVC model's potential as a tool for NP research, capable of facilitating the identification of known compounds and prioritization of novel compounds with novel mechanisms. Moreover, the distinct interclass clustering and colocalization of *in-silico* and experimental data observed in the t-SNE analysis visually corroborate the model's ability to encode and discriminate bioactive NP families based on mass spectral fingerprint data.

When the SVC model was applied to LC-MS/MS data of complex fractionated bacterial extracts, which were antifungal and suspected of containing polyenes, it was able to identify 4 times as many bacterial extracts containing polyenes as CANOPUS and nearly twice as many MFPs as polyenes within the same samples. Validating this model's amenability to complex samples could drastically improve detection of these classes.

CONCLUSION

In this study, we present a powerful approach for rapidly characterizing bioactive NPs directly from metabolomics data using machine learning models trained on molecular fingerprints derived from *in-silico* fragmentation spectra. By leveraging the structural information encoded within these fingerprints, our models effectively learned to discriminate between diverse families of antibacterial and antifungal NP scaffolds based on their unique spectral patterns.

The binary classification model for identifying polyene macrolide antifungals demonstrated superior performance compared to existing methods like CANOPUS on GNPS spectra MFs generated using known molecular formulas. Crucially, it also excelled at classification even when using predicted formulas in the fingerprint generation—a scenario that closely mimics real-world untargeted metabolomics workflows. This model's low FPR further reinforces its reliability for dereplicating these potent antifungals in complex samples.

Expanding our approach to a multiclassification model spanning 21 diverse bioactive families commonly encountered in NPs, we achieved remarkable accuracies of 93% on the GNPS experimental data set using the top-performing support vector classifier. Additionally, the support vector classifier demonstrated high specificity, with average class FPR below 0.26% on a data set of around 9,500 unrelated fingerprints.

The success of our machine learning framework hinges on the ability of the molecular fingerprints to effectively encode the key structural features associated with each bioactive scaffold's pharmacophore, enabling reliable identification of shared bioactivities. This approach removes some of the limitations of existing methods that rely solely on chemical classification, which can overlook bioactivity relationships when structural modifications occur outside the core pharmacophore.

By facilitating rapid dereplication of known bioactive scaffolds directly from metabolomics data, our machine learning models represent a powerful tool for accelerating the discovery of novel NPs. As public repositories like GNPS continue to expand, the adaptability of our approach ensures it can scale to incorporate new and emerging data sets. The reliance on *in-silico* generated training data, rather than experimental spectra, provides the flexibility to add new pharmacophore classes by assembling representative structures. This strategy can streamline the identification of genuinely new chemical entities, minimizing the redundant investment of resources in reisolating known compounds. Overall, this work demonstrates the transformative potential of machine learning coupled with the utilization of molecular fingerprints derived from *in-silico* based fragmentation MS; the synergies enabled by this coupling enable efficient dereplication of known bioactive NP classes and facilitated prioritization of novel bioactive NP scaffolds from complex extracts.

■ EXPERIMENTAL METHODS

Utilization of PubChem for Expansion of Structure Set Using Similarity Scoring. An initial collection of bioactive compounds was identified from literature reviews, with their structures represented by SMILES strings. These SMILES strings were input into PubChem's similarity search to find analogous structures with a Tanimoto score of 98 or higher. Python was used to compile, deduplicate, and remove stereochemistry from these matches (RDKit), ensuring uniqueness. Additionally, compounds stored in PubChem as charged salts, acids, or ions were simplified by stripping these components before final deduplication (RDKit).

Selection of GNPS Data Sets for Model Evaluation. To evaluate the machine learning models, MS/MS spectra were retrieved from the GNPS database using both targeted and random selection approaches to ensure a representative chemical space while maintaining computational feasibility. Targeted data sets were created by searching for specific compound names within the GNPS library. This approach was used to compile data sets for comparisons between *in-silico* and experimentally derived molecular fingerprints and evaluations of particular compound classes, such as polyene antifungals. 37 spectra for polyene antifungal compounds available in GNPS were selected to assess binary classifier performance on this specific bioactive class. 1,256 spectra were retrieved to compare molecular fingerprints generated from *in-silico* MS/MS spectra with those derived from experimental spectra. It is acknowledged that some compounds may have been missed due to some compounds having been missed.

Random data sets were used for broader evaluations, specifically false positive testing and to ensure diverse representation of the GNPS chemical space. 5,000 randomly selected spectra were used to evaluate the binary classifier's ability to identify polyene macrolides,

while a larger data set of 9,443 random spectra was compiled to assess the specificity of the multiclassification models. Random selection was performed programmatically using Python, with a random seed to ensure reproducibility. At the time of data set generation in April 2024, the GNPS database contained approximately 40,000 unique compounds. A nonexhaustive approach was adopted to avoid processing the entire database of around 600,000 spectra, which would have imposed significant computational burdens and potential redundancy in the data. Instead, the random selection was designed to encompass sufficient chemical diversity, capturing a broad representation of the GNPS library while avoiding unnecessary overlap.

All selected spectra were processed through SIRIUS 5 to generate molecular fingerprints. Spectra that failed to process due to insufficient fragmentation were excluded, resulting in slightly fewer spectra than initially intended. This workflow ensured that the data sets used for training and evaluation were robust, representative, and reproducible.

Generating Negative Training Examples from RIKEN NP Depo. Negative training examples were sourced from the RIKEN Natural Products Depository (RIKEN NP Depo) due to its diversity of natural products and synthetic derivatives, providing a meaningful counterpoint to the positive data set. Compounds in the RIKEN NP Depo have been evaluated by the MOSAIC chemical-genetic repository, ensuring quantified biological activity data.

To create the negative data set, structures were screened to exclude compounds from the 21 drug classes to remove any overlap with the positive classes. Then a subselection of the data set resulted in 2,778 unique compounds, offering broad structural diversity to enhance the model's robustness. The entire data set was not utilized to reduce data handling and computational time. Alternative data sets such as ZINC, DrugBank, COCONUT, and The NP Atlas may be explored in future studies for additional flexibility.

Mass Frontier for Batch Fragmentation of Compound Structures. Mass Frontier 8.1 v.8.1.80.8 (Thermo Fisher Scientific) was used for the batch fragmentation of structures to generate sets of predicted fragments to convert into *in-silico* mass spectra. First, the SMILES string for each compound was used in RDKit to create Structure-Data Format (SDF) files. Each molecule's structure was controlled for compatibility with the software, focusing on representing compounds in their neutral forms for accurate predictions. The Batch Fragment Generation was set to the Protonation method to mimic electrospray ionization (ESI) as would be seen in LC-MS/MS collection in positive ion mode. This method's extensive rule database allowed for the prediction of fragmentation patterns, specifying ion types, charge states, and considering radical ions to approximate experimental scenarios closely. All settings for cleavage type, rearrangements, charge retention reactions, and resonance were left unaltered. The maximum number of reaction steps set to 5 and the maximum resonance number set to 2 and the mass range for fragment generation was set from 50 to 1550 Da. Finally, the reactions limits were set to a maximum number of reactions to 3,000 with a maximum number of unique fragments set to 60. There can only be a maximum of 60 fragments to ensure that SIRIUS is able to process all provided fragments with its inborn max limit of 60 signals accepted per MS2 spectrum.

Generating the MS Files for SIRIUS Using Python. The outputs of Mass Frontier 8.0, the Structure-Data Format (SDF) files containing the computed fragments, were imported into Python using the RDKit Python package. Each SDF file had the individual fragments extracted and processed to retrieve the exact mass for each fragment. The masses were then deduplicated and arranged into unique MS-Format files (.ms) for SIRIUS 5 as seen in SIRIUS 5's acceptable input formats.

Analyzing *in-Silico* Mass Spectra with the Known Molecular Formula of Each Compound. SIRIUS 5 was employed to analyze each mass spectrum and fragmentation pattern, returning a fingerprint for each compound. The settings used for processing all the *in-silico* files were the default settings for each module. When only one formula (the known formula) was desired for prediction, the formula

was specified in the MS files and the SIRIUS settings were set to only predict 1 formula.

Compiling in-Silico Fingerprints into a Training Matrix. The final step involved collecting the molecular fingerprints generated by SIRIUS for each compound in the data set. Each fingerprint is a MACCS Key style fingerprint with 3,878 unique substructures for which probabilities are predicted. These fingerprints were compiled into a simple table. Python scripts were used to iterate over each individual sample in the SIRIUS project, extract the fingerprint, and add it to the overall table. Metadata such as the name of the sample the fingerprint came from were included in the table to ensure no mixing of fingerprints. This table serves as the input for machine learning models, with each row in the table being a unique fingerprint for every compound.

Utilizing the Scikit-Learn Python Package for Training and Testing Set Generation. Scikit-learn was used for creating the training and testing sets of the fingerprint data as well as for the initialization and testing of each model type. The primary metrics used in the evaluation of each machine learning model were precision, recall, F1, and accuracy. The polyene fingerprints were combined with the negative example fingerprints from the RIKEN NP Depo set and binarily labeled. The data set was split using an 80% and 20% ratio for the training and testing sets, respectively. For the multiclassification data set the MFPs were combined with the diverse negative set and labeled with their respective drug class or “negative”. Each multiclassification model was run with controlled random state and a decision function of “One-versus-Rest” or “One-versus-All”.

LC-MS/MS Method for Bacterial Extracts. Liquid chromatography tandem mass spectrometry (LC-MS/MS) data were acquired using a Bruker maXis II Ultra-High-Resolution LC-QTOF mass spectrometer coupled to a Waters Acquity H-Class UPLC system and operated by the Bruker HyStar 3.2 software. Chromatographic gradients were performed with a mixture of methanol and water (containing 0.1% formic acid) on an RP C-18 column (Phenomenex Kinetex 2.6 μm , 2.1 mm \times 100 mm) at 0.3 mL/min. The method was as follows: 0–1 min (10%–10% MeOH in H₂O), 1–12 min (10%–97% MeOH in H₂O), and 12–15.5 min (97% MeOH in H₂O). A mass range of m/z 50–1550 was measured in positive ESI mode for all spectra. The mass spectrometer was operated with the following parameters: capillary voltage of 4.5 kV, nebulizer pressure of 1.2 bar, dry gas flow of 4.0 L/min, dry gas temperature of 205 °C, and scan rate of 2 Hz. Tune mix (ESI-L low concentration; Agilent) was introduced through a divert valve at the end of each chromatographic run for automated internal calibration. MS/MS spectra were acquired at scan speeds of 2 Hz for signals above 1×10^4 counts and 6 Hz for signals above 1×10^6 counts. MS/MS spectra were collected using a stepping collision energy (CE) where CE increased linearly during MS/MS spectra collection. From time 0 to 32, the collision RF was 600, transfer time was 80, and CE was 70 eV. From time 33–66, the collision RF was 600, transfer time was 72, and CE was 100 eV. From time 67–100, the collision RF was 600, transfer time was 65, and CE was 130 eV. The precursor list was set to exclude precursor ions for 0.2 min after two spectra with the same precursor ion were acquired. Additionally, if the intensity of an excluded precursor ion rose 5-fold from the initial spectrum, it would be recollected.

t-SNE Visualization of Molecular Fingerprints. The *in-silico* molecular fingerprints used in training the machine learning models were combined with the molecular fingerprints of the GNPS spectra. Utilizing the t-SNE tool within the Scikit-Learn Python package the unlabeled molecular fingerprints were fit to a 2D embedding using the default parameters with the perplexity increased to 50 and the number of iterations at 750. The plot was generated using matplotlib.

Fermentation for Library Generation. For each prioritized strain, 10 mL seed cultures (25 \times 150 mm tubes) in medium DSC (5 g soluble starch, 10 g glucose, 5 g peptone, 5 g yeast extract per liter made with 50% artificial seawater) were inoculated and shaken (200 rpm, 28 °C) for 7 days. Seed cultures (2.5 mL) were used to inoculate 3 \times 100 mL of media in 500 mL baffled flasks using two distinct media (2 \times 100 mL ASW-A and 100 mL RAM2) containing Diaion HP20 (7% by weight). ASW-A was made using 20 g soluble starch, 10

g glucose, 5 g peptone, 5 g yeast extract, 5 g CaCO₃ per liter of artificial seawater; RAM2 was made using 4 g corn meal, 10 g glucose, 15 g maltose, 7.5 g Pharmamedia, 5 g yeast per liter of 50% artificial seawater. After fermentation for 7 days, the cells and HP20 were filtered using Miracloth, and the cells and HP20 were extracted with acetone (100 mL for 30 min).

Library Generation. The crude extract was dried and then dissolved using the following solvent mixture: 1 mL dimethyl sulfoxide (DMSO), 1 mL methanol, and 10 mL H₂O. Subsequently, the mixture was fractionated on an Isolute ENV+ (500 g cartridge) using a modified Gilson GX-271 liquid handler with 100% H₂O (10 mL), 25% CH₃OH/H₂O [fraction 1], 50% CH₃OH/H₂O [fraction 2], 75% CH₃OH/H₂O [fraction 3], 100% CH₃OH [fraction 4] (8 mL of each solvent). The 100% water fraction went directly to waste while the remaining four fractions were collected and subsequently dried in a speedvac. Each fraction was dissolved in DMSO and subjected to HPLC using a Gilson HPLC integrated with a Gilson 215 fitted with a 96-well plate deck capable of holding ten plates. For HPLC, a Phenomenex Monolithic C18 column (3 mm ID \times 100 mm) was used. The following HPLC gradients were used:

Fraction 1 (F1)

- 0–2 min, hold at 90% H₂O/10% CH₃CN
- 2–14.5 min, ramp to 50% H₂O/50% CH₃CN
- 14.5–19 min, ramp to 100% CH₃CN
- 19–22 min, hold at 100% CH₃CN
- 22–27 min ramp to 90% H₂O/10% CH₃CN

Fraction 2 (F2) and Fraction 3 (F3)

- 0–2 min, hold at 90% H₂O/10% CH₃CN
- 2–19 min, ramp to 100% CH₃CN
- 19–21.5 min, hold at 100% CH₃CN
- 21.5–22 min, ramp to 90% H₂O/10% CH₃CN
- 22–27 min, hold at 90% H₂O/10% CH₃CN

Fraction 4 (F4)

- 0–2 min, hold at 90% H₂O/10% CH₃CN
- 2–5 min, ramp to 70% H₂O/30% CH₃CN
- 5–19 min, ramp to 100% CH₃CN
- 19–32 min, hold at 100% CH₃CN
- 32–32.5 min, ramp to 90% H₂O/10% CH₃CN
- 32.5–37.5 min, hold at 90% H₂O/10% CH₃C

For each fraction above, 20 fractions were collected in 96-deepwell plates such that, for each extract, metabolites were arrayed across a total of 80 wells. The plates were then dried in a speedvac and DMSO (20 μL) was added to each well to dissolve the material. The contents were then transferred to Labcyte Echo plates prior to high-throughput screening.

High Throughput Screening. Next, *in vitro* high-throughput screening was applied to these HPLC purified fractions using a four-point dose response in 384 well plates with an Echo 550 acoustic droplet delivery system against *Candida auris*. Assay plates for antimicrobial testing are made ahead of time, using the Echo 550 acoustic liquid handler. 500, 250, 100, and 50 nL of natural product fraction were transferred to each quadrant of a clear 384 well plate. The following control was used for *C. albicans* (Amphotericin B 0.5 mg/mL). To prepare the test organism, a single colony of *C. albicans* was picked from a solid agar plate into 5 mL of a liquid culture and was grown for 18 h shaking at 37 °C. This culture was diluted to 0.5 McFarland units, and this stock was further diluted 1:300 for use in HTS assays. Fifty μL per well of the diluted culture was added to each well of the 384 well assay plate using the Thermo-fisher Multidrop instrument. Microorganisms are incubated with the compound overnight at 37 °C. Microorganism growth was measured by collecting an end point absorbance reading at OD₆₀₀ using a BMG CLARIOStar plate reader.

Computation of the Training Compound Chemical Properties. Compound properties were computationally generated using RDKit (version 2024.03.5). SMILES strings of compounds were processed to calculate a diverse set of molecular descriptors. For each compound, descriptors such as total atom count, exact molecular

weight, molecular formula, clogP, topological polar surface area (TPSA), number of rotatable bonds, and Lipinski's hydrogen bond donors and acceptors were computed. Additional properties included the number of aromatic rings, fraction of sp³ carbons, QED drug-likeness score, formal charge, minimal ring count, and the Murcko scaffold framework. Input data sets, comprising positive and negative classes of compounds, were merged and preprocessed to ensure valid SMILES representations. The calculated descriptors were stored in a consolidated pandas dataframe for further analysis. These computations provided a detailed molecular profile to facilitate the assessment of compound characteristics

Computation of "Drug-Like" Property using Computed Chemical Properties and Lipinski's Rules. To determine whether compounds were classified as "Drug-Like", molecular descriptors were computed for each compound and compared against class-specific representative compounds. The representative compounds for each class had their descriptors computed using advanced molecular properties such as ClogP, Topological Polar Surface Area (TPSA), Rotatable Bond Count, H-Bond Acceptors and Donors (Lipinski), and Fraction CSP3. For each class, the average values of these descriptors were calculated.

Subsequently, descriptors of the compounds within each class were compared to the class-specific average descriptors. Compounds were labeled as "Drug-Like" if at least four out of six descriptor values fell within 10% of the respective average descriptor values for their class. The percentage of matched descriptors was also calculated for each compound. This analysis allowed for the classification of compounds based on their resemblance to the pharmacophore properties of known bioactive molecules within their respective classes.

Evaluation of the Chemical Diversity within Each Training Class Using Tanimoto Similarity and Bemis-Murcko Frameworks. To evaluate the diversity within each drug class, two complementary approaches were employed: pairwise Tanimoto similarity and the enumeration of unique Bemis-Murcko frameworks. Molecular fingerprints for each compound were generated using the Python package RDKit, specifically its Morgan fingerprinting algorithm (radius 2, length 512), and Tanimoto similarity was calculated for all pairwise combinations within each class. The average Tanimoto similarity served as a measure of structural similarity, with lower averages indicating greater diversity. Additionally, Bemis-Murcko scaffolds were extracted for each compound using RDKit's MurckoScaffold.GetScaffoldForMol method, and the number of unique scaffolds within each drug class was quantified to assess scaffold diversity. Together, these analyses provided a comprehensive evaluation of the structural heterogeneity within each pharmacophore class, facilitating comparisons of diversity across drug classes.

■ ASSOCIATED CONTENT

Data Availability Statement

All of the data, code, and machine learning models to replicate the experiments done in this paper are available on GitHub at: <https://github.com/nathanbrittin/Natural-Product-Bioactivity-Classification>.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jnatprod.4c01123>.

The distribution of classes in the training data set used for the multiclassification model (Table S1); The distribution and number of GNPS spectra per class used for model evaluation (Table S2); The initial list of polyene macrolide antifungals selected for binary classifier training (Table S3); Taxonomic information on bacterial strains analyzed for polyene macrolide presence (Table S4); Identification counts for polyene identification of GNPS spectra (Table S5); Classification equivalents between pharmacophore classifications and ClassyFire or NP classification systems (Table S8);

Demonstration that utilizing mass spectra directly as machine learning input does not generalize well to experimental data (Table S9); A demonstration that changing the intensity of in-silico MSMS peaks does not impact molecular fingerprints derived from SIRIUS 5 (Figure S1); UV-vis confirmation of polyene UV patterns within bacterial extract LC-MS/MS data (Figure S2); Details on the bioactivity evaluation of fractionated bacterial library plates against *C. albicans* (Figure S3); Confusion matrices for ML model performance on 20% held-back training data and GNPS spectra are provided (Figure S4 and S5); A comparison of in-silico and GNPS spectra for polyene representative compounds using mirror matching and cosine similarity of spectra and molecular fingerprints (Figure S6); An evaluation of the diversity within all drug classes using Bemis-Murcko frameworks and average Tanimoto similarity (Figure S7) (PDF)

The list of GNPS spectrum IDs used for model performance evaluation in the form of a Microsoft word document (Table S6) (PDF)

Structure strings used to generate training and testing data sets in the form of a Microsoft excel file (Table S7) (XLSX)

Computed chemical properties for all training compounds in a Microsoft excel workbook (Table S10) (XLSX)

Computed chemical properties for representative structures of each class in a Microsoft excel notebook (Table S11) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Tim S. Bugni – Pharmaceutical Sciences Division, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States; Small Molecule Screening Facility, UW Carbone Cancer Center, Madison, Wisconsin 53792, United States; Lachman Institute for Pharmaceutical Development, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States; orcid.org/0000-0002-4502-3084; Email: tim.bugni@wisc.edu

Authors

Nathaniel J. Brittin – Pharmaceutical Sciences Division, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States; orcid.org/0009-0004-4309-7735
Josephine M. Anderson – Pharmaceutical Sciences Division, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States
Doug R. Braun – Pharmaceutical Sciences Division, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States
Scott R. Rajski – Pharmaceutical Sciences Division, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States
Cameron R. Currie – Department of Biochemistry and Biomedical Sciences, M.G. DeGroot Institute for Infectious Disease Research, David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario L8S 4L8, Canada; Department of Bacteriology, University of Wisconsin–Madison, Madison, Wisconsin 53705, United States

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jnatprod.4c01123>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant U19AI142720 and by generous contributions provided by the Marvel Family Fellowship as well as the Leon Lachman Fellowship; both administered by the Division of Pharmaceutical Sciences at UW-Madison. We also acknowledge funding from the UW-Madison Fall Competition. Finally, we thank the Analytical Instrumentation Center at the University of Wisconsin-Madison School of Pharmacy for the facilities to acquire MS data. The TOC graphic was created in BioRender. Brittin, N. (2025) <https://BioRender.com/g77x247>.

REFERENCES

- (1) *Antibiotic Resistance Threats in the United States, 2019*; CDC: Atlanta, GA, 2019. DOI: 10.15620/cdc:82532.
- (2) *COVID-19: U.S. Impact on Antimicrobial Resistance, Special Report 2022*; CDC: Atlanta, GA, 2022. DOI: 10.15620/cdc:117915.
- (3) *NIAID's Antibiotic Resistance Research Framework: Current Status and Future Directions 2019*; NIAID, 2019.
- (4) Genilloud, O. *Nat. Prod Rep* **2017**, *34* (10), 1203–1232.
- (5) De Simeis, D.; Serra, S. *Antibiotics* **2021**, *10* (5), 483.
- (6) Newman, D. J.; Cragg, G. M. *J. Nat. Prod* **2020**, *83* (3), 770–803.
- (7) Newman, D. J.; Cragg, G. M. *J. Nat. Prod* **2016**, *79* (3), 629–661.
- (8) Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Supuran, C. T.; et al. *Nat. Rev. Drug Discov.* **2021**, *20* (3), 200–216.
- (9) Gaudêncio, S. P.; Bayram, E.; LukićBilela, L.; Cueto, M.; Díaz-Marrero, A. R.; Haznedaroglu, B. Z.; Jimenez, C.; Mandalakis, M.; Pereira, F.; Reyes, F.; Tasdemir, D. *Mar Drugs* **2023**, *21* (5), 308.
- (10) Lima, N. M.; dos Santos, G. F.; da Silva Lima, G.; Vaz, B. G. *Microb. Nat. Prod. Chem.* **2023**, 101–122.
- (11) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kaponov, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34* (8), 828–837.
- (12) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *Journal of Mass Spectrometry* **2010**, *45* (7), 703–714.
- (13) Montenegro-Burke, J. R.; Guijas, C.; Siuzdak, G. *Methods Mol. Biol.* **2020**, *2104*, 149–163.
- (14) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. *Nat. Methods* **2019**, *16* (4), 299–302.
- (15) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Commun.* **2018**, *9* (1), 4035.
- (16) van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (48), 13738–13743.
- (17) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. *Nat. Methods* **2022**, *19* (7), 865–870.
- (18) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. *Nucleic Acids Res.* **2023**, *51* (D1), D1373–D1380.
- (19) Pence, H. E.; Williams, A. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124.
- (20) Mallowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Benidddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrzil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. *Nat. Rev. Drug Discov* **2023**, *22* (11), 895–916.
- (21) Zhang, R.; Li, X.; Zhang, X.; Qin, H.; Xiao, W. *Nat. Prod Rep* **2021**, *38* (2), 346–361.
- (22) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. *J. Nat. Prod.* **2021**, *84* (11), 2795–2807.
- (23) Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S. *Nat. Biotechnol.* **2021**, *39* (4), 462–471.
- (24) Dara, S.; Dhamecherla, S.; Jadvav, S. S.; Babu, C. H. M.; Ahsan, M. J. *Artif Intell Rev.* **2022**, *55* (3), 1947–1999.
- (25) Obaido, G.; Mienye, I. D.; Egbelowo, O. F.; Emmanuel, I. D.; Ogunleye, A.; Ogbuokiri, B.; Mienye, P.; Aruleba, K. *Machine Learning with Applications* **2024**, *17*, 100576.
- (26) Sarker, I. H. *SN Comput. Sci.* **2021**, *2* (3), 160.
- (27) Freund, Y.; Schapire, R. E. *Mach Learn* **1999**, *37* (3), 277–296.
- (28) Sorokina, M.; Steinbeck, C. *J. Cheminform* **2020**, *12* (1), 20.
- (29) Thermo Fisher Scientific. *Mass Frontier Spectral Interpretation Software*. <https://www.thermofisher.com/us/en/home/industry/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/mass-frontier-spectral-interpretation-software.html>.
- (30) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. *Anal. Chem.* **2021**, *93* (34), 11692–11700.
- (31) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (41), 12580–12585.
- (32) Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S. *Nat. Mach Intell* **2020**, *2* (10), 629–641.
- (33) Capecchi, A.; Probst, D.; Reymond, J.-L. *J. Cheminform* **2020**, *12* (1), 43.

- (34) Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; van der Hooft, J. J. *PLoS Comput. Biol.* **2021**, *17*, e1008724.
- (35) Huber, F.; van der Burg, S.; van der Hooft, J. J. J.; Ridder, L. *J. Cheminform* **2021**, *13* (1), 84.
- (36) Baygi, S. F.; Barupal, D. K. *J. Cheminform* **2024**, *16* (1), 8.
- (37) Goldman, S.; Wohlwend, J.; Stražar, M.; Haroush, G.; Xavier, R. J.; Coley, C. W. *Nat. Mach. Intell.* **2023**, *5* (9), 965–979.
- (38) da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (41), 12549–12550.
- (39) Van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9* (85), 2579–2605.
- (40) Hamilton-Miller, J. M. *Bacteriol. Rev.* **1973**, *37* (2), 166–196.
- (41) Eble, T. E.; Garrett, E. R. *J. Am. Pharm. Assoc.* **1954**, *43* (9), 536–538.
- (42) Garrett, E. R. *J. Am. Pharm. Assoc.* **1954**, *43* (9), 539–543.
- (43) Guruceaga, X.; Perez-Cuesta, U.; Abad-Diaz de Cerio, A.; Gonzalez, O.; Alonso, R. M.; Hernando, F. L.; Ramirez-Garcia, A.; Rementeria, A. *Toxins (Basel)* **2020**, *12* (1), 7.
- (44) DEKKER, J.; ARK, P. A. *Antibiot. Chemother.* **1959**, *9* (6), 327–332.
- (45) Kato, N.; Takahashi, S.; Nogawa, T.; Saito, T.; Osada, H. *Curr. Opin. Chem. Biol.* **2012**, *16* (1–2), 101–108.
- (46) Nelson, J.; Simpkins, S. W.; Safizadeh, H.; Li, S. C.; Piotrowski, J. S.; Hirano, H.; Yashiroda, Y.; Osada, H.; Yoshida, M.; Boone, C.; Myers, C. L. *Bioinformatics* **2018**, *34* (7), 1251–1252.
- (47) Quintana, R. P.; Lasslo, A.; Boggs, P. P. *J. Colloid Interface Sci.* **1968**, *26* (2), 166–174.
- (48) Cacho, R. A.; Chooi, Y.-H.; Zhou, H.; Tang, Y. *ACS Chem. Biol.* **2013**, *8* (10), 2322–2330.
- (49) Chicco, D.; Jurman, G. *BioData Min* **2023**, *16* (1), 4.