DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

Final report on DE-SC0021340: Discovery of Signaling Small Molecules (e.g. quorum sensing molecules) from the Microbiome.

Microbial communities are shaped through the interactions between their microbial members and the environment (microbe-microbe and host-microbe interactions). Signal transduction pathways in the microbiome are often modulated through the small molecule products of microbial biosynthetic gene clusters (BGCs). Advances in 16S rRNA profiling and shotgun metagenomics have revolutionized our understanding about the microbial composition of various communities and their BGCs. Environmental metagenomes contain thousands of BGCs with uncharacterized small molecule products that potentially play roles in signal transduction. The overarching aim of this proposal was to develop computational techniques for discovering these small molecules and characterizing their bioactivity.

We proposed to achieve this goal through three objectives. The first objective revolves around predicting the structure of microbial natural products from their biosynthetic gene cluster. The second objective is focused on an efficient strategy searching these molecular products against tandem mass spectral data, and the third objective is focused on isolation of these predicted bioactive molecules, and validation of the approach.

Our team successfully accomplished these objectives. Here I list the publications from our team focusing on each objective, and then I detail the progress made on each objective.

Publications on objective 1:

Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. Bahar Behsaz, Edna Bode, Alexey Gurevich, Yan-Ni Shi, Florian Grundmann, Deepa Acharya, Andrés Mauricio Caraballo-Rodríguez, Amina Bouslimani, Morgan Panitchpakdi, Annabell Linck, Changhui Guan, Julia Oh, Pieter C. Dorrestein, Helge B. Bode, Pavel A. Pevzner & Hosein Mohimani, **Nature Communications,** 12, 3225 (2021)

HypoRiPPAtlas as an Atlas of hypothetical natural products for mass spectrometry database search.
Yi-Yuan Lee, Mustafa Guler, Desnor N. Chigumba, Shen Wang, Neel Mittal, Cameron Miller, Benjamin Krummenacher, Haodong Liu, Liu Cao, Aditya Kannan, Keshav Narayan, Samuel T. Slocum, Bryan L. Roth, Alexey Gurevich, Bahar Behsaz, Roland D. Kersten & Hosein Mohimani, **Nature Communications,** 14, 4219 (2023).

AdenPredictor: accurate prediction of the adenylation domain specificity of nonribosomal peptide biosynthetic gene clusters in microbial genomes. Mihir Mongia, Romel Baral, Abhinav Adduri, Donghui Yan, Yudong Liu, Yuying Bian, Paul Kim, Bahar Behsaz, Hosein Mohimani, **Bioinformatics**, i40–i46 (2023).

Discovering type I cis-AT polyketides through computational mass spectrometry and genome mining with Seq2PKS, Donghui Yan, Muqing Zhou, Abhinav Adduri, Yihao Zhuang, Mustafa Guler, Sitong Liu, Hyonyoung Shin, Torin Kovach, Gloria Oh, Xiao Liu, Yuting Deng, Xiaofeng Wang, Liu Cao, David H. Sherman, Pamela J. Schultz, Roland D. Kersten, Jason A. Clement, Ashootosh Tripathi, Bahar Behsaz & Hosein Mohimani. **Nature Communications**, 15, 5356 (2024).

Publications on objective 2:

MolDiscovery: learning mass spectrometry fragmentation of small molecules, Liu Cao, Mustafa Guler, Azat Tagirdzhanov, Yi-Yuan Lee, Alexey Gurevich & Hosein Mohimani. **Nature Communications**, 12, 3718 (2021).

MS2Planner: improved fragmentation spectra coverage in untargeted mass spectrometry by iterative optimized data acquisition, Zeyuan Zuo, Liu Cao, Louis-Félix Nothia, Hosein Mohimani. **Bioinformatics**, 37, i231–i236 (2021)

Repository scale classification and decomposition of tandem mass spectral data, Mihir Mongia & Hosein Mohimani. **Scientific Reports**, 11, 8314 (2021).

Fast mass spectrometry search and clustering of untargeted metabolomics data. Mihir Mongia, Tyler M. Yasaka, Yudong Liu, Mustafa Guler, Liang Lu, Aditya Bhagwat, Bahar Behsaz, Mingxun Wang, Pieter C. Dorrestein & Hosein Mohimani, **Nature Biotechnology**, 42, 1672–1677 (2024).

Manuscripts on objective 3:

Pathogen-oriented platform for large-scale natural product discovery identifies novel antifungal targeting drug-resistant Candidiasis, Bahar Behsaz, Abhinav Adduri, Mustafa Guler, Osama G. Mohamed, Andrés Mauricio Caraballo-Rodríguez, Nirmal Chaudhary, Benjamin Krummenacher, Cameron Miller, Sitong Liu, Daniel Zamith-Miranda, Sneha P. Couvillion, Pamela J. Schultz, Kirk Broders, David H. Sherman, Ernesto S. Nakayasu, Joshua D. Nosanchuk, Pieter C. Dorrestein, Jason Clement, Ashootosh Tripathi, Hosein Mohimani. Under revision of Journal of American Chemical Society (2025).

For objective 1, we focused on three classes of natural products, nonribosomal peptides, ribosomally synthesized and post-translationally modified peptides, and polyketides. Nonribosomal peptides (NRPs) are a diverse class of natural products that include antibiotics, immunosuppressants, anticancer agents, toxins, siderophores, pigments, and cytostatics. The discovery of novel NRPs remains a laborious process because many NRPs consist of nonstandard amino acids that are assembled by nonribosomal peptide synthetases (NRPSs). Adenylation domains (A-domains) in NRPSs are responsible for selection and activation of monomers appearing in NRPs. During the past decade, several support vector machine-based algorithms have been developed for predicting the specificity of the monomers present in NRPs. These algorithms utilize physiochemical features of the amino acids present in the A-domains of NRPSs. In 2023, we developed AdenPredictor (Bioinformatics Journal), an extra trees model paired with one-hot encoding features outperforms accuracy of the existing approaches in prediction of the adenylation domain specificity of NRPs. Moreover, we showed that unsupervised clustering of 453 560 A-domains reveals many clusters that correspond to potentially novel amino acids. While it is challenging to predict the chemical structure of these amino acids, we developed novel techniques to predict their various properties, including polarity, hydrophobicity, charge, and presence of aromatic rings, carboxyl, and hydroxyl groups.

The predict molecular structure of NRPs from their BGCs, in 2021 we developed NRPminer (Nature Communications). Since the existing genome mining tools predict many putative NRPs synthesized by a given BGC, it remains unclear which of these putative NRPs are correct and how to identify post-assembly modifications of amino acids in these NRPs in a blind mode, without knowing which modifications exist in the sample. To address this challenge, we developed a modification-tolerant tool for NRP discovery from large (meta)genomic and mass spectrometry datasets. We showed that NRPminer is able to identify many NRPs from different environments, including four previously unreported NRP families from soil-associated microbes and NRPs from human microbiota. Furthermore, in this work we demonstrated the anti-parasitic activities and the structure of two of these NRP families using direct bioactivity screening and nuclear magnetic resonance spectrometry, illustrating the power of NRPminer for discovering bioactive NRPs.

The predict molecular structure of RiPPs from their BGCs, in 2023 we developed HypoRiPPAtlas (Nature Communications). HypoRiPPAtlas bridge the gap between large-scale genome mining and mass spectral datasets for natural product discovery. This Atlas is ready-to-use for in silico database search of tandem mass spectra. HypoRiPPAtlas is constructed by mining genomes using seq2ripp, a machine-learning tool for the prediction of ribosomally synthesized and post-translationally modified peptides (RiPPs). In HypoRiPPAtlas, we identify RiPPs in microbes and plants. This study paves the way for large-scale explorations of biosynthetic pathways and chemical structures of microbial and plant RiPP classes.

The predict molecular structure of Type 1 polyketide natural products from their BGCs, in 2024 we developed Seq2PKS (Nature Communications). Type 1 polyketides are a major class of natural products used as antiviral, antibiotic, antifungal, antiparasitic, immunosuppressive, and antitumor drugs. Analysis of public microbial genomes leads to the discovery of over sixty thousand type 1 polyketide gene clusters. However, the molecular products of only about a hundred of these clusters are characterized, leaving most metabolites unknown. Characterizing polyketides relies on bioactivity-guided purification, which is expensive and time-consuming. To address this, we developed Seq2PKS, a machine learning algorithm that predicts chemical structures derived from Type 1 polyketide synthases. Seq2PKS predicts numerous putative structures for each gene cluster to enhance accuracy. The correct structure is identified using a variable mass spectral database search. Benchmarks show that Seq2PKS outperforms existing methods. Applying Seq2PKS to Actinobacteria datasets, we discover biosynthetic gene clusters for monazomycin, oasomycin A, and 2-aminobenzamide-actiphenol.

To goal of objective 2 is to determine whether the predicted natural products from Aim 1 are expressed in mass spectrometry, and whether there are any unanticipated modifications in their structure. To address this challenge, in 2021 we developed MolDisocvery, a mass spectral database search tool for identification of small molecules. Identification of small molecules is a critical task in various areas of life science. Recent advances in mass spectrometry have enabled the collection of tandem mass spectra of small molecules from hundreds of thousands of environments. To identify which molecules are present in a sample, one can search mass spectra collected from the sample against millions of molecular structures in small molecule databases. The existing approaches are based on chemistry domain knowledge, and they fail to explain many of the peaks in mass spectra of small molecules. We developed molDiscovery, a mass spectral database search method that improves both efficiency and accuracy of small molecule identification by learning a probabilistic model to match small molecules with their mass spectra. A search of over 8 million spectra from the Global Natural Product Social molecular networking infrastructure shows that molDiscovery correctly identify six times more unique small molecules than previous methods.

Another challenge is objective 2 is that we usually need to search millions of predicted molecular structures against millions of mass spectra. This can be very time consuming using brute force search. To address this limitation, we developed MASST+ and Networking+, which can process datasets that are up to three orders of magnitude larger than those processed by state-of-the-art tools (Nature Biotechnology). MASST+ and Networking+ speed up the search using an indexing strategy.

To validate the pipeline (Objective 3), we collected mass spectral data on 1500 *Streptomyces* strains with genome sequence publicly available. Then we applied various computational methods we have developed for identification of novel natural products. This led to the identification of 100 novel natural products. To validate our method, we isolated one of the predicted NRPs, Edaphochelin. We demonstrated that Edaphochelin kills urgent-threat and multi-drug-resistant *Candida* pathogens by reducing the respiratory chain proteins. Structural elucidation of Edaphochelin was determined using nuclear magnetic resonance and mass spectrometry techniques. Bioactivity, safety, and efficacy were determined by in vitro assays and in vivo mouse models. The efficacy, safety, and potentially distinct mode of action from existing antifungal drugs render Edaphochelin as a promising drug candidate to combat the global threat of Candida pathogens.