Precision Discovery of Novel Inhibitors of Human Cancer Target HsMetAP*1* from Vast Unexplored Metagenomic Diversity

Oliver W. Liu[1]*, Scott Akers[1], Gabriella Alvarez[1], Stephanie Brown[1], Wenlong Cai[1], Zachary Charlop-Powers[1], Kevin Crispell[1], Ee-Been Goh[1], William W. Hwang[1], Tom H. Eyles[1], Sangita Ganesh[1], Peter Haverty[1], John L. Kulp Jr.[1], John L. Kulp III[2], Zachary Kurtz[1], Andrea Lubbe[1], Matthew Jamison[1], Aleksandr Milshteyn[1], Parisa Mokthari[1], Stephen G. Naylor[1], Samuel Oteng-Pabi[1], Ross Overacker[1], Andrew W. Robertson[1], Helen van Aggelen[1], Usha Viswanathan[3], Xiao Yang[1], Sam Yoder[1], Steven L. Colletti[1], and Devin R. Scannell[1]

[1] Zymergen, Inc., 5980 Horton Street, Ste. 105, Emeryville, CA 94608
[2] Conifer Point Pharmaceuticals, 3805 Old Easton Road, Doylestown, PA 18902
[3] Baruch S. Blumberg Institute, 3805 Old East Road, Doylestown, PA 18902
* Corresponding author: oliver@zymergen.com

ABSTRACT

Microbial natural products are specialized metabolites that have long been a rich source of human therapeutics. While the chemical diversity encoded in the genomes of microbes is believed to be large, the productivity of this modality has waned as traditional fermentation-based discovery methods have been plagued by high-rates of rediscovery, inefficient scaling, and incompatibility with target-based drug discovery. Here, we demonstrate a scalable discovery platform that couples dramatically improved assembly of deep-sequenced metagenomic samples with highly efficient, target-focused, *in silico* search strategies and synthetic biology to discover multiple novel inhibitors of human methionine aminopeptidase-1 (HsMetAP1), a validated oncology target. For one of these novel inhibitors, metapeptin B, we demonstrate sub-micromolar potency, strong selectivity for HsMetAP1 over HsMetAP2 and leverage natural congeners to rapidly elucidate key SAR elements. Our “next-gen” discovery platform overcomes many of the challenges constraining traditional methods, implies the existence of vast, untapped chemical diversity in nature, and demonstrates computationally-enabled precision discovery of modulators of human proteins of interest.

INTRODUCTION

Small molecule natural products (NPs) are specialized metabolites encoded in the genomes of bacteria, fungi, and plants. They have long been a rich source of human therapeutics. For example, over half of all small molecule drugs on the market are derived from NPs, including 65% of oncology drugs and 71% of anti-infectives (Newman and Cragg, 2020) (Stratton et al., 2015). Examples include Kyprolis (cancer), Rapamune (immune modulation), Zocor (cardiovascular) and Cubicin (infectious disease).

The outsized role NPs play in therapeutics can be attributed to the fact that these molecules have been evolutionarily selected to modulate critical cellular pathways and proteins (Chevrette et al., 2020). NPs are often exceptions to Lipinski’s Rule-of-5 indicating that natural selection has evolved bioactive molecules with drug-like properties that are otherwise difficult to conceive

(Doak et al., 2014). Additionally, the larger size and complexity of NPs – which generally contain more chiral centers and SP3 carbon atoms than synthetic small molecules – translate into rich, three-dimensional structures that occupy chemical space distinct from the simpler synthetic small molecules (Stone et al., 2022). Critically, this chemical space is otherwise difficult to explore, even with large-library technologies such as DELs, phage-display, or AI-guided drug design (Lenci et al., 2021) (Pitt and Nims, 2019) (Thomas et al., 2022).

Importantly, NPs beneficial chemistry does not come at the expense of the ability to cross cell membranes and enter cells (Salvador-Reyes and Luesch, 2015). Many of the most important undrugged cancer targets are intracellular and are thus effectively unreachable by monoclonal antibodies and other emerging modalities (Behan et al., 2019). Indeed, small molecules (including NPs) are largely unique in offering reliable delivery to diverse organs, oral administration, and competitive manufacturing (Chhabra, 2021). Innovation within the small molecule space is thus critical to address many of our most important drug targets.

Despite their track record of success, the pace of NP discovery has slowed dramatically. Over the last three decades, the majority of pharmaceutical companies have exited this modality, as the traditional fermentation-based approaches for NP discovery have been plagued by high rates of rediscovery of known compounds and inefficient scaling, resulting in lowered return on investment (Atanasov et al., 2021) (Li et al., 2019). A central limitation driving this trend is that only ~1% of microbes are readily cultured in the lab and, therefore, amenable to fermentation-based discovery (Rappé and Giovannoni, 2003) (Handelsman et al., 1998). Of these, only a small portion express molecules under any given fermentation condition, making robust screening protocols extremely challenging.

Metagenomics, which involves the capture of environmental DNA (eDNA) paired with heterologous expression systems, offers a potentially powerful alternative to traditional fermentation-based discovery by providing direct access to uncultured (and unstudied) diversity (Katz et al., 2016)) (Stevenson et al., 2019). In practice, however, efforts toward metagenomics-based NP discovery have met with relatively limited success. While congeners of already known NP scaffolds have been discovered from metagenomes (Stevenson et al., 2021) (Peek et al., 2018) (Owen et al., 2015) (Chang et al., 2013), de novo scaffolds with novel bioactivities of interest have remained elusive outside of a few examples (Wang et al., 2022). These efforts have been hindered by two main challenges: difficulty of obtaining good metagenomic sequence assemblies and a lack of computational approaches to efficiently identify desirable biosynthetic gene clusters (BGCs).

The immense size and complexity of soil metagenomic environments, which are estimated to contain 10^4-10^6 unique phylotypes per gram of soil, limit the efficacy of shotgun sequencing approaches. Short-read technologies, like Illumina, do not generate sequence assemblies that are large enough to be broadly useful for the discovery of natural-product encodingBGCs which can range in size from 10 to 100+kb (Xu et al., 2022). Long-read technologies, on the other hand, still lack the throughput to provide sufficient coverage of very complex samples and/or suffer from high error rates (Delahaye and Nicolas, 2021) (Tedersoo et al., 2021). Hybrid sequencing strategies have improved the quality of metagenomic assemblies from soil but are still insufficient for broad BGC discovery (Xu et al., 2022).

Even with higher-quality sequence information, one needs a strategy to identify the small number of BGCs encoding NPs with therapeutic value out of the multitude of BGCs that can be found in the metagenome. While BGCs and their constituent genes can be bioinformatically annotated, there has been limited success in developing computational approaches that can accurately predict the structure of the encoded small molecule NP. The ideal would be a bioinformatic strategy that can leverage DNA sequence alone to down-select and prioritize the BGCs with high-value therapeutic potential.

In this paper, we present an end-to-end metagenomic NP discovery platform that enables the targeted discovery of novel NPs that can modulate a specific human protein target of interest. We demonstrate the power of this approach by leveraging this platform for targeted discovery of novel NP inhibitors of human methionine aminopeptidase-1 (HsMetAP1), a key translational regulator with strong associations across a wide range of solid tumor cancers (Frottin et al., 2016) (Behan et al., 2019). We detail our efforts to sequence and catalog metagenomic soil diversity on a scale that enables unprecedented access to the massive biosynthetic potential encoded in the soil microbiome, including a metagenomics database that contains >1.4Tb of assembled sequences from contigs greater than 10kb in length and >6.8M predicted BGCs across six different soil samples. To identify BGCs in this vast collection with potential therapeutic value, we utilized a bioinformatic strategy leveraging the presence of self-resistance enzymes (SREs) to identify 35 BGCs in the database that are predicted to encode distinct, novel NP inhibitors of HsMetAP1, including two that were selected for further analysis. Downstream technologies for heterologous expression, untargeted discovery of novel metabolites, and computational approaches to assign observed bioactivity to specific metabolites enabled the production, identification, and isolation of encoded molecules that validate our functional predictions, resulting in the discovery of a novel cyclic depsipeptide inhibitor of HsMetAP1, which we call metapeptin B.

RESULTS

Sequencing of large-insert cosmid metagenomic libraries enables access to massive metagenomic diversity including millions of novel BGCs

Soils are among the richest ecological sources of microbial diversity (Curtis et al., 2002) (Rappé and Giovannoni, 2003). The repertoire of yet undiscovered microbial NPs that exist within this environment is vast (Gavriilidou et al., 2022). However, the immense size and complexity of the soil metagenomes have severely limited the efficacy of shotgun metagenomic sequencing strategies. The complexity makes it computationally challenging to assemble short-read data to sufficient lengths for BGC discovery (*e.g.*, >10kb), while the diversity and the lack of population uniformity make it difficult to obtain sufficient coverage using long-read technologies to capture sequence information from less abundant species (Tedersoo et al., 2021). To assess the state of the art in soil metagenome sequencing, we analyzed the top five most deeply sequenced soil metagenomes, as measured by assembled size, generated by the Joint Genome Institute (JGI) of the U.S. Department of Energy (https://img.jgi.doe.gov/cgi-bin/m/main.cgi). These datasets range from 73.0Gb to 22.5Gb of assembled sequence; however, on average, only 3.7% (1.2%-5.6% range) of the assembled sequence in these datasets are found on contigs >10kb, indicating that the vast majority of these data are not useful for BGC discovery (Figure 1a).

To enable metagenomic BGC discovery, we hypothesized that we could generate higher quality assemblies of soil metagenomes by first partitioning the initial sample into smaller, lower-complexity DNA sub-pools. The resulting reduction in size and complexity would make the sub-pools more amenable to shotgun sequencing and the higher-quality assemblies from each DNA sub-pool could then be joined to give a higher resolution view into the overall soil metagenome. To this purpose, we built large-insert (35-40kb) cosmid libraries from soil eDNA. The libraries were arrayed as sub-pools of 6,000-25,000 cosmids each, thus reducing the complexity of the sequence assembly challenge from a mixture of 10^4-10^6 microbial genomes in the initial eDNA sample to roughly 50-250 genome equivalents per sub-pool (Figure 1b). In this paper, we describe 6 metagenomic libraries constructed from soil samples collected from across the United States and containing ~12-20M cosmids clones each (Figure S1).

Each sub-pool was sequenced to an average coverage of 25x using Illumina short-read sequencing technology. Raw reads were assembled using standard bioinformatic pipelines and assembled contigs were annotated and ingested into a scalable, custom database for further analysis. On average, we generated ~444Gb of assembled sequence per library with an average of ~229Gb (~51%) of data contained on contigs >10kb (Figure 1a). In comparison to the 5 largest soil metagenomic JGI datasets, on average, we generated 11.6X the amount of assembled sequence per library with >140X more assembled data contained on contigs >10kb, demonstrating the utility of our approach to increase the quality of the assembled sequence. Merging contigs across sub-pools can further improve assembly quality, work not discussed here. In total, across the six libraries, we generated ~2.7Tb of assembled sequence, of which ~1.4Tb is on assemblies >10kb.

In order to assess the relative diversity found in each library, we annotated open reading frames in each library using Prodigal and compared the protein content (de-replicated at 90% amino acid identity) of the six libraries to each other as well as to the UniRef90 dataset. On average, each complete library contained 151M protein coding sequences, which is approximately the size of the most recent release of the UniRef90 dataset (release 2022_02). We saw <0.6% overlap between any of the libraries with UniRef90, consistent with the idea that the vast majority of soil microbial diversity has never been cultured and would not be found in public databases populated primarily with cultured organisms (Figure 1c). Strikingly, we saw only a 2.4% average overlap across all pairwise combinations of metagenomic libraries suggesting that the microbial diversity in each soil sample was largely orthogonal to the other samples. In total, across the six libraries, we annotated 901.3M open reading frames (de-replicated at 90% amino acid identity by library).

We used antiSMASH (Blin et al., 2019) to predict a total of >6.8M BGCs in the six sequenced libraries (Figure 1d). Of these, NRPS systems represent the most common class (~2.5M), followed by terpenes (~1.3M), RiPP's (~1.2M), and polyketides (~1.1M). While these numbers represent non-deduplicated BGC counts, we anticipate that, ultimately, these data contain millions of distinct and novel BGCs based on the low protein coding sequence overlap within our own libraries as well as with public sequence. To our knowledge, this is the single largest database of BGCs in the world and it continues to grow. For comparison, a recent analysis of ~170,000 genomes in NCBI RefSeq database and ~47,000 metagenome assembled genomes from various sources identified a total of ~1.2M non-deduplicated BGCs (Gavriilidou et al., 2022).

A resistance gene-based search strategy can be used to rapidly search metagenomic diversity for BGCs encoding predicted bioactivities of interest

Our database of 6.8M BGCs provides access to a vast and untapped universe of novel chemistry and bioactivity. The challenge, then, becomes one of prioritization. How can one identify the small fraction of BGCs within this database that encode for molecules with the potential to have meaningful therapeutic activity?

An attractive approach to address this challenge is to leverage the presence of self-resistance enzymes (SREs) found within some BGCs (Tran et al., 2019) (Culp et al., 2022). An SRE often encodes for a variant copy of the essential enzyme targeted by the NP encoded by the BGC. This variant copy provides resistance to the toxic effects of the NP and enables the host to survive the production of these molecules. For example, within the BGC encoding for the proteasome inhibitor salinosporamide is an SRE encoding for a variant of the beta-subunit of the proteasome that is resistant to the effects of salinosporamide (Kale et al., 2011). For the purposes of genome-mining, the presence of an SRE within a novel BGC serves as a strong predictor for the function of the encoded NP and could be used to specifically identify inhibitors of a desired protein target. However, the utility of this approach has been limited by the relative rarity of SREs within characterized BGCs and the modest sizes of existing BGC databases..

To assess the efficacy of mining our metagenomic database with an SRE-based strategy, we selected human methionine aminopeptidase type 1 (HsMetAP1) as a protein target of interest. HsMetAP1 cleaves the N-terminal methionine residues of nascent peptides and plays an important role in protein regulation. It has also been identified as a target for potential antitumor compounds (Frottin et al., 2016) (Behan et al., 2019), and there is one example of a bacterial NP inhibitor of HsMetAP1, bengamide, that is encoded by a BGC that contains a methionine peptidase SRE (White et al., 2017). To identify novel BGCs from the metagenome that encode HsMetAP1 inhibitors, we searched a subset of our metagenomic database (~1.2M BGCs) for BGCs that contain a gene encoding a methionine aminopeptidase within the predicted boundaries of the cluster. The resulting BGCs were de-replicated and computationally prioritized (Figure 2a). In total, we identified 35 BGCs that met our criteria. These BGCs spanned a broad range of molecular classes, sizes, and predicted taxonomies (Table S1). Notably, none of the identified BGCs resemble any characterized biosynthetic systems found in the MiBIG database

(Kautsar et al., 2019) including that for bengamide, highlighting the novelty of metagenomic diversity.

Heterologous expression of BGCs containing putative MetAP1 resistance genes produced lysates with predicted inhibitory bioactivity

In order to assess the accuracy of these bioinformatic predictions, we selected two of these biosynthetic pathways for heterologous expression studies The first BGC, ZYM301, contains two predicted biosynthetic genes encoding for an NRPS (*mtpB*) and a methyltransferase (*mtpD*), a transporter gene (*mtpC*), two genes of unknown function (*mtpA* and *mtpE*), and the methionine aminopeptidase putative resistance gene (*mtpF*) (Figure 2b; Table S2). The domain structure of the NRPS (*mtpB*) is made up of three modules predicted by antiSMASH to incorporate N-methyl-L-tyrosine, N-methyl-L-threonine, and N-methyl-L-valine, followed by a thioesterase domain. The second BGC, ZYM302, contains a single-module PKS (*orf8*), an NRPS-like gene (*orf25*), a varied set of thirteen biosynthetic genes, two transporters, seven regulators and the methionine aminopeptidase putative resistance gene (*orf22*) (Figure 2b; Table S3). Very little can be predicted about the building blocks used in the biosynthesis of the ZYM302 metabolites, however analysis of the PKS and NRPS-like genes allow for some predictions. The single module PKS (*orf8*) is predicted to produce and incorporate 6-methylsalicylic acid based on its sequence similarity to ChlB1 in the chlorothricin gene cluster (51% ID) (Shao et al., 2006). The NRPS-like gene consists of an adenylation and thiolation domain predicted by antiSMASH to incorporate phenylalanine.

Cosmids containing the complete ZYM301 and ZYM302 BGCs were isolated from our eDNA libraries and the BGCs were subsequently subcloned into expression vectors. These BGC-containing vectors, as well as an empty vector control, were conjugated into the host strain, *Streptomyces albus* J1074. The resulting exconjugates were fermented in mO42 media and crude organic extracts were screened for HsMetAP1 inhibitory activity using an established colorimetric assay. We detected novel *Hs*MetAP1 inhibitory activity in both sets of extracts, consistent with our expectation that the putative resistance genes identified by our bioinformatic search can be used to predict the biological activity of NPs encoded by novel BGCs encoded in the metagenome (Figure 3a).

Identification of novel metabolites encoded by ZYM301 and ZYM302

LC-MS/MS analysis of the active extracts confirmed the presence of clone-specific metabolites in both *S.albus:ZYM301* and *S.albus:ZYM302*. For *S.albus:ZYM301*, differential analysis yielded a list of 20 features in positive ionization mode that were detected in the *S.albus:ZYM301* samples, but not in the empty vector control (Figure 3b). Molecular networking analysis demonstrated that 19 of these features form a network based on the similarity of their MS/MS spectral patterns (Figure S2). Further manual analysis and grouping of features that were in-source fragments or adducts of the same compound, yielded a final list of 12 novel compounds. These 12 compounds consist of one major species (m/z 867.5217), which we call metapeptin A, and 11 relatively minor species that include metapeptin B-E, all eluting between 3.5 and 4.4 min (Table 1).

Singly- and doubly-charged ions were detected for all compounds in full scan mode. Data-dependent MS2 spectra were triggered for nine of the 12 compounds (Figure S3). All compounds had a common fragment of m/z 164.1082, suggesting a fragment with composition $C_{10}H_{14}NO$ derived from an N-methylated tyrosine residue. The range of molecular weights between 834 and 898 Da indicated that more than one of each amino acid building block was incorporated into the molecule. A series of three compounds had differences of m/z 14.015, suggesting differential methylation patterns of the same scaffold. No matches to known compounds were found after dereplication against publicly available mass spectral databases. Searches against an in-house mass spectral database of previously observed features in this heterologous host also yielded no matches.

For *S.albus:*ZYM302, differential analysis yielded a list of 9 features in positive ionization mode that were detected in the *S.albus:ZYM302* samples, but not in the empty vector control (data not shown). Dereplication against internal and external mass spectral databases yielded no matches for these features. Apart from two features with identical masses (m/z 865.4912), but different retention times, no obvious similarities in MS/MS fragmentation patterns were observed for these 9 features. Bioactivity tracked with these two ions of interest in fractionation experiments, but, due to low titers and compound instability, we chose to focus on ZYM301.

Orthogonal fractionation strategy efficiently identifies bioactive species within complex mixtures

A major challenge in bioactive NP discovery is the unambiguous assignment of bioactivity to a specific molecule within a complex mixture. As with the ZYM301 BGC, it is common for a newly characterized strain and/or BGC to produce many novel or unknown metabolites. Rather than attempt to purify each of the 12 novel molecules produced in *S.albus:ZYM301*, we opted for a modified biochemometric strategy, similar to those developed by the Cech and Dorrestein laboratories (Caesar et al., 2019) (Nothias et al., 2018). The goal of this approach is to statistically link LC-MS/MS chemical metabolite profiles generated across a series of chromatographically generated fractions with bioactivity measurements from the same fractions, resulting in statistical correlations between specific metabolites and bioactivity. As a result of the high sensitivity of the MS-analysis, these methods are particularly useful at discerning activity associated with low titer metabolites that may be overlooked during purification due to coelution with more abundant molecules which can result in incorrect assignment of bioactivity (Kellogg et al., 2016).

We used three different types of fractionation in parallel rather than iterative rounds of bioactivity guided analysis, which both accelerated the workflow as well as reduced the risk of unsuccessful separation of novel metabolites during fractionation. An 8L fermentation of *S.albus:ZYM301* was extracted and fractionated using normal phase (silica), reverse phase (C18), and size-exclusion (LH20) chromatography. An identically prepared set of fractions were prepared from a 4L fermentation of an *S. albus* containing an empty vector control to serve as a negative control within the bioassay. Each fractionation generated seven fractions for a total of

21 experimental fractions across three types of columns, each of which were subjected to bioassay and metabolomics analyses (Figure 4a).

We used a Partial Least Squares (PLS) Regression model to prioritize differentially expressed compounds based on the selectivity ratio. The selectivity ratio measures the explained variance versus residual variance and has been shown to be successful at identifying bioactive compounds (Kellogg et al., 2016). By visualizing the selectivity ratios on the network plot for all features in the fractions, we clearly see one connected component of features that have the highest selectivity ratios (Figure 4b). Notably, the six differentially expressed features with the highest selectivity ratios, including several in-source fragments, were associated specifically with metapeptin B, one of the novel minor compounds produced in *S.albus:ZYM301* (Figure 4c), suggesting that most, if not all, of the observed bioactivity can be attributed to metapeptin B.

Metapeptin B is a novel cyclic depsipeptide that differs from metapeptin A by only a single methylation

Given higher production titers, we first elucidated the structure of metapeptin A using a combination of high-resolution electrospray ionization mass spectrometry (HRESIMS) and 1D and 2D NMR data (Figure 5). HRESIMS analysis demonstrated an ion peak at m/z of 867.5217, consistent with a molecular formula of $C_{46}H_{70}N_6O_{10}$. The ^1H NMR spectrum of metapeptin A has signal distribution consistent with a typical peptide, containing α and β protons, aromatic proton signals, as well as additional *N*-methyl signals. Extensive analysis of 2D NMR spectra, including the correlation spectroscopy (COSY) and heteronuclear single quantum coherence (HSQC), demonstrated the amino acid components as *N,N*-dimethyl tyrosine, *N*-methyl threonine and *N*-methyl leucine, consistent with our bioinformatic prediction. Heteronuclear multiple bond correlation (HMBC) spectra of metapeptin A established the linkage of amino acids, and backbone of the 14-member ring of a cyclic peptide (Table S5, Figure S4). A detailed structural elucidation is provided in the Supplementary Text.

The structure of metapeptin B was determined by tandem mass spectrometry (MS/MS) based on comparative analysis of the HRMS/MS fragmentation data with metapeptin A (Figure 5, Figure S5). The molecular formula of metapeptin B was established as $C_{45}H_{68}N_6O_{10}$ by HRESIMS. Compared to metapeptin A, the Δm/z = 14.0155 implicated the loss of a single CH_2 group. In order to determine its location, we first analyzed the tandem mass spectrometry spectrum of metapeptin A (protonated molecular ion of m/z 867) under optimized conditions. Metapeptin A produced six main product ions at m/z 676.4280, 434.2649, 416.2544, 289.1547, 164.1070 and 126.0550. These ions and the fragmentation pathways were assigned (Figure S5). Metapeptin B, under the identical mass spectrometry conditions, produced not only the same product ions as metapeptin A, but also a series of pairing ions with Δm/z = 14 pattern (Figure S5). The fragment pair with m/z 164.1070 and 150.0913 suggested that the missing CH_2 is associated with the *N,N*-dimethyl-Tyr moiety. While mass spectrometry was not able to further confirm which carbon is missing, we propose that metapeptin B is an *N*-mono-methylated congener of metapeptin A for two reasons. First, the first adenylation domain of the NRPS (*mtpB*) is predicted to utilize tyrosine as a substrate by anitSMASH, making the incorporation of *N,N*-dimethyl-hydroxyphenylglycine (HPG) highly unlikely. Second, an NRPS-embedded

N-methyltransferase domain is known to be “leaky” and to produce demethylated shunt products in addition to the mature products under culture conditions (Fukuda et al., 2004), or in the absence of S-adenosyl-L-methionine (Billich and Zocher, 1987). Notably, the yield of demethylated congeners of beauvericins, depsipeptides encoded by a dimodule NRPS, was much lower than that of the fully methylated products (Fukuda et al., 2004) and this result agrees with our observation that the yield of metapeptin B is ~22 fold lower than of metapeptin A.

Metapeptin B is a sub-micromolar inhibitor of EcMetAP and is highly selective for HsMetAP1 over HsMetAP2

To characterize their bioactivity, we ran the purified metapeptin A and metapeptin B material through a set of enzyme inhibition assays. As predicted by our biochemometric analysis, metapeptin B inhibits HsMetAP1 with an IC_{50} of ~50uM while metapeptin A shows no activity even at the highest concentration tested (2mM) (Figure 6a-b), highlighting the importance of the single methyl group difference between metapeptin A and B. Furthermore, metapeptin B, but not metapeptin A, had inhibitory activity against the *E. coli* methionine aminopeptidase homolog, EcMetAP, that was ~100-fold greater (IC_{50}=~500nM) than that observed against HsMetAP1 (Figure 6c-d), as might be expected for an inhibitor whose native target is the bacterial protein.

Notably, we saw no inhibition of the other methionine peptidase in the human genome, HsMetAP2, by metapeptin B at 2mM (Figure 6e) indicating that unlike other known HsMetAP1 inhibitors such as bengamide (García-Ruiz and Sarabia, 2014), metapeptin B is highly selective for HsMetAP1 over HsMetAP2.

Asymmetric methylation of metapeptin B stabilizes the interaction of the cyclized dimer with HsMetAP1

Availability of natural congeners at the point of discovery enables rapid SAR insights, and, in this case, highlights the importance of both asymmetric methylation and macrocycle ring formation. To investigate the latter, we designed and synthesized the singly methylated tripeptide monomeric structure (NMe-Monomer) (Figure S6). The linear NMe-Monomer demonstrated no detectable inhibition at 2mM, confirming the cyclization of metapeptin B is critical for its target engagement, most likely due to conformational rigidity (Figure 7a).

To gain further insight into metapeptin B SAR, molecular docking, MMGBSA energy calculations, molecular dynamics, and pKa calculations were performed. All methods indicated better binding for the asymmetric monomethyl metapeptin B. Induced-fit docking on HsMetAP1 (PDB:6LZC) found that the conformations of the A and B variants show differences in both the methyl amine sites and a flipping of the methyl amide linkers (Figure 7b). As a result of these differences, the MMGBSA energies show a significant separation, with metapeptin B having a lower, more favorable level, largely due to a change in DDGs solvation energy. Additionally, the strength of the overall dipole of the two conformations differ by 7.05D (metapeptin A) versus 7.34D (metapeptin B) resulting in stronger electrostatic interaction with the metal ions for metapeptin B. Metapeptin B also has a favorable orientation pointing at the metal ions in the pocket (Figure S7). When we look at the effects of pKa, we find that the pKa of metapeptinA (7.76) versus metapeptinB (7.46) implies that a larger fraction of metapeptin A will be positively charged and have an unfavorable interaction with the two positively charged metal ions. Finally, molecular dynamics simulations indicate a more stable binding for metapeptin B. An RMSD of the length of the key hydrogen bond of the monomethyl amine of metapeptin B to H212 indicates a higher interaction strength, as quantified by the frequency of occurrences in the molecular dynamics trajectory, compared to the dimethyl amine of metapeptin A (Figure 7c).

In summary, the combination of experimentation and cheminfomatic modeling based on the naturally available congeners, enables rapid insight into key structure-activity relationships and provides a foundation for the design of advanced analogs

Metapeptin B does not inhibit mtpF confirming self-resistance enzymes can be used to predict the function of unknown BGCs and identify structurally novel inhibitors

Our bioinformatic search strategy is based on our ability to identify bonafide SREs. To confirm that the putative SRE gene in the metapeptin BGC (*mtpF*) functions as a resistance gene, we tested the ability of metapeptin B to inhibit the activity of the methionine peptidase variant encoded by *mtpF*. Consistent with its predicted role in alleviating metapeptin B toxicity, we found that metapeptin B had no effect on the activity of the methionine peptidase encoded by *mtpF* at the highest concentrations tested (200mM) (Figure 8)

DISCUSSION

A “next-gen” natural product discovery platform

In this paper we describe a *de novo* NP discovery platform that enables targeted discovery of novel NP modulators of protein targets of interest. We leverage access to vast, high-quality metagenomic data from soil microbiomes, bioinformatic and data science approaches to identify novel BGCs predicted to encode molecules with the desired bioactivity, and synthetic biology workflows to produce and characterize these novel molecules. By doing so, we are able to surmount key challenges of traditional fermentation-based NP discovery and identify a vast untapped source of advantaged chemistry. We also demonstrate how to access this diversity in a highly targeted manner for precision drug discovery.

Novelty: Fermentation-based NP discovery has been plagued by high rates of rediscovery, in large part due to the dependence on cultured strain collections and standard fermentation processes that leave many BGCs silent (Tomm et al., 2019). We are able to bypass these limitations by enabling unprecedented access to metagenomic diversity. Our approach to capturing and sequencing soil microbiomes generates orders of magnitude more high-quality sequence information (contigs >10kb) than standard shotgun sequencing. This, in turn, enables the annotation of millions of BGCs that, importantly, come from diversity that is orthogonal to widely-explored cultured diversity. We find that there is little overlap between our sequenced metagenomic libraries and public databases (<0.6%), consistent with estimates that >99% of microbial diversity in soil is not easily cultured. Strikingly, there is also little overlap between our six metagenomic samples as well (~2.4% between pairs of libraries), suggesting that we have only begun to scratch the surface of metagenomic diversity (discussed more below). As such, our data demonstrate that metagenomic diversity represents a very distinct and deep source of novel NPs. Furthermore, with the sequence of metagenomic BGCs in hand, it becomes straightforward to identify which metagenomic BGCs have high similarity to known BGCs and may encode the same or similar molecules.

Indeed, in this paper, our *in silico* analysis predicted 35 novel BGCs as likely to encode methionine peptidase inhibitors. These BGCs encode a broad range of biosynthetic classes (e.g. NRPS, PKS, terpenes, ladderanes, etc.) (Table S1), and none have similarity to characterized BGCs in the MiBIG database. Both ZYM301 (NRPS) and ZYM302 (NRPS-like) demonstrated the predicted bioactivity and ZYM301 was found to encode metapeptin B, a novel depsipeptide inhibitor. Taken together, these results demonstrate the power of this approach for finding novel, chemically diverse bioactive molecules for a specific target protein of interest.

Target-focused and efficient: Fermentation-based NP discovery typically relies on libraries of extracts that generally are not compatible with high-throughput, target-based assays favored at most pharmaceutical companies (Atanasov et al., 2021). Mixed extracts pose issues around liquid handling, assay interference, toxicity, background inhibition, and more (Henrich and Beutler, 2013). Instead, these libraries of extracts are more often screened in phenotypic assays with readouts such as cell death, cell morphology changes, or changes in gene expression (Wilson et al., 2020). The challenge posed by phenotypic screening, however, is that the observed effects are mediated through unknown mechanisms and require extensive down-stream mechanistic studies before relevance to a particular protein target or pathway of interest can be determined. For many pharmaceutical companies, which are increasingly target-focused in their approach towards drug discovery, this level of effort before determining relevance is unacceptable.

By leveraging a resistance-gene search strategy, we place validated targets at the center of our approach. In this paper, we selected HsMetAP1, a key translational regulator with strong associations across a wide range of solid tumor cancers, as our protein target of interest. Rather than screening through thousands of fermentation extracts with assays subject to significant false positive and false negative rates, we quickly down-selected from >1 million BGCs to 35 candidates *in silico* to construct ultra-enriched and highly compact libraries of advantaged chemical matter that can then be subjected to sensitive analysis.

We believe a resistance-gene based search strategy will be applicable for a broad set of human protein targets. First, there is abundant precedent for bacterial NPs being approved as modulators of human therapeutic targets. Rapamune (based on rapamycin) and Kyprolis (based on epoxomicin) are two well known examples. Indeed, when we applied our resistance gene search approach to find proteasome inhibitors, we recovered the epoxomicin family, the salinosporamide family, as well as putative novel inhibitors (data not shown). Second, this approach does not appear to be particularly sensitive to evolutionary distance, perhaps reflecting the importance of the core biology carried out by deeply conserved protein families. Methionine peptidases are approximately ~40% identical between humans and bacteria while proteasome subunits are only ~30% conserved. Despite the relatively low sequence identity, in both cases, we see conservation of protein function and pharmacological tractability. Lastly, while this paper has focused on our bacterial metagenomic database, we have built a similar fungal metagenomic database. We foresee that some targets that are less tractable with bacterial NPs will be readily tractable with NPs that have evolved specifically to target eukaryotic proteins.

Capabilities needed to successfully operate this “next-gen” platform

While this “next-gen” NP discovery approach is extremely attractive for the reasons just discussed, it is important to note that in order to turn this workflow into a platform that can be run repeatedly, successfully, and at scale requires the development of a broad and non-trivial set of capabilities.

Metagenomic diversity: Access to large-scale, well-assembled metagenomic DNA is central to our approach. Not only is improved assembly quality critical to enable BGC discovery (described above), but significant database scale is required to leverage resistance genes as a search strategy. In the case of the methionine peptidase family analyzed in this study, we estimate that no more than 1 in 20,000-80,000 BGCs would meet our criteria for further exploration. While this number will no doubt vary for other target families, it implies that millions of BGCs are required to confidently pursue a drug discovery program around a given target given their rarity. The converse is also informative. Only one bacterial NP inhibitor of HsMetAP1, bengamide, was previously known from public databases. Our ability to identify 35 additional novel BGCs encoding putative HsMetAP1 inhibitors (including two validated BGCs) suggests that the field has only uncovered the tip of the iceberg of resistance gene-containing BGCs. To discover more, however, will require significant investments to expand metagenomic diversity.

Our results contrast sharply with the inherent limitations of strain collections. Indeed, our data indicate we are far from saturating metagenomic diversity. In addition to the lack of overlap between different metagenomic samples on a protein level, we only found the ZYM301 and ZYM302 BGCs once in our metagenomic sequence data. We did not find the bacterial BGC encoding bengamide at all, despite the amount of data we generated. Taken together, we have detected no evidence of diminishing returns in terms of protein diversity, BGC diversity, and presumably chemical diversity, and we do not anticipate being able to reach saturation for the foreseeable future.

Data Infrastructure and Data Science: This approach requires sufficient data infrastructure to not only capture and store enormous amounts of information, but to enable rapid search, retrieval, and manipulation of the information as well. Well-conceived data architecture, computational infrastructure, and efficient bioinformatic tools are critical, especially as the amount of sequence information continues to increase.

On the experimental side, data science approaches are imperative for prioritizing work on the BGCs and molecules with the most promising activities. In the case of BGC selection, data science will improve the accuracy of resistance gene identification and assessments of BGC quality by leveraging experimental data and a range of criteria embedded in the sequence information. Similarly, the application of data science to analytical data will improve the sensitivity and accuracy of molecule detection and characterization as well as drive strategies to increase titer.

Synthetic Biology: Lastly, heterologous expression of BGCs is required to access metagenomic diversity. In this paper, we focused on two BGCs, ZYM301 and ZYM302, that were conjugated into *S. albus* without promoter refactoring. Both expressed sufficiently to confirm activity but only ZYM301 expression was adequate for isolation. ZYM302 illustrates the importance of high titers to a robust discovery workflow. In this case, all experiments were done in a single host (*S. albus*) with one media (mO42) and with modest process optimization reserved for ZYM301. While optimization of these parameters can no doubt result in titer improvements, reliance on fermentation optimization bears similarities to historical pharmaceutical approaches and is operationally challenging to scale. In our experience, synthetic biology - specifically high-throughput multi-edit genetic engineering of BGCs - is key to controlled heterologous expression and enabling a robust metagenomic discovery workflow.

Relatedly, expression titers become exponentially more important as programs advance. As part of this work, we have been developing a total synthesis of Metapeptin B (not shown). While this is highly enabling for analog generation during lead identification, the efficiency of ring cyclization is low, and would not support downstream lead optimization, preclinical development, nor API supply. A semi-synthetic process based on fermenting a key intermediate will therefore be required. In our experience, genome-wide host engineering (involving hundreds to tens of thousands of edits) on top of intensive cluster engineering is required to achieve commercial production economics. This is an important consideration for the success of our overall approach.

Conclusion

NPs have provided some of the most important small molecule drugs in our pharmacopeia and have impacted human health for decades and arguably millennia (Newman and Cragg, 2020) Today, we are in an era of AI-enabled drug discovery (Chopra et al., 2022), using DELs and other large-library formats to explorechemical space (Gironda-Martínez et al., 2021) and going “beyond the rule-of-5” (Caron et al., 2020) to advance small molecule drug discovery. Our “next-gen” approach to NP discovery allows us to tap into billions of years of evolution and thus provides an orthogonal though complementary approach to expand chemical space and identify advantaged starting points for drug discovery. Indeed, NPs can be powerful starting points to enable AI-optimization (Begnini et al., 2021) (Bergner et al., 2019). Most importantly, we have demonstrated two things. First, the universe of NPs that are relevant for targeted human drug discovery is likely much larger than previously imagined. Second, by combining metagenomics, data science and synthetic biology it is possible to do precision, data-driven, and scalable NP drug discovery.

MATERIALS AND METHODS

eDNA libraries construction and sequencing

Soil eDNA libraries were constructed following a protocol adapted from the Brady protocol (Brady, 2007). Briefly, eDNA was released from ~125g to 500g of soil sifted through a 2mm sieve by suspending it in an SDS detergent buffer and heating the mixture at 70°C for 2 hours. After removing debris by centrifugation, DNA was extracted by alcohol or PEG precipitation and further purified and size selected by field inversion gel electrophoresis. DNA in the 35-40kb fraction was end-repaired and ligated into an appropriate library vector. Cosmid libraries were constructed using MaxPlax™ Lambda Packaging Extracts (Lucigen Corp, Middleton, WI) as sub-pools of between 6,000 and 25,000 cosmid clones. Each library was built out to contain ~20 million clones, with the exception of the Radiant library, which contained ~12 million clones.

Prior to preparing sequencing libraries the cosmid backbone was removed by digesting each DNA sub-pool with an appropriate restriction enzyme and removing the small backbone fragments. Purified insert sub-pools were then tagmented and barcoded using unique combinations of Illumina i5 and i7 index primers. Final barcoded DNA libraries were size selected and pooled at 1-4X 96-well plates per NovaSeq run, depending on the library complexity.

Sequence Annotation

Following assembly, using metaSPAdes (Nurk et al., 2017), open reading frames (ORFs) were called using Prodigal (Hyatt et al., 2010) and taxonomic classification for each contig was assigned using Kaiju (Menzel et al., 2016). ORFs were further annotated with PFAM domains using hmmer (Eddy, 1998) and PFAM-A database (Mistry et al., 2021). Biosynthetic gene clusters were predicted on all contigs using antiSMASH v5 (Blin et al., 2019).

JGI analysis

To evaluate how Zymergen soil metagenomic libraries compare to the largest publicly available soil metagenomic datasets, we looked to the Joint Genome Institute's Integrated Microbial Genomes & Microbiomes (IMG/M) database (https://img.jgi.doe.gov/cgi-bin/m/main.cgi). We identified the 5 largest datasets in IMG/M based on the assembled genome size, corresponding to IMG Genome IDs 3300050821, 3300043331, 3300042731, 3300043313, and 3300045391. The assembly lengths of all contigs larger or equal to 10kb were extracted from the coverage statistics file associated with each dataset and summed to yield the total assembly lengths contained on contigs >10kb.

Library comparisons

To evaluate the overall novelty of sequence in the libraries, we first clustered each library independently at 90% amino acid sequence using MMseqs2 algorithm (Steinegger and Söding, 2017), such that the clusters were composed of sequences containing at least 11 residues and at least 80% overlap with the seed sequence of the cluster (*i.e.,* the longest sequence in the cluster). Subsequently, the representative cluster outputs from individual libraries were co-clustered together with the UniRef90 dataset (Release: 2020_01).

MetAP1 resistance gene search

To identify homologues of HsMetAP1, a BLASTp search was run against the Radiant and CA29 metagenomic libraries using the HsMetAP1 gene (P53582) as the query. To allow the search to include distant homologues, a permissive cut off of e-value < 10 was chosen. The resulting hits were narrowed down to those that share a contig with a BGC, as called by antiSMASH (Blin et al. 2021). The BGCs closely associated with the MetAP1 homologues were run through BigScape to group together similar clusters, allowing dereplication at a cutoff of 0.7 (Navarro-Muñoz et al., 2020). Finally, the remaining BGCs were prioritized based on a number of factors including the similarity of the putative resistance gene to HsMetAP1, the distance of the putative resistance gene from the BGC, the presence of the resistance gene in an operon with a biosynthetic gene, and the predicted completeness of the cluster. This resulted in a list of 35 “high quality” clusters that were candidates for producing a inhibitor for methionine aminopeptidases.

Isolation of cosmids from eDNA libraries

To isolate cosmids containing BGCs of interest, primers specific to the clusters of interest were designed to generate a 400-500kp amplicon. These primers were used to track the cosmids through multiple rounds of serial dilutions by PCR. Briefly, a library well containing the cosmid of interest was used to inoculate an overnight culture in LB supplemented with 100 µg/ml carbenicillin. Using OD600nm, approximately 30 cells/well were inoculated in a 384-deep well plate and grown overnight. Positive wells as assayed by PCR were then plated to single colonies on LB (100 µg/ml carbenicillin) Q-trays to select for colony-PCR positive clones. Isolated clones were sequence verified by Illumina NGS sequencing via tagmentation.

Assembly of BGCs into heterologous expression vectors

For heterologous expression of ZYM301 and ZYM302 after isolating the cosmids containing the BGCs, yeast homologous recombination was used to transfer the BGCs to integrative (pTARw) expression vectors containing yeast replication origins and selection markers. Briefly, pTARw were digested with I-SceI and PacI to linearize and expose ends that have homology to cosmids sequences flanking the eDNA encoding ZYM301 and ZYM302 respectively. These digested vectors were co-transformed with the cosmids of interest into *Saccharomyces cerevisiae* (BMA64) using a standard LiAc/SS carrier DNA/PEG method (Gietz and Schiestl, 2007). DNA from PCR-positive yeast colonies were isolated using the ChargeSwitch™ Plasmid Yeast Mini Kit (Invitrogen) and electroporated into epi300 electrocompetent cells. DNA extracted from positive colonies checked by cPCR was sequence verified by Illumina NGS sequencing via tagmentation.

Heterologous expression

Heterologous expression plasmids for ZYM301 and ZYM302 were transformed into *Escherichia coli* S17.1 cells and transferred into *Streptomyces albus* via conjugation. Exconjugant colonies that grow on Mannitol Soya (MS) agar plates with nalidixic acid (30 µg/ml) and apramycin (50 µg/ml) were restreaked to single colonies. Four colonies that passed colony PCR verification were glycerol stocked and then used to inoculate triplicate 3mL seed cultures (starting OD450=0.05) in Tryptic Soy Broth (TSB) + apramycin (50 µg/ml) along with appropriate vector and media only controls. Seed cultures were grown for 3 days at 30°C and then diluted 1:10 in 4 different media (O42, mO42, R5A and ISP4) for fermentation. After 7 days of incubation at 30°C for 7 days the cultures were then extracted as described below for AC analysis.

MetAP1 enzymatic assay validation

The primary methionine aminopeptidase colorimetric activity assay is based on a commercial kit (R&D Systems) and further developed and optimized as follows. The reaction occurs in two steps: enzyme activation and product detection. *Enzyme activation*, each well contains a 25 uL mixture of activation buffer: 50 mM HEPES, 0.1 mM CoCl2, 0.1 M NaCl, pH 7.5; 100 uM of fluorogenic tripeptide substrate Methionine-Glycine-Proline-7-amido-methylcoumarin (R&D Systems ES017); and 2 ug/mL of MetAP enzyme (R&D Systems 3537-ZN). *Product detection*, each reaction well is supplemented with 25 uL of 2 ng/mL DPPIV/CD26 (serine exopeptidase diluted in activation buffer).

The assay was performed at room temperature in an opaque 384-well polystyrene plate and measured in a black, flat bottom, 384-well plate. The assay is initiated as MetAP is mixed with the tripeptide substrate and incubated for 5 minutes.The initial reaction will result in the cleavage of methionine, yielding a dipeptide product, Gly-Pro-AMC. MetAP is then inactivated by heating the reaction to 100 ºC for 5 min and cooled on ice for an additional 5 min. Detection of the dipeptide product is carried out via two orthogonal methods; degradation of the dipeptide product via DPPIV protease and targeted LC-MS analysis of the dipeptide product. Incubation of the reaction mixture with a DPPIV solution at room temperature for 10 minutes results in the hydrolysis of the dipeptide and release amido-methylcoumarin, which is measured at 380/460 nm excitation/emission on a Tecan Spark microplate reader. The release of AMC, measured via fluorescence, corresponds to the activity of the MetAP under analysis. Alternatively, the dipeptide-AMC product is measured via LC-MS, using a standard (Sigma Aldrich G2761) to quantify MetAP activity. Both analysis methods have been proven to yield comparable signals, validating either method as a tool characterizing enzyme activity.

MetAP1 enzymatic assay background controls

As is typical in enzyme kinetics, the initial rate of enzyme hydrolysis is used to measure enzyme activity, with background hydrolysis being subtracting from the enzymatic output to observe accurate MetAP activity. Background controls were vetted by removing key reagents from the reaction mixture to ensure that activity was dependent on the expected agents (substrate, enzyme, co-factors) and when any one was not present, enzyme activity above background was not observed. All MetAP enzymes tested tolerated up to 10% DMSO and 5 % Methanol without significant enzyme inactivation.

Two additional controls were established to confirm that observed inhibition is specific to MetAP and not the DPPIV protease used in the colorimetric readout. First, bioactive fractions/molecules (including metapeptin B) were incubated with the dipeptide product and the DPPIV protease. In all cases, the fractions/molecules showed no inhibition of the hydrolysis of the dipeptide by the DPPIV protease and release of AMC, indicating that the observed inhibition is specific to MetAP. Second, to address concerns about potential false positives with the colorimetric assay, LC-MS analysis was also used to confirm the increase in the MetAP cleavage product, GP-AMC, in positive control reactions.

Finally, *S. albus* fermentation extracts can have overlap in fluorescence with the AMC readout used to measure activity in the colorimetric assay. To control for this potential interference, an inhibitor control was established by incubating the test inhibitor in the activation buffer and measuring its fluorescence. This value was then subtracted from the MetAP+inhibitor reaction to determine the true impact of the inhibitor on enzyme activity. Conversely, *S. albus* extracts may also cause non-specific inhibition at relatively high concentration within the assay. We determined that 0.1-0.5 mg/ml was an acceptable range for the working concentration of extracts/fractions that enable the detection of inhibition while maintaining relatively low background fluorescence.

Sample Preparation for UPLC-MS/MS analysis

Three mL of fermentation broth was extracted twice with 3 mL ethyl acetate (HPLC grade, Fischer) by shaking for 1 minute at 1000 rpm followed by sonication for 15 minutes. Samples were centrifuged, and the organic layer was collected and pooled to yield 6 mL of extract. The ethyl acetate was removed under reduced pressure in a Speedvac Savant (Thermo). Dry samples were suspended in 120 uL methanol (LC-MS grade, Fisher) and transferred to HPLC vials. A pooled sample for each medium was generated by combining 30 uL aliquots from replicate samples of each medium type.

UPLC-MS(/MS) data acquisition

Samples were subjected to ultra performance liquid chromatography mass spectrometry on a Q-Exactive Mass Spectrometer (Thermo Fisher) connected to a Vanquish Liquid Chromatography system (Thermo Fisher). A gradient of water (mobile phase A) and acetonitrile (mobile phase B), each containing 0.1% formic acid, was employed with a flow rate of 0.5 mL/min on a Zorbax Eclipse Plus C18 RRHD 2.1 x 50 mm, 1.8 µm column (Agilent), operated at 40C. The gradient started at 2% mobile phase B, holding for 1 min, followed by a linear gradient to 100% mobile phase B over 7 minutes, and then held at 100% mobile phase B for 2 minutes, returning to initial conditions over 0.1 min and holding for 0.9 min for a total run time of 11 min. The mass spectrometer was operated at spray voltage: 3.6 kV, capillary temperature: 275, sheath gas flow rate: 25, auxiliary gas flow rate: 10, S-lens RF level: 70. Full scan mass spectra were acquired in positive and negative ionization mode from m/z 200-1500 at 70K resolution and ACG target of $3e^6$, and maximum ion fill time of 200 ms. Data-dependent MS2 spectra were acquired in positive and negative ionization mode for pooled samples, collecting a full MS scan from m/z 200-1500 at 70K resolution and ACG target of $3e^6$. The top five most abundant ions per scan were selected for MS/MS with a resolution of 17.5K and ACG target of $1e^5$, and stepped collision energies of 10, 20 and 40 NCE. Maximum ion fill time was 50 ms, dynamic exclusion was 3 sec, and an isolation window of 1 m/z was used.

Untargeted Data analysis

Positive and negative ionization mode datasets were obtained by acquiring full scan mode data of each sample, as well as data-dependent MS2 data of each pooled sample. Raw data were exported to Compound Discoverer Software (v3.1, Thermo) for deconvolution, alignment and annotation. Putative novel feature dereplication was performed against an in-house database of previously acquired features, using a custom Python script which matched features within a mass and retention time threshold of 5 ppm and 0.2 min, respectively. MS2 data were converted to mzml format using Compound Discoverer and exported to Ometa Labs Flow Analysis Platform. Molecular networking analysis was performed using the Classical Networking workflow (see supplemental information for workflow parameters). MS2 spectra of related compounds were grouped within the dataset according to similarity, and searched against reference spectral libraries (GNPS, NIST, MoNa).

Extraction Methodology for Orthogonal Fractionation

ZYM301: The 8 L of ZYM301 and 4 L of empty plasmid control were processed identically. Bacterial cells were first removed via centrifugation (5000 RPM, 15 min) and discarded. The clarified broth was extracted with a 5% (w/v) addition of activated HP20 resin, and allowed to gently stir overnight. The resin was filtered from the aqueous broth, then extracted with methanol (2 × 1 L), followed by acetone (2 × 1 L). The organic fractions were combined and dried *in vacuo*, yielding a thick aqueous suspension. The aqueous layers were diluted to 500 mL using distilled water, and then partitioned against an equal volume of ethyl acetate four times (4 × 500 mL). The ethyl acetate layer was dried over $MgSO_4$, filtered, and finally dried *in vacuo* yielding 522.18 mg of BGC extract, and 249.55 mg of control extract. These extracts were reconstituted in methanol, and partitioned into three roughly equal aliquots for fractionation.

ZYM302: ZYM302 (2 L) and the corresponding empty vector control (2 L) were extracted identically. Cells were removed via centrifugation (5000 RPM, 15 min) and discarded. Clarified broth was extracted via liquid-liquid partition using an equal volume of a 4:1 mixture of ethyl acetate and isopropanol (3 × 2 L). Organic layers were combined, and dried *in vacuo* yielding thick brown oily residues for both ZYM302 and its corresponding empty vector control (3.585 g and 1.571 g respectively). This material was partitioned into ~200 mg aliquots for further processing.

Orthogonal Fractionation

Silica fractionation ZYM301: Flash chromatography was performed using a Biotage Selekt automated chromatography system utilizing pre-packed Biotage Sfar HC Duo (10 g) silica columns. Both the empty vector control extract (79.35 mg), and the ZYM301 extract (173.21 mg) were fractionated identically. Material was fractionated using a flow rate of 40 mLmin^{-1}, collecting 60 mL fractions (4 CV). Material was eluted using a three-solvent system, consisting of hexanes (solvent A), ethyl acetate (solvent B), and methanol (solvent C). The column was initiated with a linear increasing gradient from 30% to 100% solvent B in solvent A for 12 CVs (F1-F3). This was followed by an isocratic elution using 100% solvent B for 4 CVs (F4). This was followed by another linear increasing gradient from 10% to 80% solvent C in solvent B over 8 CVs (F5-F6). Finally, an 80% isocratic wash of solvent C in solvent B was performed, over 8 CVs. This generated another 2 fractions that were combined into a single final fraction (F7), yielding 7 fractions in total for both extracts. Fractions were dried into pre-weighed vials using a V10-touch evaporator (Biotage) coupled with a Gilson GX-271 Liquid Handler. Fractions were used for both bioactivity assessment and MS-analysis without further purification. For MS-analysis, samples were brought up to a concentration of 1 mgmL^{-1}, and 4 μL was injected and run using the UPLC-MS/MS method previously described.

Silica fractionation ZYM302: Fractionations were carried out identically for both ZYM302 (204.8 mg) and empty vector control (201.1 mg). Fractionations and MS-analyses were carried out using the same gradient, flow rate, drying procedure, and sample concentrations as previously described for ZYM301. The only difference was 30 mL fractions were collected (2 CV each). Using the same solvents and gradient for elution as described above, the fractions were generated as follows, 12 CVs (F1-F6), 4 CVs (F7-F8), and 8 CVs (F9-F12). The final 8 CVs were divided into 2 × 4 CV blocks (F13-F14), yielding 14 fractions total.

C18 Fractionation ZYM301: Flash chromatography was performed using a Biotage Selekt automated chromatography system utilizing pre-packed Biotage Sfar C18 (12 g) columns. Both the empty vector control extract (83.98 mg), and the ZYM301 extract (174.10 mg) were fractionated identically. Material was eluted with a flow rate of 12 mLmin^{-1} collecting 68 mL fractions (4 CV). Material was eluted using a simple 2-solvent gradient system, consisting of H_2O (solvent A), methanol (solvent B). The column was first washed with 5% methanol in H_2O for 4 CV (F1), followed by a linear increasing gradient from 5% to 100% methanol over 20 CV (F2-F6). An isocratic gradient of 100% methanol was then applied for 8 CV, and this wash was combined into one fraction (F7), yielding 7 fractions in total. Fractions were dried into pre-weighed vials using a V10-touch evaporator (Biotage®) coupled with a Gilson GX-271 Liquid Handler. Fractions were used for both bioactivity assessment and MS-analysis without further purification. For MS-analysis, samples were brought up to a concentration of 1 mgmL^{-1}, and 4 μL was injected and run using the UPLC-MS/MS method previously described.

C18 Fractionation ZYM302: Fractionations were carried out identically for both ZYM302 (208.5 mg) and empty vector control (200.8 mg). Fractionations and MS-analyses were carried out using the same gradient, flow rate, drying procedure, and sample concentrations as ZYM301. The only difference was 34 mL fractions were collected (2 CV each). Using the same solvents and gradient for elution as described above, the fractions were generated as follows, 4 CVs (F1-F2), and 20 CVs (F3-F12). The final 8 CVs were divided into 2 × 4 CV blocks (F13-F14), yielding 14 fractions total.

LH20 Size-Exclusion Fractionation ZYM301: Size-exclusion chromatography was performed with a hand packed LH20 column (15.9 × 600 mm). Both the empty vector control extract (86.22 mg), and the ZYM301 extract (174.87 mg) were fractionated identically. Material was eluted using 100% methanol, with a flow rate of ~0.5 mLmin^{-1} collecting 4 mL fractions over 180 mL, yielding 45 initial fractions. Due to mass limitations, fractions 1-10 were combined (F1), and then every 4 fractions from 11-30 (F2-F5), and all remaining fractions (31-45) were combined (F7) yielding 7 fractions in total. Fractions were dried using a Speedvac Vacuum Concentrator, and used without further purification. For MS-analysis, samples were brought up to a concentration of 1 mgmL^{-1}, and 4 µL was injected and run using the UPLC-MS/MS method previously described.

Orthogonal fractionation data analysis

The input features for the PLS analysis were all features from the gene cluster and empty vector sample for which MS2 data was captured. Features that were more than 0.9 cosine similarity were combined into a consensus spectrum. For each feature, the peak area was determined via XIC integration with 0.2 tolerance on the retention time. Peak areas were then normalized to sum to equal amounts for each fraction. For the union of all 38523 features across all 21 fractions (7 fractions for 3 different fractionation methods), the normalized peak areas were cast into a feature matrix of dimension 21 x 38523. The corresponding bioactivity vector of dimension 21 was composed by taking the average inhibition percentage across the three replicates and subtracting the inhibition observed in the same fraction for the empty vector.

A PLS analysis with two components and standard scaling was then run on the resulting feature matrix and bioactivity matrix, and the selectivity ratios were calculated from the resulting PLS vectors. Features with a differential expression ratio (measured by the sum of peak areas across all fractions for the gene cluster sample versus empty vector control) less than 100 were disregarded. The resulting selectivity ratios were tabled and plotted on top of a network plot for all gene cluster and empty vector sample features by scaling the node size for each feature. The network plot was run with a cosine similarity cutoff of 0.7, 8 matching peaks and a maximum shift of 250.

Purification

After 7-day culture period, 180L worth culture flasks were combined and centrifuged (4,000 rpm for 10 min). Mycelial portion was discarded. The supernatant was absorbed onto HP20 resin (5%, w/v) for overnight overhead spinning. The resin was filtered and washed with water to remove water soluble components. The resin was extracted in ethyl acetate (12L). The organic phase was dried *in vacuo* to afford 70g of dried crude extract.

Crude extract was subjected to a size exclusion column packed with Sephadex LH-20 and manually fractionated in methanol. The fractions were screened by LC-MS. Fractions containing the compounds of interest were combined and subjected to preparative HPLC purifications. The first round of HPLC was conducted on a C18 column (250 X 10 mm, phenomenex, CH_3CN-H_2O, 0.1% FA, flow rate: 8ml/min), and the second round of HPLC was conducted on a phenyl hexyl column (250 X 10 mm, phenomenex, CH_3CN-H_2O, 0.1%FA, flow rate: 8ml/min).

HPLC fractions containing the compounds were dried to yield metapeptin A (25mg) and metapeptin B (1.8mg).

Metapeptin A, white, amorphous powder, (+)-HR-ESIMS m/z = 867.5224, $[M + H]^+$ (calcd for $C_{46}H_{71}N_6O_{10}{}^+$, 867.5226)

Metapeptin B, white, amorphous powder, (+)-HR-ESIMS m/z = 853.5069, $[M + H]^+$ (calcd for $C_{45}H_{69}N_6O_{10}{}^+$, 853.5070)

NMR

All NMR spectra were recorded on a Bruker AVANCE at 900 MHz (^{1}H NMR) and 226 MHz (^{13}C NMR). NMR spectra were analyzed using Mestrenova 9.0.1.

Monomer synthesis

The NMe-Monomer was designed with all natural (L) stereochemistry based upon bioinformatic prediction from the encoded BGC analysis. The molecule was synthesized in 8 steps as illustrated in Figure S6. Detailed synthetic procedures and characterization are available upon request.

Molecular modeling

Most simulations were carried out with Schrodinger's Small Molecule Drug Discovery Suite version 2022-1. Glide docking was done with the XP precision level with post-docking minimization and applying strain correction terms. MM-GBSA calculations used the VSGB solvation model with OPLS4 force field. Induced-fit docking was completed with the Extended Sampling protocol generating up to 80 poses, residues within 15 Å of the binding site were refined, and the docking was done with SP precision. Jaguar was used for the pKa calculations using B3LYP-D3 level of theory and the 6-31G** basis set. Maestro was used for computing dipoles and making the docking poses picture. The dipole result was a property calculated by QikProp. Desmond was used for the molecular dynamics simulations using the NPT ensemble class at a temperature of 300 K for 25ns. PDB ID 6LZC was used for the HsMetAP1 protein.

References

Atanasov, A.G., Zotchev, S.B., Dirsch, V.M., the International Natural Product Sciences Taskforce, and Supuran, C.T. (2021). Natural products in drug discovery: advances and opportunities. Nat. Rev. Drug Discov. *20*, 200–216. https://doi.org/10.1038/s41573-020-00114-z.

Begnini, F., Poongavanam, V., Over, B., Castaldo, M., Geschwindner, S., Johansson, P., Tyagi, M., Tyrchan, C., Wissler, L., Sjö, P., et al. (2021). Mining Natural Products for Macrocycles to Drug Difficult Targets. J. Med. Chem. *64*, 1054–1072. https://doi.org/10.1021/acs.jmedchem.0c01569.

Behan, F.M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C.M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., et al. (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature *568*, 511–516. https://doi.org/10.1038/s41586-019-1103-9.

Bergner, A., Cockcroft, X., Fischer, G., Gollner, A., Hela, W., Kousek, R., Mantoulidis, A., Martin, L.J., Mayer, M., Müllauer, B., et al. (2019). KRAS Binders Hidden in Nature. Chem. – Eur. J. *25*, 12037–12041. https://doi.org/10.1002/chem.201902810.

Billich, A., and Zocher, R. (1987). N-Methyltransferase function of the multifunctional enzyme enniatin synthetase. Biochemistry *26*, 8417–8423. https://doi.org/10.1021/bi00399a058.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., and Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. *47*, W81–W87. https://doi.org/10.1093/nar/gkz310.

Brady, S.F. (2007). Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nat. Protoc. *2*, 1297–1305. https://doi.org/10.1038/nprot.2007.195.

Caesar, L.K., Kellogg, J.J., Kvalheim, O.M., and Cech, N.B. (2019). Opportunities and Limitations for Untargeted Mass Spectrometry Metabolomics to Identify Biologically Active Constituents in Complex Natural Product Mixtures. J. Nat. Prod. *82*, 469–484. https://doi.org/10.1021/acs.jnatprod.9b00176.

Caron, G., Digiesi, V., Solaro, S., and Ermondi, G. (2020). Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. Drug Discov. Today *25*, 621–627. https://doi.org/10.1016/j.drudis.2020.01.012.

Chang, F.-Y., Ternei, M.A., Calle, P.Y., and Brady, S.F. (2013). Discovery and synthetic refactoring of tryptophan dimer gene clusters from the environment. J. Am. Chem. Soc. *135*, 17906–17912. https://doi.org/10.1021/ja408683p.

Chevrette, M.G., Gutiérrez-García, K., Selem-Mojica, N., Aguilar-Martínez, C., Yañez-Olvera, A., Ramos-Aboites, H.E., Hoskisson, P.A., and Barona-Gómez, F. (2020). Evolutionary dynamics of natural product biosynthesis in bacteria. Nat. Prod. Rep. *37*, 566–599. https://doi.org/10.1039/c9np00048h.

Chhabra, M. (2021). Biological therapeutic modalities. In Translational Biotechnology, (Elsevier), pp. 137–164.

Chopra, H., Baig, A.A., Gautam, R.K., and Kamal, M.A. (2022). Application of Artificial intelligence in Drug Discovery. Curr. Pharm. Des. *28*. https://doi.org/10.2174/1381612828666220608141049.

Culp, E.J., Sychantha, D., Hobson, C., Pawlowski, A.C., Prehna, G., and Wright, G.D. (2022). ClpP inhibitors are produced by a widespread family of bacterial gene clusters. Nat. Microbiol. *7*, 451–462. https://doi.org/10.1038/s41564-022-01073-4.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002). Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. U. S. A. *99*, 10494–10499. https://doi.org/10.1073/pnas.142680199.

Delahaye, C., and Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. PloS One *16*, e0257521. https://doi.org/10.1371/journal.pone.0257521.

Doak, B.C., Over, B., Giordanetto, F., and Kihlberg, J. (2014). Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chem. Biol. *21*, 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinforma. Oxf. Engl. *14*, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Frottin, F., Bienvenut, W.V., Bignon, J., Jacquet, E., Jacome, A.S.V., Van Dorsselaer, A., Cianferani, S., Carapito, C., Meinnel, T., and Giglione, C. (2016). MetAP1 and MetAP2 drive cell selectivity for a potent anti-cancer agent in synergy, by controlling glutathione redox state. Oncotarget *7*, 63306–63323. https://doi.org/10.18632/oncotarget.11216.

Fukuda, T., Arai, M., Yamaguchi, Y., Masuma, R., Tomoda, H., and Omura, S. (2004). New Beauvericins, Potentiators of Antifungal Miconazole Activity, Produced by Beauveria sp. FKI-1366: I. Taxonomy, Fermentation, Isolation and Biological Properties. J. Antibiot. (Tokyo) *57*, 110–116. https://doi.org/10.7164/antibiotics.57.110.

García-Ruiz, C., and Sarabia, F. (2014). Chemistry and Biology of Bengamides and Bengazoles, Bioactive Natural Products from Jaspis Sponges. Mar. Drugs *12*, 1580–1622. https://doi.org/10.3390/md12031580.

Gavriilidou, A., Kautsar, S.A., Zaburannyi, N., Krug, D., Müller, R., Medema, M.H., and Ziemert, N. (2022). Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat. Microbiol. *7*, 726–735. https://doi.org/10.1038/s41564-022-01110-2.

Gietz, R.D., and Schiestl, R.H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat. Protoc. *2*, 31–34. https://doi.org/10.1038/nprot.2007.13.

Gironda-Martínez, A., Donckele, E.J., Samain, F., and Neri, D. (2021). DNA-Encoded Chemical Libraries: A Comprehensive Review with Successful Stories and Future Challenges. ACS Pharmacol. Transl. Sci. *4*, 1265–1279. https://doi.org/10.1021/acsptsci.1c00118.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. *5*, R245-249. https://doi.org/10.1016/s1074-5521(98)90108-9.

Henrich, C.J., and Beutler, J.A. (2013). Matching the power of high throughput screening to the chemical diversity of natural products. Nat. Prod. Rep. *30*, 1284–1298. https://doi.org/10.1039/c3np70052f.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

Kale, A.J., McGlinchey, R.P., Lechner, A., and Moore, B.S. (2011). Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. ACS Chem. Biol. *6*, 1257–1264. https://doi.org/10.1021/cb2002544.

Katz, M., Hover, B.M., and Brady, S.F. (2016). Culture-independent discovery of natural products from soil metagenomes. J. Ind. Microbiol. Biotechnol. *43*, 129–141. https://doi.org/10.1007/s10295-015-1706-6.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., et al. (2019). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. gkz882. https://doi.org/10.1093/nar/gkz882.

Kellogg, J.J., Todd, D.A., Egan, J.M., Raja, H.A., Oberlies, N.H., Kvalheim, O.M., and Cech, N.B. (2016). Biochemometrics for Natural Products Research: Comparison of Data Analysis Approaches and Application to Identification of Bioactive Compounds. J. Nat. Prod. *79*, 376–386. https://doi.org/10.1021/acs.jnatprod.5b01014.

Lenci, E., Baldini, L., and Trabocchi, A. (2021). Diversity-oriented synthesis as a tool to expand the chemical space of DNA-encoded libraries. Bioorg. Med. Chem. *41*, 116218. https://doi.org/10.1016/j.bmc.2021.116218.

Li, F., Wang, Y., Li, D., Chen, Y., and Dou, Q.P. (2019). Are we seeing a resurgence in the use of natural products for new drug discovery? Expert Opin. Drug Discov. *14*, 417–420. https://doi.org/10.1080/17460441.2019.1582639.

Menzel, P., Ng, K.L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. *7*, 11257. https://doi.org/10.1038/ncomms11257.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. *16*, 60–68. https://doi.org/10.1038/s41589-019-0400-9.

Newman, D.J., and Cragg, G.M. (2020). Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J. Nat. Prod. *83*, 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

Nothias, L.-F., Nothias-Esposito, M., da Silva, R., Wang, M., Protsyuk, I., Zhang, Z., Sarvepalli, A., Leyssen, P., Touboul, D., Costa, J., et al. (2018). Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. J. Nat. Prod. *81*, 758–767. https://doi.org/10.1021/acs.jnatprod.7b00737.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834. https://doi.org/10.1101/gr.213959.116.

Owen, J.G., Charlop-Powers, Z., Smith, A.G., Ternei, M.A., Calle, P.Y., Reddy, B.V.B., Montiel, D., and Brady, S.F. (2015). Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. Proc. Natl. Acad. Sci. U. S. A. *112*, 4221–4226. https://doi.org/10.1073/pnas.1501124112.

Peek, J., Lilic, M., Montiel, D., Milshteyn, A., Woodworth, I., Biggins, J.B., Ternei, M.A., Calle, P.Y., Danziger, M., Warrier, T., et al. (2018). Rifamycin congeners kanglemycins are active against rifampicin-resistant bacteria via a distinct mechanism. Nat. Commun. *9*, 4147. https://doi.org/10.1038/s41467-018-06587-2.

Pitt, A., and Nims, Z. (2019). Peptide Display Technologies. Methods Mol. Biol. Clifton NJ *2001*, 285–298. https://doi.org/10.1007/978-1-4939-9504-2_13.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394. https://doi.org/10.1146/annurev.micro.57.030502.090759.

Salvador-Reyes, L.A., and Luesch, H. (2015). Biological targets and mechanisms of action of natural products from marine cyanobacteria. Nat. Prod. Rep. *32*, 478–503. https://doi.org/10.1039/c4np00104d.

Shao, L., Qu, X.-D., Jia, X.-Y., Zhao, Q.-F., Tian, Z.-H., Wang, M., Tang, G.-L., and Liu, W. (2006). Cloning and characterization of a bacterial iterative type I polyketide synthase gene encoding the 6-methylsalicyclic acid synthase. Biochem. Biophys. Res. Commun. *345*, 133–139. https://doi.org/10.1016/j.bbrc.2006.04.069.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

Stevenson, L.J., Owen, J.G., and Ackerley, D.F. (2019). Metagenome Driven Discovery of Nonribosomal Peptides. ACS Chem. Biol. *14*, 2115–2126. https://doi.org/10.1021/acschembio.9b00618.

Stevenson, L.J., Bracegirdle, J., Liu, L., Sharrock, A.V., Ackerley, D.F., Keyzers, R.A., and Owen, J.G. (2021). Metathramycin, a new bioactive aureolic acid discovered by heterologous expression of a metagenome derived biosynthetic pathway. RSC Chem. Biol. *2*, 556–567. https://doi.org/10.1039/d0cb00228c.

Stone, S., Newman, D.J., Colletti, S.L., and Tan, D.S. (2022). Cheminformatic analysis of natural product-based drugs and chemical probes. Nat. Prod. Rep. *39*, 20–32. https://doi.org/10.1039/D1NP00039J.

Stratton, C.F., Newman, D.J., and Tan, D.S. (2015). Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorg. Med. Chem. Lett. *25*, 4802–4807. https://doi.org/10.1016/j.bmcl.2015.07.014.

Tedersoo, L., Albertsen, M., Anslan, S., and Callahan, B. (2021). Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl. Environ. Microbiol. *87*, e0062621. https://doi.org/10.1128/AEM.00626-21.

Thomas, M., Boardman, A., Garcia-Ortegon, M., Yang, H., de Graaf, C., and Bender, A. (2022). Applications of Artificial Intelligence in Drug Design: Opportunities and Challenges. Methods Mol. Biol. Clifton NJ *2390*, 1–59. https://doi.org/10.1007/978-1-0716-1787-8_1.

Tomm, H.A., Ucciferri, L., and Ross, A.C. (2019). Advances in microbial culturing conditions to activate silent biosynthetic gene clusters for novel metabolite production. J. Ind. Microbiol. Biotechnol. *46*, 1381–1400. https://doi.org/10.1007/s10295-019-02198-y.

Tran, P.N., Yen, M.-R., Chiang, C.-Y., Lin, H.-C., and Chen, P.-Y. (2019). Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl. Microbiol. Biotechnol. *103*, 3277–3287. https://doi.org/10.1007/s00253-019-09708-z.

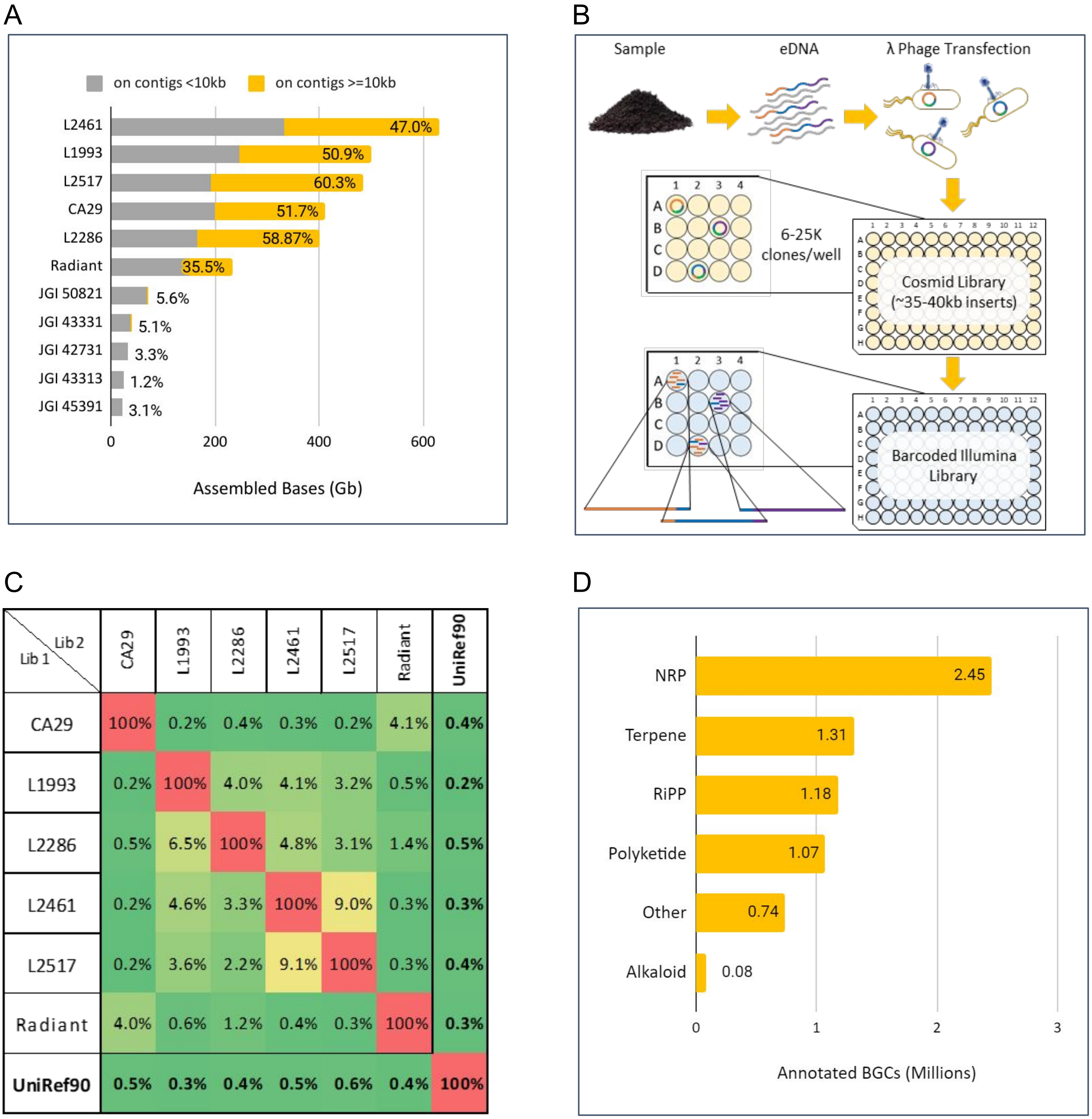
Wang, Z., Forelli, N., Hernandez, Y., Ternei, M., and Brady, S.F. (2022). Lapcin, a potent dual topoisomerase I/II inhibitor discovered by soil metagenome guided total chemical synthesis. Nat. Commun. *13*, 842. https://doi.org/10.1038/s41467-022-28292-x.

White, K.N., Tenney, K., and Crews, P. (2017). The Bengamides: A Mini-Review of Natural Sources, Analogues, Biological Properties, Biosynthetic Origins, and Future Prospects. J. Nat. Prod. *80*, 740–755. https://doi.org/10.1021/acs.jnatprod.6b00970.

Wilson, B.A.P., Thornburg, C.C., Henrich, C.J., Grkovic, T., and O'Keefe, B.R. (2020). Creating and screening natural product libraries. Nat. Prod. Rep. *37*, 893–918. https://doi.org/10.1039/C9NP00068B.

Xu, G., Zhang, L., Liu, X., Guan, F., Xu, Y., Yue, H., Huang, J.-Q., Chen, J., Wu, N., and Tian, J. (2022). Combined assembly of long and short sequencing reads improve the efficiency of exploring the soil metagenome. BMC Genomics *23*, 37. https://doi.org/10.1186/s12864-021-08260-3.

FIGURE 1



Lib 2 / Lib 1	CA29	L1993	L2286	L2461	L2517	Radiant	**UniRef90**
CA29	100%	0.2%	0.4%	0.3%	0.2%	4.1%	**0.4%**
L1993	0.2%	100%	4.0%	4.1%	3.2%	0.5%	**0.2%**
L2286	0.5%	6.5%	100%	4.8%	3.1%	1.4%	**0.5%**
L2461	0.2%	4.6%	3.3%	100%	9.0%	0.3%	**0.3%**
L2517	0.2%	3.6%	2.2%	9.1%	100%	0.3%	**0.4%**
Radiant	4.0%	0.6%	1.2%	0.4%	0.3%	100%	**0.3%**
UniRef90	**0.5%**	**0.3%**	**0.4%**	**0.5%**	**0.6%**	**0.4%**	**100%**

Figure 1. Metagenomic libraries contain vast, orthogonal diversity. A) Total assembly lengths for the 6 soil metagenomic libraries described in this paper and the 5 largest soil metagenomics datasets from JGI. Percentages indicate the fraction of assembled sequence contained on contigs >10kb. Libraries described in this paper contain, on average, 11.6X more total assembled sequence and 140X assembled sequence on contigs >10kb. **B)** Schematic diagram for generating reduced-complexity metagenomic libraries. eDNA cosmid library containing ~35-40kb of eDNA/cosmid is generated via λ phage transfection as sub-pools containing 6-25K clones. Barcoded sequencing libraries are generated and sequence is assembled by sub-pool to generate long assemblies. **C)** Pairwise comparison of overlap between protein coding sequences in metagenomic libraries clustered at 90% aa identity as well as UniRef90 (in bold). **D)** NRP, terpene, RiPP, and polyketide clusters are the most common major biosynthetic gene classes represented in the metagenomic libraries (non-deduplicated data).

FIGURE 2

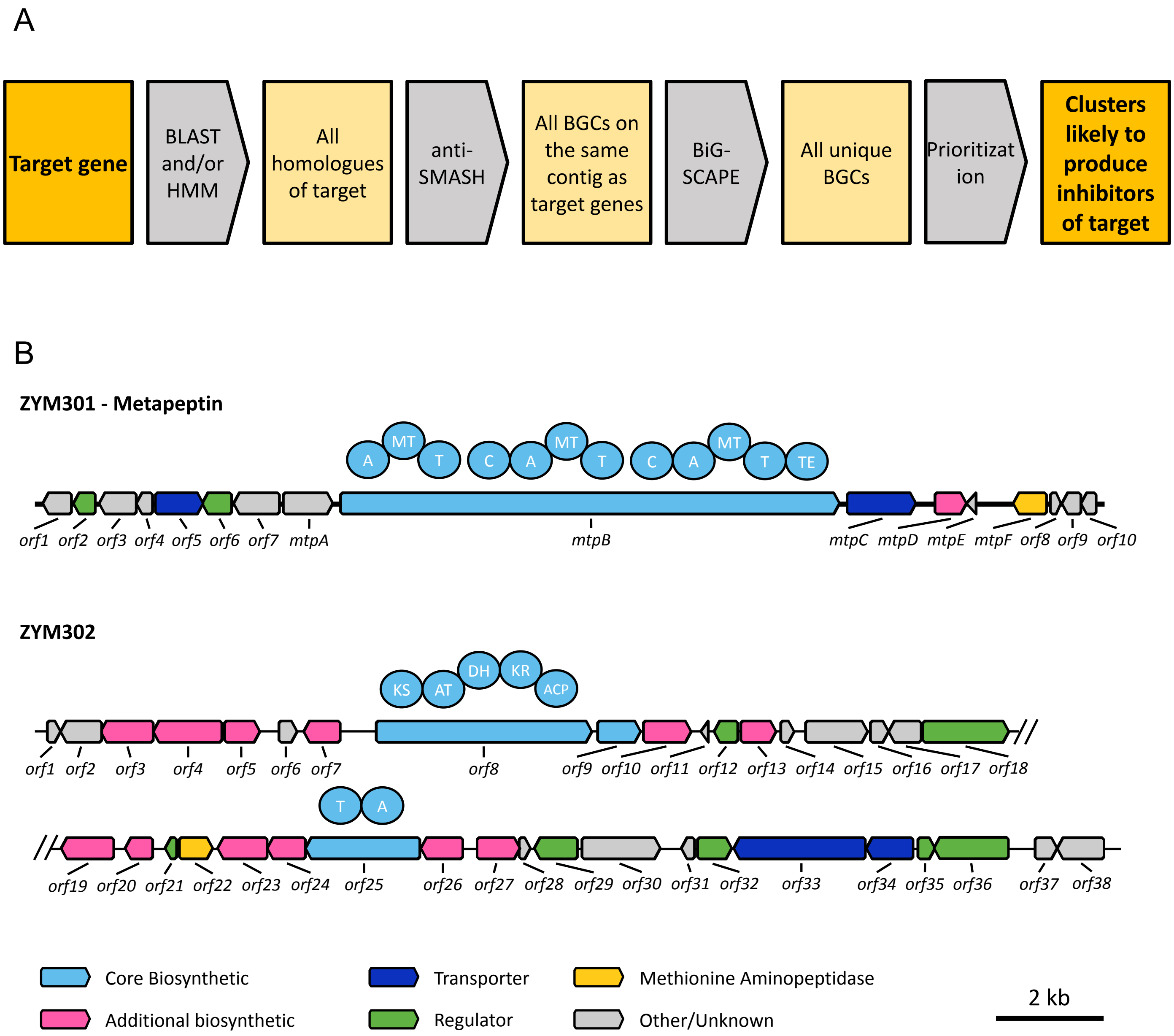

A
Target gene
BLAST and/or HMM
All homologues of target
anti-SMASH
All BGCs on the same contig as target genes
BiG-SCAPE
All unique BGCs
Prioritization
Clusters likely to produce inhibitors of target
B
ZYM301 - Metapeptin
A
MT
T
C
A
MT
T
C
A
MT
T
TE
orf1 orf2 orf3 orf4 orf5 orf6 orf7 mtpA
mtpB
mtpC mtpD mtpE mtpF orf8 orf9 orf10
ZYM302
KS
AT
DH
KR
ACP
orf1 orf2 orf3 orf4 orf5 orf6 orf7
orf8
orf9 orf10 orf11 orf12 orf13 orf14 orf15 orf16 orf17 orf18
T
A
orf19 orf20 orf21 orf22 orf23 orf24 orf25
orf26 orf27 orf28 orf29 orf30 orf31 orf32 orf33 orf34 orf35 orf36 orf37 orf38
Core Biosynthetic
Additional biosynthetic
Transporter
Regulator
Methionine Aminopeptidase
Other/Unknown
2 kb

Figure 2. Identification of BGCs containing putative MetAP resistance genes. A) A flowchart showing the generalized workflow for identifying clusters that will produce inhibitors of target genes of interest. **B)** ZYM301 (metapeptin) and ZYM302 gene clusters. ZYM302 is displayed over two lines for clarity, split in the 65 bp intergenic region between *orf18* and *orf19*. Genes are colored based on the predicted function assigned by gene identity and antiSMASH (Tables S2-S3), with the methionine aminopeptidase target genes called out in yellow. The domain organization of PKS and NRPS-like genes are shown above the genes, labeled as adenylation (A), methyltransferase (MT), thiolation (T), condensation (C), thioesterase (TE), ketosynthase (KS), acyltransferase (AT), dehydratase (DH), ketoreductase (KR), or acyl carrier protein (ACP).

FIGURE 3

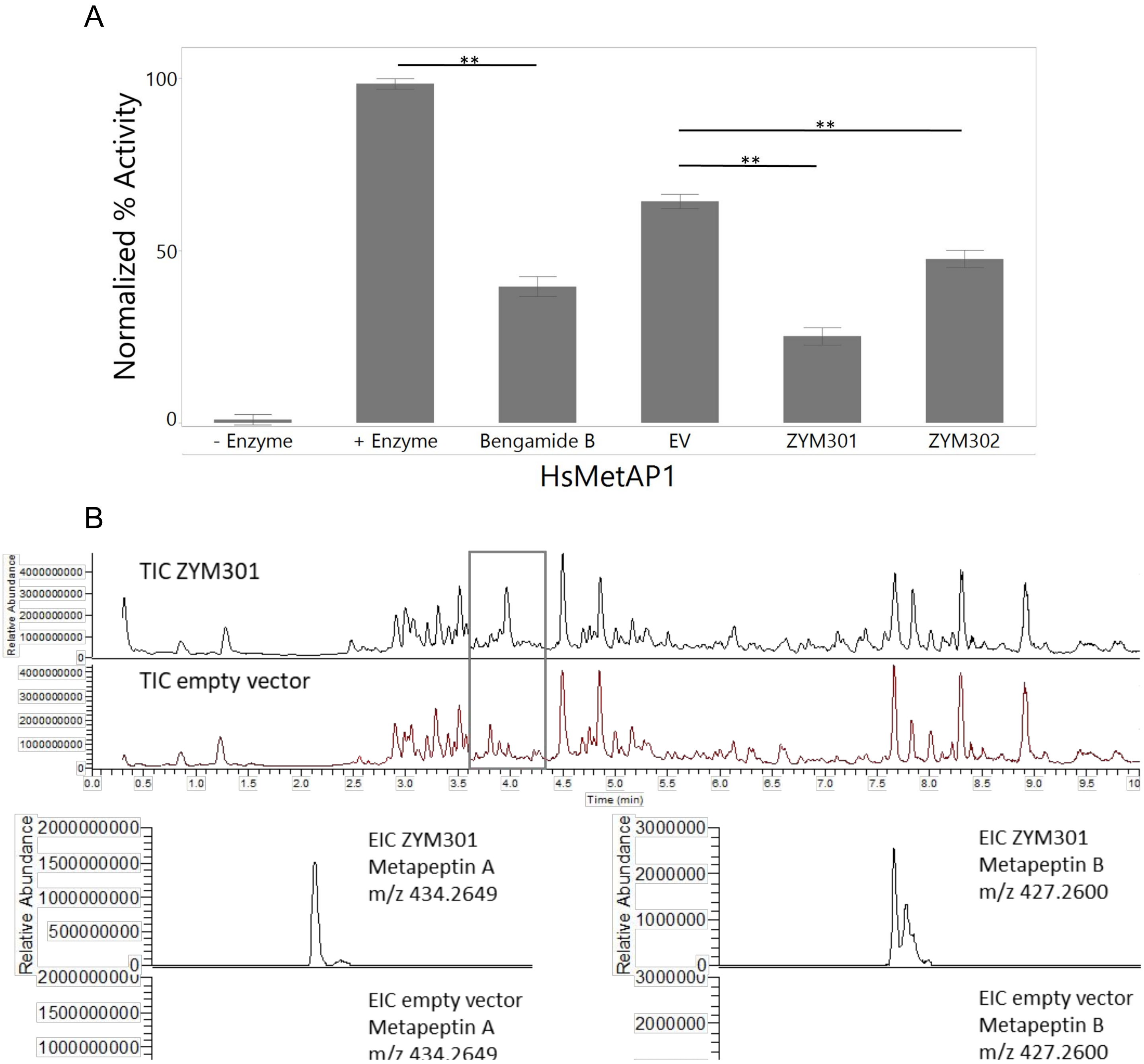
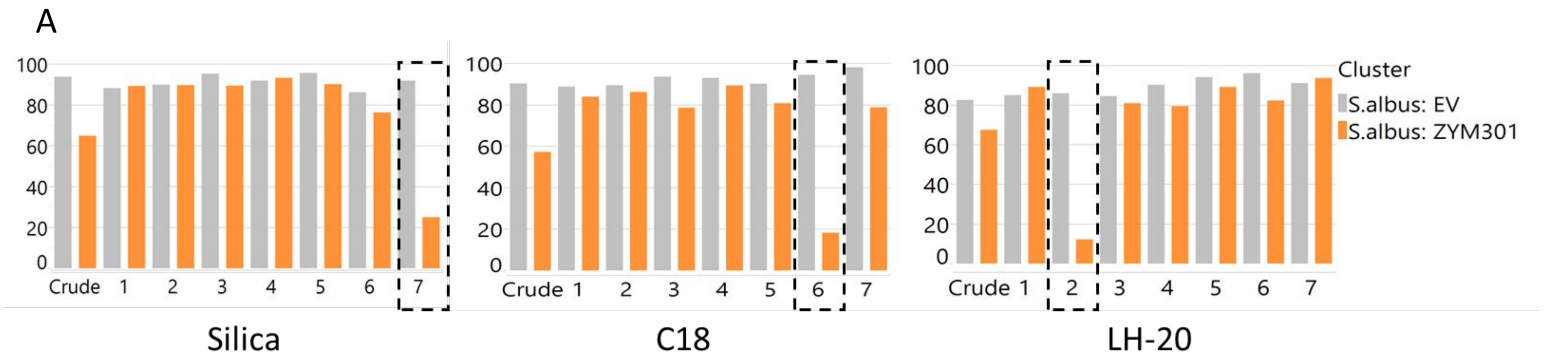


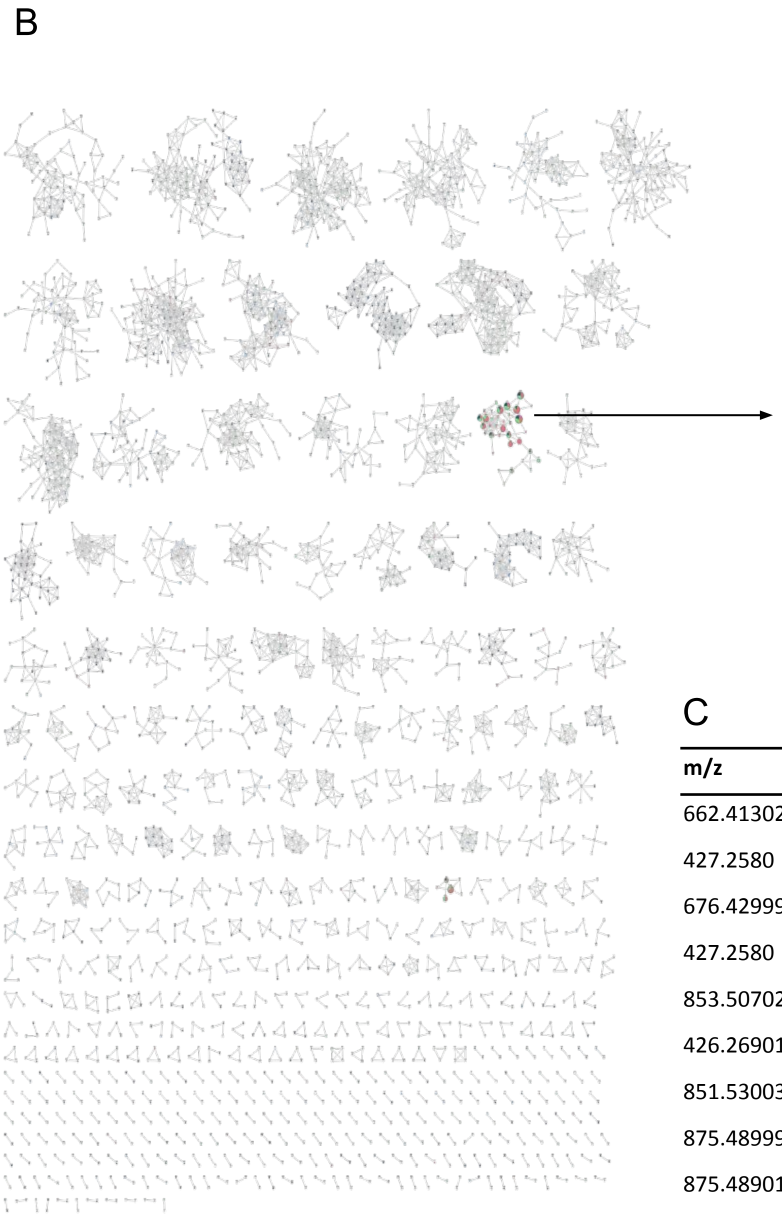
Figure 3. Crude extracts from *S.albus:ZYM30*1 and *S.albus:ZYM302* inhibit HsMetAP1. A) Crude extracts (0.5mg/ml) from *S. albus*:*ZYM301*, *S. albus*:*ZYM302*, *S. albus* empty vector control (EV) and Bengamide B control (100 μM) were incubated with HsMetAP1 and tripeptide substrate for 20 minutes and assayed for methionine peptidase activity. Controls included a DMSO vehicle control (+enzyme) and reactions that did not contain hsMetAP1 (-enzyme). Observed Inhibitory activity were separately validated via LC-MS. Error bars represent the standard deviation of the mean, n=3. **Statistical significance assessed via Dunnett's, p-value (<0.001). **B)** Total Ion Chromatograms of a representative *S. albus:ZYM301* and empty vector control sample highlighting the region where novel features were detected (above), and Extracted Ion Chromatograms of Metapeptin A and B (below).

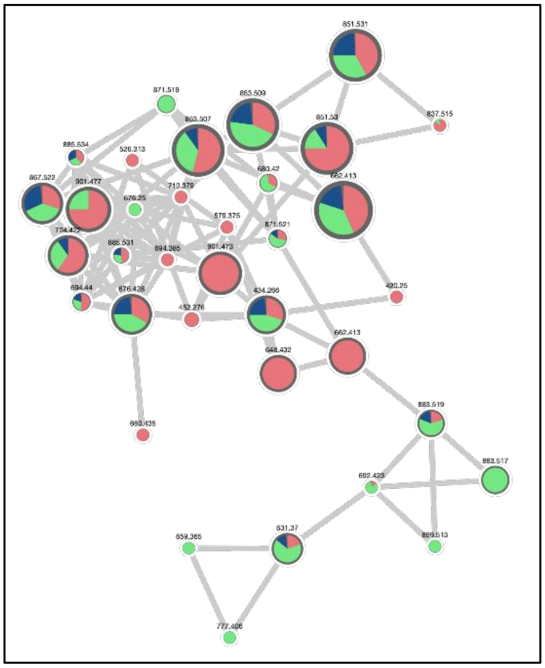
Table 1. Novel compounds identified by untargeted LC-MS/MS analysis.

Metapeptin	Observed m/z	Calculated Molecular Weight	Retention time
	901.4738 $[M+H]^{+1}$, 451.2436 $[M+H]^{+2}$	900.466	3.58
C	839.4909 $[M+H]^{+1}$, 420.2483 $[M+H]^{+2}$	838.4831	3.92
B	853.5067 $[M+H]^{+1}$, 427.2600 $[M+H]^{+2}$	852.4989	3.95
	889.5046 $[M+H]^{+1}$, 445.2537 $[M+H]^{+2}$	888.4963	3.97
A	867.5217 $[M+H]^{+1}$, 434.2649 $[M+H]^{+2}$	866.5139	3.98
	885.5313 $[M+H]^{+1}$, 443.2696 $[M+H]^{+2}$	884.5235	4.01
	855.4876 $[M+H]^{+1}$, 428.2474 $[M+H]^{+2}$	854.4798	4.06
E	883.5175 $[M+H]^{+1}$, 442.2617 $[M+H]^{+2}$	882.5097	4.09
	899.5128 $[M+H]^{+1}$, 450.2625 $[M+H]^{+2}$	898.5049	4.16
D	851.5274 $[M+H]^{+1}$, 426.0726 $[M+H]^{+2}$	850.5195	4.20
	869.5346 $[M+H]^{+1}$, 435.2707 $[M+H]^{+2}$	868.5268	4.20
	835.5323 $[M+H]^{+1}$, 418.2700 $[M+H]^{+2}$	834.5244	4.40

FIGURE 4





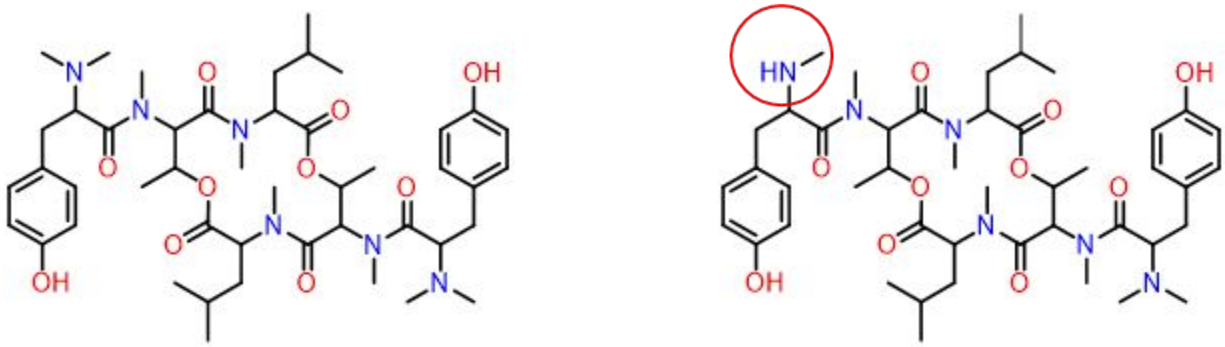


C

m/z	Retention time	Compound	selectivity_ratio
662.41302	3.73	Metapeptin B	3.49
427.2580	3.71	Metapeptin B	3.15
676.42999	4.14	Metapeptin B	3.14
427.2580	3.7	Metapeptin B	3.06
853.50702	3.76	Metapeptin B	3.05
426.26901	4.13	Metapeptin B	3.04
851.53003	4.22	Metapeptin D	3.04
875.48999	3.73	Metapeptin B	2.88
875.48901	3.79	Metapeptin B	2.88
451.2420	3.27	MW 900	2.5

Figure 4. Biochemometric strategy indicate metapeptin B is largely responsible for observed bioactivity. A) Bioactivity assays of the 21 fractions generated using C18, Silica and LH20 size exclusion chromatography columns. Significant inhibition by *S.albus:ZYM301* fractions are outlined in a dotted box. **B)** A molecular network of all the features detected with MS2 data across 21 fractions, and a magnified view on the connected component that contains the compounds with the highest selectivity ratio. The node size indicates the selectivity ratio and the pie chart indicates the fractionation method it was observed in (green: Silica, red: C18, blue: LH20). **C)** Top 10 differentially expressed features sorted by selectivity ratio and the compounds with which they are associated.

FIGURE 5



Metapeptin A

Metapeptin B

Figure 5. Structures of metapeptin A and B. The red circle highlights the single methyl difference between the two molecules.

FIGURE 6

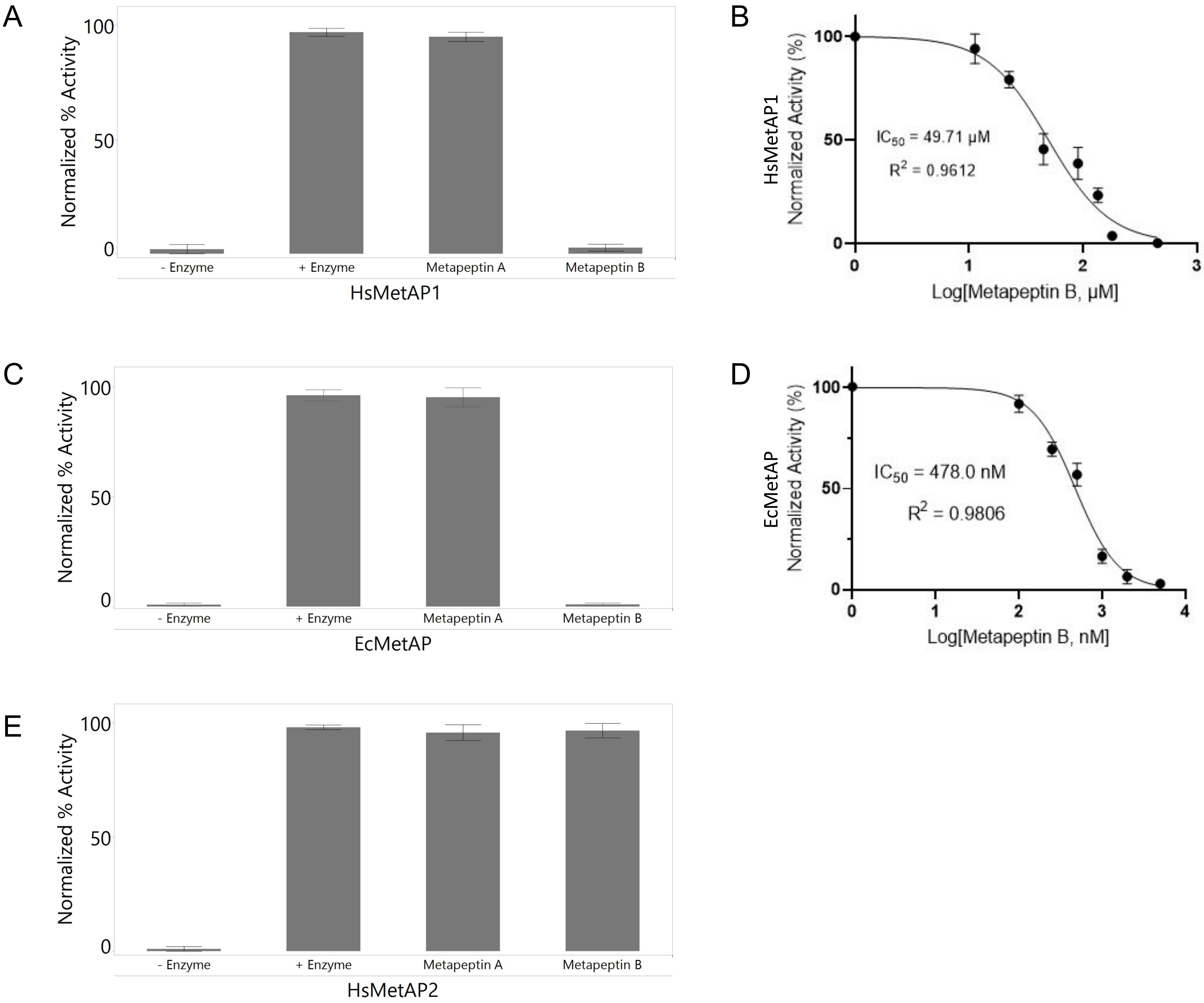
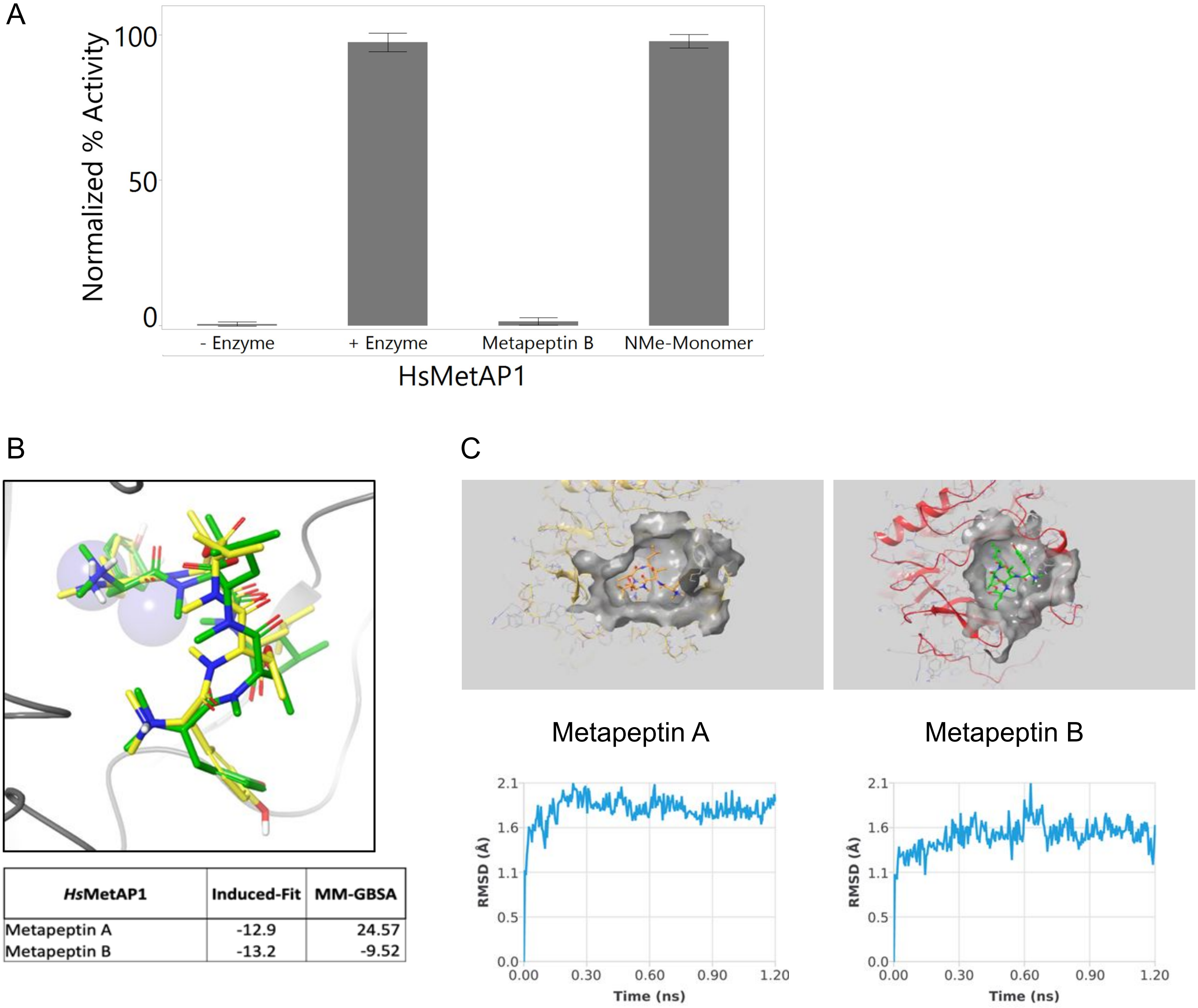


Figure 6: Metapeptin B inhibits HsMetAP1 and EcMetAP, but not HsMetAP2. Metapeptin A and metapeptin B (100 μM) were incubated with **A)** HsMetAP1, **C)** *E. coli* MetAP (EcMetAP), **E)** HsMetAP2 and tripeptide substrate for 20 minutes and assayed for methionine peptidase activity. Controls included a DMSO vehicle control (+enzyme) and reactions that did not contain enzyme (-enzyme). Error bars represent the standard deviation of the mean, n=3. **B)** Dose response curve of Metapeptin B (0.5 μM – 500 μM) against hsMetAP1. Non-linear regression (variable slope) analysis used to fit the curve. Error bars represent the standard deviation of the mean, n=9. Statistical significance assessed via ANOVA, p-value (<0.001). **D)** Dose response curve of Metapeptin B (100 nM – 5000 nM) against EcMetAP1. Non-linear regression (variable slope) analysis used to fit the curve. Error bars represent the standard deviation of the mean, n=9. Statistical significance assessed via ANOVA, p-value (<0.001).

FIGURE 7



*Hs*MetAP1	Induced-Fit	MM-GBSA
Metapeptin A	-12.9	24.57
Metapeptin B	-13.2	-9.52

Figure 7. SAR demonstrate the importance of the cyclization and asymmetric methylation of metapeptin B for its interaction with HsMetAP1. A) Metapeptin B and the NMe-Monomer linear tripeptide (100μM), were incubated with HsMetAP1 and tripeptide substrate for 20 minutes and assayed for methionine peptidase activity. Controls included a DMSO vehicle control (+enzyme) and reactions that did not contain enzyme (-enzyme). No inhibition was observed with NMe-Monomer, highlighting the importance of cyclization of metapeptin B for bioactivity. **B-C)** Three different molecular simulations (docking, MMGBSA, molecular dynamics) support stronger binding for metapeptin B. **B)** Induced fit docking shows conformational differences resulting in better docking score and MMGBSA energies (more negative is more favorable) including a large DDGs solvation change. **C)** Lower RMSD for distance to binding pocket residues for metapeptin B indicates more stable binding.

FIGURE 8

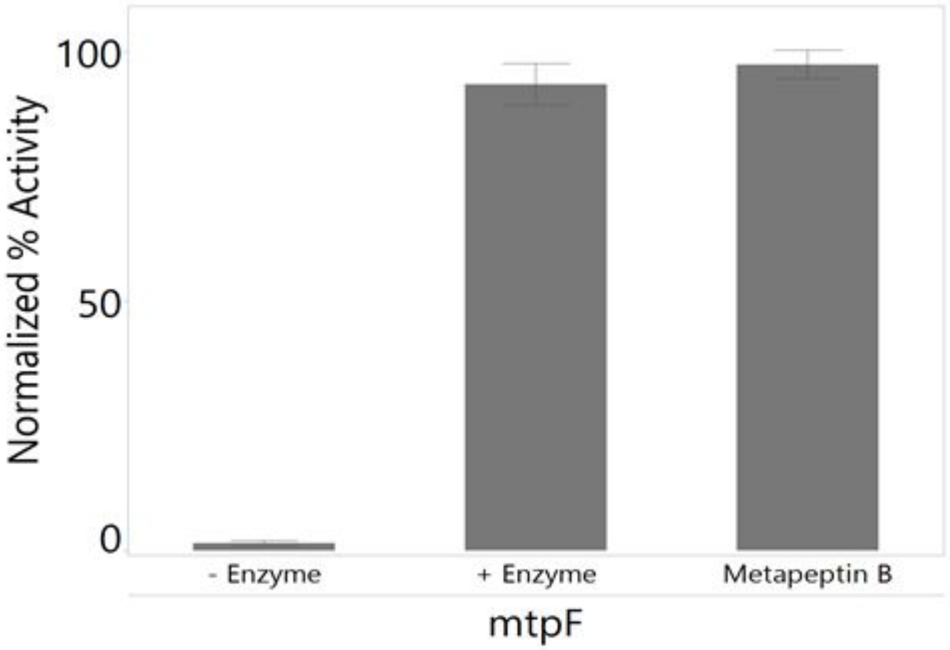


Figure 8: Methionine peptidase encoded by the putative resistance gene in metapeptin cluster (*mtpF*) is not inhibited by metapeptinB. Metapeptin B was incubated with the methionine peptidase encoded by *mtpF* and tripeptide substrate for 20 minutes and assayed for methionine peptidase activity. Controls included a DMSO vehicle control (+enzyme) and reactions that did not contain enzyme (-enzyme). Error bars represent the standard deviation of the mean, n=3.