Investigating Bioactive Molecule Function from Novel Machine Learning Guided Discovery

Rana A. Haidous[1], Adrian Russ[2], Allison S. Walker[2]

1. *Hillsboro High School, Nashville, TN, United States 37215.*
2. *Department of Chemistry, Vanderbilt University, Nashville, TN, United States 37235.*

KEYWORDS. Machine, Algorithm, Bioactivity.

BRIEF. Utilizing machine-learning guided algorithm to discover bioactive molecule functions.

ABSTRACT. Antimicrobial resistance is a recurring issue in identifying drug compounds that is derived from over prescribed antibiotic prescriptions. Developing new drugs is costly and time-consuming which is why machine learning is a significant asset in discovering viable drugs that are safe and effective. This creates large demand compounds and relies greatly on identifying unknown bioactivity within biosynthetic gene clusters. This study aims to use a machine learning algorithm to create predictions on various bacterial strains. The bacterial strains used for this experiment include *P. rubra*, *R. rhizogenes*, and *H. cretacea*. Over the course of six weeks, metabolite extraction workflows, including bioactivity assays on each strain, were carried out to test the bioactive compounds. Gel analyses of polymerase chain reaction products were performed to amplify DNA from the bacterium's genome to determine if it is the correct uncontaminated species. Although the algorithm predicted high bioactivity for these strains, the assays resulted in low inhibition in the conditions tested. Overall, low bioactivity levels were observed, indicating there are few compounds to discover or the conditions were not optimal for high inhibition results. We interpret these results to indicate that higher concentrations are needed to obtain bioactivity.

INTRODUCTION.

According to the Center for Disease Control, healthcare providers attributed approximately 236.6 million antibiotic prescriptions to Americans in 2022, with at least 28% of those prescribed medications deemed unnecessary [1]. Despite the publics sincere need for prescription medicine, overuse of antibiotic prescriptions is an important issue that can lead to antimicrobial resistance. Antimicrobial resistance (AMR) occurs when bacteria, viruses, and fungi no longer respond to antimicrobial medicines. It is estimated that by 2050, over 39 million deaths will be caused by antibiotic-resistant infections [2].

Research and discovery costs range from $314 million to $2.8 billion and take 10 years to develop, often more per antibiotic [2]. Given the high cost and time to develop new drugs, machine learning-guided discovery is a significant asset in discovering viable drug prospects safely and efficiently. Machine learning-guided discovery algorithms predict the number of biosynthetic gene clusters (BGCs) per genome [3]. Access to these BGC sequences aids in predicting novel activities from natural products for drug discovery [3].

Over 50% of FDA-approved drugs are either natural products or derived from them, such as secondary metabolites [4]. Many antibiotics used since 1943 are produced by the bacterium *Streptomyces*. *Streptomyces* are not a single bacterium but a group of related bacteria which produces a variety of secondary metabolites used in antibiotics. However, since *Streptomyces* are utilized as a source in many antibiotics, discovering new natural product activities becomes increasingly challenging as most have already been identified. Combining the power of machine learning and studying strains that are less researched and studied compared to *Streptomyces*, could help lead to the discovery of new natural product activities. Therefore, contributing and utilizing machine learning data assists in the discovery of new natural product activities, ultimately leading to the development of new drugs, antibiotics, and treatments.

Streptomyces in Drug Discovery. Streptomyces is identified as a group of related bacteria with a large production of secondary metabolites, which are commonly used in human drugs [5]. Almost all FDA-approved drugs originate from the Streptomyces strain due to its diverse array of biosynthetic gene clusters (BGCs) [5]. A microbiologist, Selman Waksman, discovered Streptomycin in 1943 and is considered the first effective antibiotic treatment against tuberculosis [6]. Although *Streptomyces* has been extremely useful as it produces a high number of compounds, the drug development success rate has continued to fall, reaching an all-time low 6.3% composite success rate in 2022 [7]. This occurs because the same novel compounds have been rediscovered multiple times in similar strains of bacteria [5]. Since *Streptomyces* is one of the bacterial strains that has been used for such a long time, most of its bioactive compounds, such as lanthipeptides, terpenes, and type 1 polyketide synthases, have already been identified [5]. These compounds exhibit large antimicrobial and antifungal activities, exacerbating their importance in the world of new antibiotic drug discovery [8].

Additionally, the constant rediscovery of the same bioactive compounds creates a large demand for undiscovered compounds, and the need for new treatments will continue to grow [9]. The immense burden of induced resistance exacerbates the need for new antibiotics and treatments. Although some bacterial strains can foster many biosynthetic gene clusters (BGCs), like *Streptomycin* with 35, not all these products can be processed in laboratories as some may be silent gene clusters where they may require specific environmental cues such as stress conditions to "activate" [10]. The two main types of metabolites include primary and secondary. Primary metabolites are essential for an organism's survival, while secondary metabolites, which serve as enhancers such as defense mechanisms, are the ones harvested since the organism can live without them.

Antimicrobial Resistance Challenges. Along with rediscovery, antimicrobial resistance (AMR) is one of the two greatest issues in identifying novel drug compounds [2]. If the human body is treated with the same drug or antibiotic over time, germs develop the ability to defeat the drugs designed to kill them. To develop new antibiotics to combat AMR novel compounds need to be identified to create a brand-new drug. However, the process of drug discovery transcends beyond identifying novel compounds. As mentioned before, research and discovery costs range immensely, taking 10 years to develop, often more [2]. After this process, drugs still need to go through clinical trials where they are tested through three main phases and can vary in time with an average of 10 years for most drugs [10].

The Role of Machine Learning in Drug Discovery. Discovering new natural product activities is extremely time-consuming, especially as the chance of rediscovering biosynthetic gene clusters increases after a new one is introduced. To combat this issue, machine learning models can predict bioactive molecules, reducing the amount of time spent repeating the same cycle [2]. Biosynthetic gene clusters (BCGs) are a

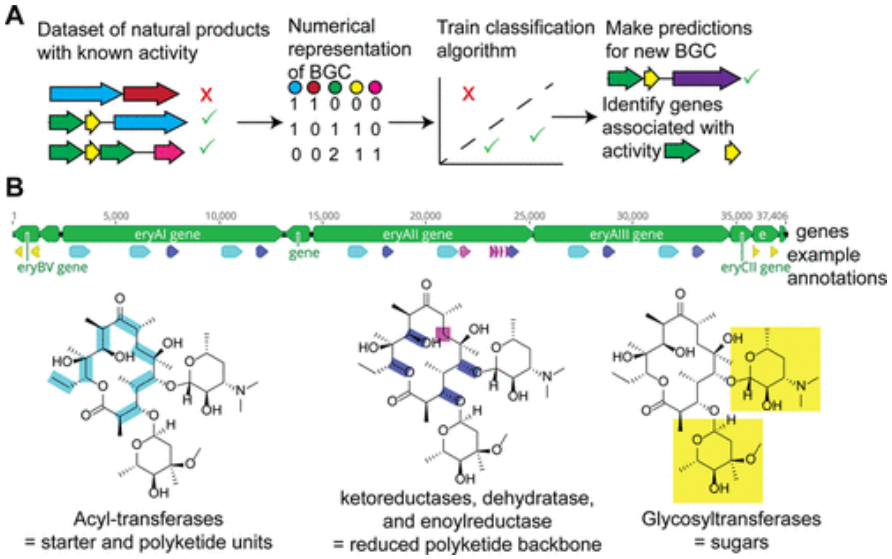

Figure 1. Machine learning method which includes the dataset for BCG's (A) and natural product activities for the gene clusters (B) (Figure by A. Walker).

group of two or more genes located closely together to encode a secondary metabolite [2]. To predict this new activity, a set of known BCGs and the natural products they produce are represented in a sequence [2] as seen in Figure 1. While not all drugs from machine learning are synthesized, the predictions are still important because not all drugs are derived from natural products. Some drugs are synthesized, which also addresses the unmet need for predictions. Therefore, Machine learning-guided discovery of bioactive compounds from carefully screened and purified bacterial strains will lead to a high potential of identified bioactive compounds. Through this study, the identification of these compounds will commence through bioactivity assays and metabolite isolation.

MATERIALS AND METHODS.

Bi-Phasic Extraction of Bacterial Growth Liquid-Liquid Extraction. A liquid-liquid extraction was performed to isolate compounds in the aqueous and organic phases to isolate metabolites for bioactivity assays. To prepare a sample for extraction, the supernatant was added to a beaker with as the aqueous layer. After this, the organic solvent, ethyl acetate, was added to the beaker. When these two phases were mixed, there were hydrophobic and hydrophilic compounds, making the hydrophobic compounds ultimately go towards the organic liquid phase. The beaker was then covered with aluminum foil, to avoid solvent evaporation. To transfer compounds that reside in both phases, such as hydrophobic compounds, into the organic phase from the aqueous phase, the beaker was sonicated for 45 minutes with periodic stirring of the beaker every 10 minutes. Therefore, compounds from both phases would be thoroughly transferred. After the beaker had been sonicated, the phases went through separation again so compounds would be extracted. To separate the phases, the mixture was transferred to a separating funnel, where the lower density liquid and the high-density liquid reside on the bottom of the funnel, separating as they did before the compounds were mixed. The purpose of separating both layers again is to have distinct compounds from each phase. The organic layer and aqueous layers were collected in round separate flasks; after this, a layer of bubbles was formed at the solvent interface due to mixed hydrophobic and hydrophilic properties and was treated as a third layer, collected separately in its own flask. To fully extract the compounds, each round flask containing each phase was put into a rotary evaporator to concentrate metabolites through evaporating any ethyl acetate so only hydrophobic compounds remain. Once the organic layer had dried, acetonitrile was added to dissolve the sample, making it more suitable to use for an assay. To fully dry the mixture, the remaining liquid was transferred to a 50 µL falcon tube to dry in a speed-vac. The metabolites were stored in a -20 degrees Celsius freezer after done drying in the speedvac.

Agar Plate Bio-Activity Assay. We utilized bioactivity assays to evaluate certain antimicrobial properties of extracted metabolites derived from natural products. To ensure full integrity of assay results, every step was done by an open flame to avoid contamination. In this assay, A *Bacillus subtilis* was selected from a fully grown liquid growth in a culture tube since it demonstrates high growth characteristics throughout determining microbial inhibition. *Bacillus subtilis* was the one bacterium used in this process because it had very sensitive characteristics against the antimicrobial compounds and was also one of the bacterium that was easy to culture [10]. The use of this bacteria ultimately made it easier to get more consistent results as it can identify how effective specific metabolites are. We plated everything together and labeled each section of metabolites to spot specific placement and for future reference. To have sufficient bacterial coverage, 100 µL of *Bacillus subtilis* was pipetted onto the petri dish. A sterile spreader was then utilized to distribute the bacteria evenly across the agar plate so that there is a constant, even layer for the different metabolites from each layer to grow on. 5 µL of each extracted metabolites suspended in solution were pipetted onto each corresponding label. Additionally, 5 µL of the solvent was utilized for inhibition as the control, and to ultimately distinguish effects with any solvent influences. After pipetting the metabolites, the agar plate was then closed and stored face down in a solid incubator to eliminate any condensation that may reach the surface of the agar and disrupt growth results. The plate was monitored daily for colony growth on areas where metabolites were pipetted.

Gel Electrophoresis Analysis of Polymerase Chain Reaction (PCR) Product. To identify DNA in bacteria, PCR was conducted to amplify the 16S rRNA gene which was used for sequences that tend to be species-specific oriented. PCR is when the samples go through cycles of different temperature in order to separate DNA strands, anneal primers, then build brand new DNA strands. With these steps, numerous copies of DNA were made of the target gene which assists in the gene amplification.

Pelleted bacterial cells were resuspended in the PrepMan solution and heated at 100 degrees Celsius to release the DNA. After this, the DNA is then combined with primers that are specific to the 16S rRNA gene and Taq DNA polymerase which is an enzyme that synthesizes brand new DNA strands. Over the time span of three hours, the PCR reaction amplified the DNA with the touchdown program in a thermocycler.

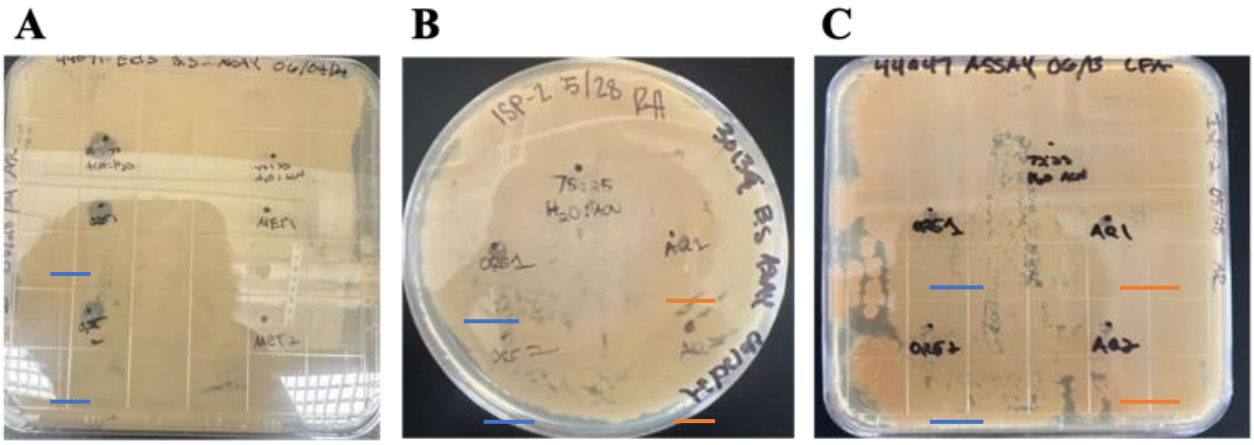

Figure 2. Solid agar plates with two solutions and one control. "Met 1" & "Met 2" represent methanol and is used in place of the aqueous phase **(A)**. In addition to 72:25 H2O:ACN, 25:75 CAN:H2O is used for controls **(A)**. "Org 1" and "Org 2" (blue) stand for the organic phase, "Aq1" & "Aq2" (orange) represent the aqueous phase **(B, C)**. 72:25 H20:ACN is the control used **(B, C)**.

We verified amplification through separated PCR products by gel electrophesis. This was done by preparing an agarose gel including SYBR safe stain which helped with visualization under UV light. To sort the DNA fragments and see which ones were bigger than the others, the samples were run with a DNA ladder alongside them at 100V. To indicate the successful amplification of the gene that is targeted, bands were revealed under a UV light and demonstrate bands that were around 1500.

RESULTS.

Agar Plate Bio-activity Assay Growths. The metabolites and compounds of each bacterial strain were tested against *Bacillus subtilis*. To determine the bioactivity of the metabolites produced by these bacteria, a bioactivity assay is a technique used to evaluate a specific strain's growth and to learn more about the strain. The key feature of a bioactivity assay is the inhibition zone, a clear circular area that surrounds the antimicrobial agent. In figure 2A we see that there is a small zone of inhibition which measures the susceptibility of the bacteria around the spot the compounds were pipetted in. We verified this the most in figure 2B, where inhibition zones are prominent in the Org 1 & Org 2 spots as well as its control of 25:75 ACN:H2O. In figure 2C, the only inhibition zone that is seen in this plate is around Org 1.

Amplifying the DNA of Various Bacterial Strains. In Figure 3, an image of a gel electrophoresis demonstrates the DNA amplification for bacterial strains using a PCR product made up of *Polymorphospora rubra, Herbidospora sakaeratensis, Herbidospora cretacea, Herbidospora galbida, Longispora albida, Longispora fulva and Longispora urticae*. We use a UV light, to show and represent DNA bands visually. Figure 3 demonstrates a DNA ladder that represents each DNA fragment present and is labeled with kilobases (kb) which essentially references the size of each DNA fragment. Gel shows bands at 1.5 kb indicating that the PCR was successful for each bacterial strain. Lack of smearing in the bands indicate purity of the DNA amplification. With this, these results indicate novel bacteria strains that we can identify in future studies.

DISCUSSION.

The strains *Polymorphospora rubra*, *Herbidispora cretacea*, and *Rhizobium rhizogenes* were selected for this study. The strains were grown and each of their metabolites were extracted and tested for bioactivity, revealing differing levels of inhibition, growth, and bioactivity. The machine learning algorithm demonstrated that strains with more predictions for biosynthetic gene clusters per genome were more favored. The machine learning method not only predicts the number of clusters but predicts the activity of the strain given their BGC profile. The goal was to select strains based on their predicted activity, and incubation length due to the time constraints of the duration of this project. As a result, finding a balance of selecting strains that grew relatively fast, but also had a promising prediction according to the algorithm was a vital aspect. However, a strain may have many predicted BCGs, some of which may have already been discovered, increasing the risk of rediscovery. It is more important to spend time researching strains that do not have many clusters that are identified within them since the stake of rediscovery tends to get higher as the number of identified BCGs increases.

All agar bioactivity plates showed partial or low amounts of inhibition, which suggests a need to explore the effectiveness of the substances used in this study further. If the strains did not exhibit increased inhibition across the plate, different conditions might be required to achieve maximum potential growth to achieve optimal metabolite expression from BGCs. Under the tested conditions, these strains may not effectively demonstrate their antibiotic activity, but other conditions could reveal greater inhibition and growth. In the first assay, *Polymorphospora rubra* showed very little inhibition, which may indicate that the strain has low bioactivity and requires different conditions, such as alternative substances or extraction methods. In the second assay, *Rhizobium rhizogenes*, partial inhibition zones were demonstrated, which means that the bacteria are somewhat effective towards the antimicrobial agent used. The third assay, with *Herbidispora cretacea*, showed low growth and inhibition, similar to the second assay. Overall, while growth was minimal in this study, we conclude that is a promising endeavor and may benefit from testing more conditions.

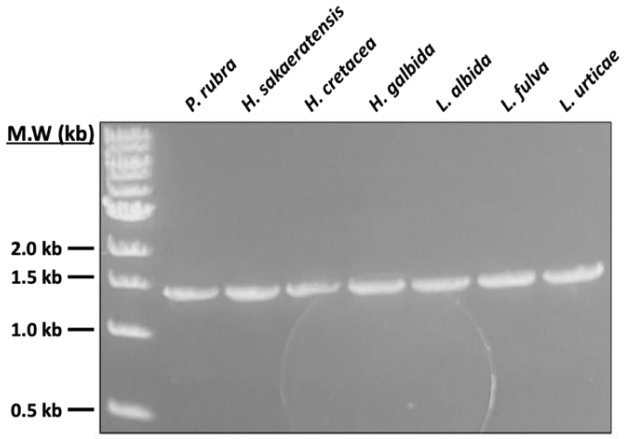

Figure 3. Gel Electrophesis image of various bacteria's and their molecular weights.

The hypothesis for this study verified that machine learning-guided discovery of bioactive compounds from carefully screened and purified bacterial strains will lead to the identification of novel bioactive compounds. The algorithm would ultimately determine whether or not a strain would have large amounts of predicted activities, and from the strains chosen they had a predicted activity of around 30 BGCs cumulatively. However, the results in this study did not align with this hypothesis as limited amounts of inhibition were present throughout the three bacterial strains *Polymorphospora rubra, Herbidispora cretacea, and Rhizobium rhizogenes.* The hypothesis was not accurate as the machine learning algorithm predicted something different from the results, and this could be due to the certain methods or testing conditions used which may not be completely optimal for the metabolites to have high activity in.

Implications and applications. An imperative aspect in drug discovery lies unidentified compounds in unresearched bacterial strains. Although the results of all three bacterial strains demonstrated low inhibition levels, these new insights still count as new knowledge since it is learned to test these antimicrobial different conditions. These findings are applicable as they influence future studies that could be performed on these same strands and can inform those studies in altering certain cultivation conditions. For example, through lengthening periods of incubation this could overall enhance how novel metabolites are expressed through the bioactivity assays against certain microbial properties. Ultimately, resistant pathogens are slowly growing and applying this research to the future is imperative as the issue of AMR only increases with time. Therefore, the optimization of increasing incubation times, or continuous streaking on plates would help enhance the expressed activity metabolites have. This study could also be an approach to begin conducting bioactivity assays differently. We identified the tested conditions were unsuitable with these strains. Experimentations to know which assay method is best for all strains can be implemented so future research can further excel. We inferred that the improved methodology helps inform future studies with the goal of drug discovery in mind, shortening the time of compound discovery.

FUTURE DIRECTIONS.

To improve the discovery of bioactive compounds in all three strains, several future directions can be implemented to achieve higher inhibition and growth results. These include scaling up the growths for *Rhizobium rhizogenes* and *Herbidispora cretacea* by increasing their incubation time in order to develop their metabolite expression to obtain a greater number of compounds during growth. Doing this would address the number of limited compounds that were observed during previous assays. Previous microbial studies showed microbial fermentation being scaled up, resulting in an increased yield of compounds that demonstrate bioactivity. With this, longer periods of incubation would assist in expressing a higher yield of bioactive compounds in this experiment which addresses low inhibition. As low inhibition arose throughout all three bioactivity assays, another future direction would be streaking on plates as a useful tactic for better colony isolation and successful extraction. When each colony is more isolated and separate, this links with growth patterns that tend to be more defined and easier to analyze visually. The technique allows the individual colonies to grow without any type of interference with other colonies surrounding with close proximity. This ultimately minimizes the "competition" for every colony and helps it flourish on its own – also leading to more selective colony development. Another future aim would be reproducing these growths in higher volumes would allow the strains to react better to a higher dosage, leading to more refined effects. Scaling up these cultures would assist in higher inhibition as it would during increased incubation periods, accuracy and responses would be more consistent as well since large fermentations have been proven to have more accurate results within bioactivity compounds.

CONCLUSION.

The objective of this experiment was to extract novel compounds from selected strains of bacteria using machine learning algorithms to aid in new drug discovery. The experiment combined bioactivity assays with machine learning-guided discovery. Although the algorithm predicted high bioactivity for these strains, the assays resulted in low inhibition and bioactivity due to the conditions tested. Overall, low colony levels were observed, indicating either that the strains inherently have low bioactivity with few novel compounds to discover or that the conditions were not optimal for high inhibition results, which would lead to increased bioactivity detection. *Polymorphospora rubra* showed visual microbial growth, while *Rhizobium rhizogenes* and *Herbidispora cretacea* demonstrated low growth and inhibition. These results highlight the need for higher concentration levels or enhanced sterilization techniques to obtain more credible and significant results.

ACKNOWLEDGMENTS.

I thank the Research Experience for High School Students program and thank Dr. Walker, Adrian Russ, Dr. Swartz and Dr. Means for their guidance and support.

REFERENCES.

[1] Centers for Disease Control and Prevention. Outpatient antibiotic prescribing in the United States. *CDC Gov* (2024).
[2] R. C. Mohs, N. H. Greig, Drug discovery and development: Role of basic biological research. *Clin Interv* **3**, 651–657 (2017).
[3] O. Riedling, A. S. Walker, A. Rokas, Predicting fungal secondary metabolite activity from biosynthetic gene cluster data using machine learning. *Microbiol Spectr* **12**, e03400-23 (2024).
[4] J. Newman, G. M. Cragg, Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
[5] K. C. Belknap, C. J. Park, B. M. Barth, C. P. Andam, Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Sci. Rep.* **10**, 2003 (2020).
[6] H. B. Woodruff, Selman A. Waksman, Winner of the 1952 Nobel Prize for Physiology or Medicine. *Appl Environ Microbiol* **80**, 2–8 (2014).
[7] D. Sun, W. Gao, H. Hu, S. Zhou, Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* **12**, 3049–3062 (2022).
[8] J. Knerr, W. A. Van Der Donk, Discovery, Biosynthesis, and Engineering of Lantipeptides. *Annu. Rev. Biochem.* **81**, 479–505 (2012).
[9] L.-F. Nothias, M. Nothias-Esposito, R. Da Silva, Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J. Nat. Prod.* **81**, 758–767 (2018).
[8] M. W. Mullowney, K. R. Duncan, S. S. Elsayed et al., Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discovery* **22**, 895–916 (2023).
[9] A. S. Walker, J. Clardy, A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* **61**, 2560–2571 (2021).
[10] F. Xu, Y. Wu, C. Zhang, et al., genetics-free method for high-throughput discovery of cryptic microbial metabolites. *Nat Chem Biol* **15**, 161–168 (2019).

Rana Haidous is a student at Hillsboro High School in Nashville, Tennessee. She participated in a research internship through the Research Experience for High School Students.