Computational Metabolomics for Bioactivity-Driven Classification in Natural Products Drug
Discovery and Opioid Detection

By

Nathaniel Jacob Brittin

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Pharmaceutical Sciences)

at the

UNIVERSITY OF WISCONSIN-MADISON

2025

Date of final oral examination: 05/15/25

The dissertation is approved by the following members of the Final Oral Committee:
    Tim S. Bugni, Professor, Pharmaceutical Sciences
    Lindsay R. Kalan, Assistant Professor, Medical Microbiology and Immunology
    Jason C. Kwan, Associate Professor, Pharmaceutical Sciences
    Ting Fu, Assistant Professor, Pharmaceutical Sciences

Dedication

This thesis is dedicated to all those in my life who have supported me and continue to do so at every stage of my journey. To my parents, Jill and Stephen Brittin, for raising and supporting me to be an open book, always ready to absorb every new scrap of information to fill the pages of my mind. It has shaped me into the person I am today, and I hope you are as proud of me as I am of myself. To my siblings, Olivia, Liam, and Alan, who are all growing up so fast and becoming incredible adults all by themselves, you all push me to be an example for you and to be a brother you can be proud of. To my partner in life, Carolyn, who, through thick and thin, has stood by me for twelve and a half years. I couldn't have achieved any of this without you—without your support, I wouldn't have moved across the country, bought and sold our first home, or adopted two wonderful cats. You make all this worth doing and bring an order to my life that, while I may complain about it occasionally, supports both of us to accomplish our goals. To my grandparents, Richard and Libby Barwis, who are wonderfully supportive and who are the inspiration for many of my artistic ventures. To all my other family, aunts, uncles, and cousins, thank you for everything! To my friends, who make my life more than just work and home—whether it's camping trips, biking adventures, or wild game nights, my life is richer because of all of you. Finally, I dedicate this thesis to myself. I have persevered through years of hard work and personal growth to produce the work presented here. I've taken every opportunity and challenge as fuel for growth, both as a person who constantly seeks knowledge and as a scientist eager to contribute to solving the world's challenges.

Abstract

The rise of multidrug-resistant (MDR) pathogens and the proliferation of synthetic opioids pose urgent threats to global health, demanding innovative approaches in both drug discovery and forensic analysis. This dissertation presents a multidisciplinary framework that leverages high-resolution mass spectrometry, yeast chemical genomics (YCG), and machine learning (ML) to address these dual challenges across two different but connected domains: natural product discovery and clinical toxicology (Chapter 1).

To accelerate the identification of structurally novel and mechanistically distinct antifungal agents, we developed a high-throughput screening pipeline that integrates LC-MS/MS-based metabolomics with YCG profiling (Chapter 2). Nearly 40,000 bacterial extract fractions from diverse microbiomes were screened for antifungal activity, with hits prioritized based on both chemical-genetic interaction signatures and computational metabolomics identifications. This dual-platform approach enabled functional dereplication by linking mass spectral features to both known drug classes and unexplored bioactive compounds, streamlining the prioritization of lead scaffolds for further development.

To further enhance the dereplication and prioritization of small molecules, we trained ML classifiers on *in-silico* MS/MS fingerprints to assign natural products to 21 pharmacophore-defined drug classes (Chapter 3). These classifiers achieved >93% accuracy in multiclass tasks and consistently outperformed state-of-the-art tools (e.g., CANOPUS) across *in-silico*, GNPS, and experimental datasets. When applied to microbial extracts, the models enabled rapid classification of bioactivity, even in the absence of direct structural matches, expanding the accessible chemical and functional space for antimicrobial discovery.

The second major focus of this dissertation is the development of ML-based models for

opioid detection from clinical LC-MS/MS data (Chapter 4). Classifiers for morphinan, fentanyl, and nitazene opioids were trained on simulated molecular fingerprints and validated using GNPS spectra, CDC FAS Kit reference materials, and anonymized clinical blood and urine samples. Morphinan and fentanyl models demonstrated exceptional performance, increasing compound detection and sample coverage by 90–600% and 33–400%, respectively, compared to traditional spectral library searches. Although the nitazene classifier performed well in controlled settings, clinical application highlighted the need for improved coverage and training data.

Collectively, these studies illustrate how integrating high-resolution metabolomics with machine learning and functional genomics can transform both small molecule discovery and public health surveillance, enabling scalable, structure-informed detection of bioactive natural products and emerging synthetic opioids.

Acknowledgments

I would like to thank my advisor Dr. Tim Bugni for being a great mentor to me. Over my five years here at UW – Madison, he has consistently provided opportunities for me to grow as a scientist and a person. He gave me guidance when needed, he gave me space to explore my ideas and provided a space that encouraged growth and achievement. Thank you for being my mentor, I couldn't imagine doing a Ph.D. without you as my advisor and mentor.

I would like to thank the members of my thesis committee for consistently providing me with feedback and encouragement throughout my time here. Your insight and questions have continued to improve my understanding of science and the quality of my work.

To all my Bugni lab members, past and present, thank you for all being wonderful people who always provided a helping hand or a listening ear. Bailey, Chris, Josephine, Imraan, Doug, Scott, Shukria, Tae Hyun Lee, Changyeol, Shaurya, Chris T., Fan, Mitasree, Qihao, Dave, and Spence.

To everyone I haven't listed or are reading this, thank you!

Table of Contents

Table of Figures

Chapter 1: Introduction

1.1 INTRODUCTION

The global rise in multidrug-resistant (MDR) pathogens and the opioid crisis represent two of the most pressing public health challenges of the 21st century. The persistence and spread of MDR bacteria and fungi, combined with the emergence of novel synthetic opioids, continue to strain healthcare systems and compromise public safety [1-5]. This dissertation presents a multidisciplinary approach to address these crises, focusing on three research areas: (1) the discovery of novel antifungal agents using liquid chromatography-tandem mass spectrometry (LC-MS/MS) and yeast chemical genomics, (2) the machine learning-based classification of natural products (NPs) for bioactivity prediction, and (3) the development of a machine learning-driven framework for opioid detection in clinical metabolomics data. Through these studies, we aim to advance our ability to combat MDR pathogens and enhance opioid detection strategies in clinical and forensic settings.

1.1.1 The Challenge of Multidrug-Resistant Pathogens

MDR bacterial and fungal pathogens present a critical challenge to global health, especially in the context of rising antibiotic resistance [6,7]. Antibiotics, once considered a miracle of modern medicine, have lost their efficacy as bacteria and fungi evolve rapidly to resist commonly used drugs. The World Health Organization (WHO) has recognized MDR pathogens as a leading threat to global health, with drug-resistant bacterial and fungal infections currently responsible for hundreds of thousands of deaths each year and a growing number of illnesses [8]. A recent report estimates that there are about 6.5 million invasive fungal infections each year with nearly 4 million deaths [9].

In the case of bacterial pathogens, many species have emerged as particularly resistant to commonly prescribed antibiotics. The spread of these resistant pathogens has made routine surgeries and treatments, such as cancer chemotherapy or organ transplantation, increasingly dangerous, as infections that were once easily treatable now require more aggressive and toxic therapies.

Fungal pathogens that are particularly concerning include *Candida auris*, a pan-resistant fungal pathogen which is immune to all antifungal classes, and *Aspergillus fumigatus*, an opportunistic pathogen that has resistance to first-line antifungal drugs such as triazoles [10-13]. For immunocompromised patients, such as those undergoing chemotherapy, organ transplantation, or living with HIV/AIDS, these infections pose significant mortality risks—with rates ranging from 30 to 95% mortality and hospital stays extending from 46 to 140 days [2,8]. The increasing prevalence of MDR fungi further complicates the treatment landscape, as these pathogens have developed resistance to the four currently available antifungal drug families [14,15].

Addressing the threat of MDR pathogens requires innovative approaches for discovering novel antimicrobial agents. This dissertation contributes to these efforts by integrating LC-MS/MS, a powerful tool for high-resolution chemical analysis, with chemical genomics or machine learning (ML) to uncover new natural products with potential antimicrobial activity [16].

1.1.2 *Novel Antifungal Discovery: LC-MS/MS and Chemical Genomics*

The need for novel antifungal agents is urgent, as existing classes of antifungals are rendered ineffective by resistant pathogens [14,15]. Conventional antifungal therapies often have limited mechanisms of action and are unable to keep pace with the rapid mutation and adaptation of pathogens. As a result, there is a critical need to explore new sources for antifungal compounds, particularly from natural products (NPs), which have historically been a rich source of bioactive

molecules [17-19].

Natural product discovery, however, is hindered by the challenge of dereplication, which involves distinguishing novel compounds from those already known. To overcome these barriers, we integrate LC-MS/MS with yeast chemical genomics (YCG) to discover novel antifungals. LC-MS/MS allows for rapid and detailed analysis of complex mixtures, enabling the characterization of metabolites with high sensitivity [16]. Chemical genomics, on the other hand, provides functional insights by linking the dose response of an antifungal agent to impacted areas of fungal biology, helping to pinpoint the likely mechanisms of action in a high-throughput manner.

Our approach leverages bacterial strains from underexplored niches, such as marine environments and insect microbiomes, which are known to harbor a wealth of unique metabolites. By combining LC-MS/MS with YCG, we can rapidly screen thousands of natural product extracts for antifungal activity, prioritize bioactive bacterial extracts containing no known antifungals, and leverage functional genomics to focus on promising new mechanisms. This integrated pipeline not only works to enhance the efficiency of antifungal discovery but also prioritizes compounds that exhibit novel mechanisms of action, thus offering a strategic approach to combating MDR fungal infections.

1.1.3 *Machine Learning for Bioactivity-Driven Classification of Natural Products*

Despite the promise of NPs as a source of novel therapeutics, the discovery process is often slow and inefficient due to the vast chemical and structural diversity [17-21]. Recent methods for compound classification based on chemical classification hierarchy or biosynthetic pathways are insufficient for identifying bioactive compounds in complex mixtures, particularly when structural features outside a compound's pharmacophore can impact chemical and biosynthetic classifications [22,23].

Machine learning has emerged as a transformative tool in this domain, enabling the analysis of complex datasets such as LC-MS/MS spectra to uncover underlying patterns in bioactivity [24]. By employing ML algorithms to analyze molecular fingerprints (MFPs) derived from *in-silico* fragmentation spectra, we can develop predictive models that classify natural products according to their pharmacophore, defined as the structural features essential for biological activity. This pharmacophore-based approach offers a more biologically relevant method for NP classification, focusing on the chemical features directly linked to bioactivity rather than broad structural or biosynthetic classifications [25].

In this dissertation, we apply ML techniques to classify 21 distinct bioactive drug classes, enabling rapid prioritization of natural products for further testing. We use *in-silico* generated MS/MS spectra and molecular fingerprints as training data, generated through SIRIUS 5, to expand the available chemical space for our models, thereby overcoming the limitation of finite experimental data [26-29]. By focusing on bioactivity rather than structural similarity, our ML framework provides a more robust and efficient strategy for identifying known bioactive drugs, thus accelerating the dereplication and prioritization of novel antifungals, antibiotics, and other bioactive natural products.

1.1.4 Addressing the Opioid Crisis with Machine Learning-Driven Detection

The opioid crisis has reached unprecedented levels, with opioid overdose deaths continuing to rise globally [30-34]. Traditional detection methods, such as immunoassays and gas chromatography-mass spectrometry (GC-MS), struggle to keep pace with the proliferation of known and novel synthetic opioids (NSO), including fentanyl analogs and nitazenes, which are not always detectable with the standard workflows [35-38]. As illicit drug manufacturers modify the chemical structures of opioids to evade legal restrictions, detection methods, or discover new

compounds, there is a critical need for innovative analytical approaches to improve opioid detection and characterization [36,39-41].

Machine learning offers a promising solution to this problem. The approach used for classification of antimicrobial natural products was highly successful; therefore, extension of those methods to identify drugs of abuse, such as opioids, was logical. By training ML models on an expansive set of *in-silico* generated mass spectral data combined with high-resolution experimental LC-MS/MS data, we developed a system capable of detecting both known and unknown opioids based on shared structural information. Our approach leverages *in-silico* fragmentation data and computationally generated MFPs, which allow for the identification of opioids in clinical samples even when the precise molecular structure is not known. This ability to generalize across different opioid classes, such as morphinan, fentanyl, and nitazenes, enhances the sensitivity and specificity of opioid detection, providing a more scalable and adaptable solution for opioid detection.

1.1.5 Summary

The research presented in this dissertation contributes to the urgent need for novel antimicrobial agents and improved opioid detection methods. By integrating high-resolution LC-MS/MS with chemical genomics and machine learning, we offer a comprehensive strategy for overcoming the limitations of traditional approaches in drug discovery and opioid detection. The novel dereplication pipeline, which incorporates LC-MS/MS and YCG, accelerates the discovery of antifungal agents, while our ML-driven framework for bioactivity classification allows for efficient identification of known bioactive pharmacophores within natural products. Furthermore, the opioid detection system we have developed represents a significant advancement in the ability to monitor and respond to the novel synthetic opioid crisis, providing a scalable and data-driven solution to track and identify novel synthetic opioids.

1.2   REFERENCES

(1)   Fisher, M.C.; Alastruey-Izquierdo, A.; Berman, J.; Bicanic, T.; Bignell, E.M.; Bowyer, P.; Bromley, M.; Brüggemann, R.; Garber, G.; Cornely, O.A.; et al. Tackling the Emerging Threat of Antifungal Resistance to Human Health. Nat. Rev. Microbiol. 2022, 20, 557–571.

(2)   Brown, G.D.; Denning, D.W.; Gow, N.A.R.; Levitz, S.M.; Netea, M.G.; White, T.C. Hidden Killers: Human Fungal Infections. Sci. Transl. Med. 2012, 4, 165rv13.

(3)   Lee, Y.; Puumala, E.; Robbins, N.; Cowen, L.E. Antifungal Drug Resistance: Molecular Mechanisms in Candida albicans and Beyond. Chem. Rev. 2021, 121, 3390–3411.

(4)   Gnat, S.; Łagowski, D.; Nowakiewicz, A.; Dyląg, M. A Global View on Fungal Infections in Humans and Animals: Opportunistic Infections and Microsporidioses. J. Appl. Microbiol. 2021, 131, 2095–2113.

(5)   O'Donnell J, Gladden RM, Mattson CL, Hunter CT, Davis NL. Vital Signs: Characteristics of Drug Overdose Deaths Involving Opioids and Stimulants - 24 States and the District of Columbia, January-June 2019. MMWR Morb Mortal Wkly Rep. 2020 Sep 4;69(35):1189-1197.

(6)   Lamoth F., Lockhart S.R., Berkow E.L., Calandra T. Changes in the Epidemiological Landscape of Invasive Candidiasis. J. Antimicrob. Chemother. 2018;73:i4–i13.

(7)   Rabaan A.A., Sulaiman T., Al-Ahmed S.H., Buhaliqah Z.A., Buhaliqah A.A., AlYuosof B., Alfaresi M., Al Fares M.A., Alwarthan S., Alkathlan M.S., et al. Potential Strategies to Control the Risk of Antifungal Resistance in Humans: A Comprehensive Review. Antibiotics. 2023;12:608.

(8)   WHO . WHO Fungal Priority Pathogens List to Guide Research, Development and Public Health. World Health Organization; Geneva, Switzerland: 2022.

(9)   Denning DW. Global incidence and mortality of severe fungal disease. Lancet Infect Dis. 2024 Jul;24(7):e428-e438. doi: 10.1016/S1473-3099(23)00692-8.

(10)   Chowdhary A, Jain K, Chauhan N. Candida auris Genetics and Emergence. Annu Rev Microbiol. 2023 Sep 15;77:583-602.

(11)   Pallotta F, Viale P, Barchiesi F. Candida auris: the new fungal threat. Infez Med. 2023 Sep 1;31(3):323-328.

(12)   Earle K, Valero C, Conn DP, Vere G, Cook PC, Bromley MJ, Bowyer P, Gago S. Pathogenicity and virulence of Aspergillus fumigatus. Virulence. 2023 Dec;14(1):2172264.

(13)   Dladla M, Gyzenhout M, Marias G, Ghosh S. Azole resistance in Aspergillus fumigatus-

comprehensive review. Arch Microbiol. 2024 Jun 15;206(7):305.

(14)   Lee Y, Puumala E, Robbins N, Cowen LE. Antifungal Drug Resistance: Molecular Mechanisms in Candida albicans and Beyond. Chem Rev. 2021 Mar 24;121(6):3390-3411.

(15)   Perlin DS, Rautemaa-Richardson R, Alastruey-Izquierdo A. The global problem of antifungal resistance: prevalence, mechanisms, and management. Lancet Infect Dis. 2017 Dec;17(12):e383-e392.

(16)   Brittin NJ, Aceti DJ, Braun DR, Anderson JM, Ericksen SS, Rajski SR, Currie CR, Andes DR, Bugni TS. Dereplication of Natural Product Antifungals via Liquid Chromatography-Tandem Mass Spectrometry and Chemical Genomics. Molecules. 2024 Dec 28;30(1):77.

(17)   Chevrette M.G., Carlson C.M., Ortega H.E., Thomas C., Ananiev G.E., Barns K.J., Book A.J., Cagnazzo J., Carlos C., Flanigan W., et al. The Antimicrobial Potential of Streptomyces from Insect Microbiomes. Nat. Commun. 2019;10:516.

(18) De Simeis D, Serra S. Actinomycetes: A Never-Ending Source of Bioactive Compounds-An Overview on Antibiotics Production. Antibiotics (Basel). 2021 Apr 22;10(5):483.

(19)   Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J. Nat. Prod 2020, 83 (3), 770– 803.

(20)   Carroll AR, Copp BR, Grkovic T, Keyzers RA, Prinsep MR. Marine natural products. Nat Prod Rep. 2024 Feb 21;41(2):162-207.

(21)   Arnold A, Alexander J, Liu G, Stokes JM. Applications of machine learning in microbial natural product drug discovery. Expert Opin Drug Discov. 2023 Jul-Dec;18(11):1259-1272.

(22)   Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E, Greiner R, Wishart DS. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016 Nov 4;8:61.

(23)   Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, van der Hooft JJJ, Dorrestein PC, Gerwick WH, Cottrell GW. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J Nat Prod. 2021 Nov 26;84(11):2795-2807.

(24)   Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, Meijer D, Terlouw BR, Biermann F, Blin K, Durairaj J, Gorostiola González M, Helfrich EJN, Huber F, Leopold-Messer S, Rajan K, de Rond T, van Santen JA, Sorokina M, Balunas MJ, Beniddir MA, van Bergeijk DA, Carroll LM, Clark CM, Clevert DA, Dejong CA, Du C, Ferrinho S, Grisoni F, Hofstetter A, Jespers W, Kalinina OV, Kautsar SA, Kim H, Leao TF, Masschelein J, Rees ER,

Reher R, Reker D, Schwaller P, Segler M, Skinnider MA, Walker AS, Willighagen EL, Zdrazil B, Ziemert N, Goss RJM, Guyomard P, Volkamer A, Gerwick WH, Kim HU, Müller R, van Wezel GP, van Westen GJP, Hirsch AKH, Linington RG, Robinson SL, Medema MH. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023 Nov;22(11):895-916. doi: 10.1038/s41573-023-00774-7.

(25)   Brittin NJ, Anderson JM, Braun DR, Rajski SR, Currie CR, Bugni TS. Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics. J Nat Prod. 2025 Feb 28;88(2):361-372.

(26)   Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. Nat. Methods 2019, 16 (4), 299– 302

(27)   Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J. Nat. Prod. 2021, 84 (11), 2795– 2807

(28)   Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. Proc. Natl. Acad. Sci. U. S. A. 2015, 112 (41), 12580– 12585

(29)   Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S. Database-Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC. Nat. Mach Intell 2020, 2 (10), 629– 641

(30)   Ringuette AE, Spock M, Lindsley CW, Bender AM. DARK Classics in Chemical Neuroscience: Carfentanil. ACS Chem Neurosci. 2020 Dec 2;11(23):3955-3967.

(31)   Devereaux AL, Mercer SL, Cunningham CW. DARK Classics in Chemical Neuroscience: Morphine. ACS Chem Neurosci. 2018 Oct 17;9(10):2395-2407.

(32)   Burns SM, Cunningham CW, Mercer SL. DARK Classics in Chemical Neuroscience: Fentanyl. ACS Chem Neurosci. 2018 Oct 17;9(10):2428-2437.

(33)   Mella-Raipán J, Romero-Parra J, Recabarren-Gajardo G. DARK Classics in Chemical Neuroscience: Heroin and Desomorphine. ACS Chem Neurosci. 2020 Dec 2;11(23):3905-3927.

(34)   Gardner EA, McGrath SA, Dowling D, Bai D. The Opioid Crisis: Prevalence and Markets of Opioids. Forensic Sci Rev. 2022 Jan;34(1):43-70.

(35)   Uljon S. Advances in fentanyl testing. Adv Clin Chem. 2023;116:1-30.

(36)   Palmquist KB, Truver MT, Shoff EN, Krotulski AJ, Swortwood MJ. Review of analytical methods for screening and quantification of fentanyl analogs and novel synthetic opioids in biological specimens. J Forensic Sci. 2023 Sep;68(5):1643-1661.

(37)   Broussard LA. Choosing a Fentanyl Immunoassay. J Appl Lab Med. 2024 Sep 3;9(5):871-873.

(38)   Bates MN, Helm AE, Barkholtz HM. Screening for Forensically Relevant Drugs Using Data-Independent High-Resolution Mass Spectrometry. Chem Res Toxicol. 2024 Apr 15;37(4):571-579.

(39)   Tabarra I, Soares S, Rosado T, Gonçalves J, Luís Â, Malaca S, Barroso M, Keller T, Restolho J, Gallardo E. Novel synthetic opioids - toxicological aspects and analysis. Forensic Sci Res. 2019 Jul 3;4(2):111-140.

(40)   Papsun DM, Krotulski AJ, Logan BK. Proliferation of Novel Synthetic Opioids in Postmortem Investigations After Core-Structure Scheduling for Fentanyl-Related Substances. Am J Forensic Med Pathol. 2022 Dec 1;43(4):315-327.

(41)   Clinton HA, Thangada S, Gill JR, Mirizzi A, Logan SB. Improvements in Toxicology Testing to Identify Fentanyl Analogs and Other Novel Synthetic Opioids in Fatal Drug Overdoses, Connecticut, January 2016-June 2019. Public Health Rep. 2021 Nov-Dec;136(1_suppl):80S-86S.

Chapter 2: Dereplication of Natural Product Antifungals via Liquid Chromatography–Tandem Mass Spectrometry and Chemical Genomics

2.1   INTRODUCTION

Invasive fungal infections have become an increasing global health concern, with escalating rates of infection by multi-drug resistant (MDR) strains and a limited number of antifungals for treatment [1,2,3,4]. This particularly impacts sensitive populations such as immunocompromised patients and organ transplant recipients, with mortality rates ranging from 30 to 95% and hospital stays extending from 46 to 140 days [2,5]. Treatments are limited to three classes of antifungal used for invasive fungal infections: polyenes, azoles, and echinocandins. Thus, the development of double and triple resistance by dangerous pathogens such as *Candida albicans*, *C. auris*, *C. glabrata*, and *Aspergillus fumigatus* is highly concerning [6,7,8]. Novel antifungals with new mechanisms are clearly needed.

Addressing this urgent need, we present a novel discovery pipeline that integrates high-resolution Liquid Chromatography–Tandem Mass Spectrometry (LC-MS/MS) with Yeast Chemical Genomics (YCG) [9,10,11]. This innovative platform seamlessly combines the structural dereplication capabilities of LC-MS/MS [12,13] with the functional insights provided by YCG, offering an efficient approach to identify and characterize antifungal compounds. Central to our strategy is the prioritization of bacterial strains sourced from under-explored niches—
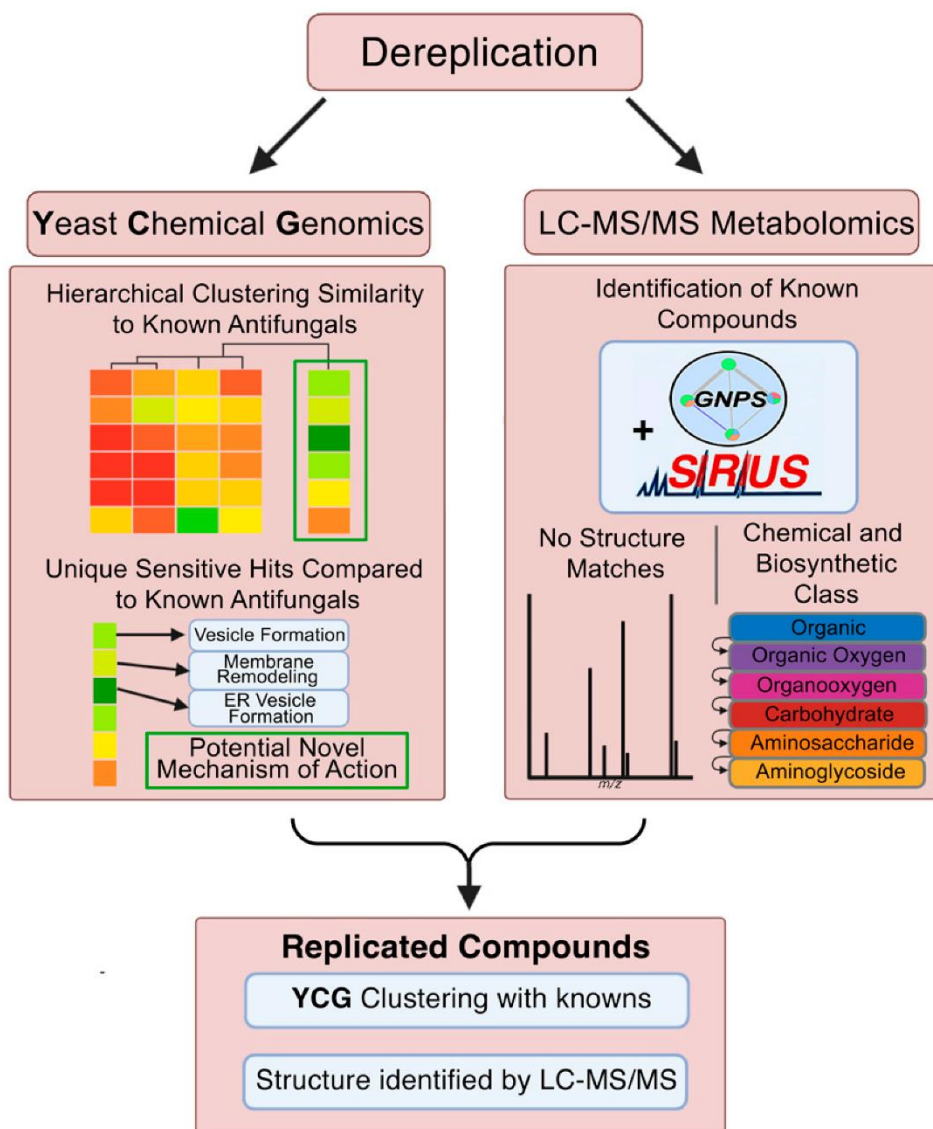
Figure 2.1 Identification and dereplication of antifungal natural product fractions using YCG and LC-MS/MS. Bacterial isolates from under-explored microbiomes were cultured. Cells and spent growth media were extracted, extensively fractionated, and assayed for inhibition of *Candida albicans* strain K1. Positive fractions were further screened against MDR strains *C. auris* B11211 and *C. glabrata* 4720. YCG and LC-MS/MS were then used in parallel to identify fractions with known or otherwise undesirable MoA profiles and chemical structures.

marine invertebrates, insects, and human microbiomes—which are renowned for their rich diversity of unique metabolites [11,14]. Unlike conventional methods that often rely on single-dimension screens, our integrated approach simultaneously assesses both the compound structure and mechanism of action (MoA), enhancing discovery accuracy. The prioritized bacterial strains

were processed to generate extracts, which underwent a two-step fractionation process to yield high-purity natural product fractions suitable for high-throughput screening [15]. Fractions were then screened at four concentrations against *Candida* albicans K1, followed by tests against MDR strains *C. glabrata* and *C. auris*, and counter-screened for hemolytic toxicity against red blood cells. Out of over 40,000 fractions screened, 450 samples displayed activity against MDR fungal pathogens. By leveraging the dual functionality of LC-MS/MS and YCG, our pipeline not only minimizes the rediscovery of known compounds but also focuses efforts on promising candidates by integrating structural data with MoA insights (Figure 2.1).

2.2   RESULTS & DISCUSSION

LC-MS/MS spectral data from extract fractions were analyzed using GNPS [16] and SIRIUS 5 (ver. 5.8.0) [17,18,19,20]; GNPS contained annotated experimental data for ~600,000 molecules. SIRIUS 5 employed database-independent structure predictions, thus expanding comparative abilities to >110,000,000 unique structures present in databases like PubChem and ChemSpider. In parallel, YCG provided mechanism of action (MoA) insights [21,22,23]. For YCG, pools of DNA-barcoded *Saccharomyces cerevisiae* single-gene knockout strains were grown in the presence of extract fractions with antifungal activity. Strain populations were quantified by barcode DNA sequencing to generate lists of hypersensitive and resistant strains, which provided a chemical genomic profile. These characteristic profiles were then compared with those from known antifungals and other fractions for dereplication. To leverage YCG for our screening library, we optimized the system for use with 50 µL cultures in 384-well plates in a semi-automated fashion. Since each well in our fractionated library contained only ~100 µg of material, we opted to employ 384 well format over the more traditional 96 well format.

To evaluate LC-MS/MS and YCG for the reliable identification of known compounds in

this system, we spiked bacterial cultures that were devoid of antifungal activity with a variety of clinically used antifungals, namely the polyenes amphotericin B and natamycin, the echinocandins caspofungin and micafungin, and the azoles itraconazole and voriconazole. Extraction and fractionation were followed by testing for antifungal activity against *C. albicans* to identify active wells, which were then analyzed by LC-MS/MS and queried via GNPS and SIRIUS 5 for the positive identification of the antifungals (Supplementary Material Datasets S1.1–S1.7). In parallel, active wells were tested by YCG using the "Diagnostic" library of 310 DNA-barcoded yeast knockouts [23]. After multiplexed PCR, amplicons from all the wells of a 384-well plate were combined and sequenced. Sequence data were analyzed using BEAN-counter (ver. 2.6.1) [24]. The unique YCG output for each compound, a vector related to the growth of all the knockout strains, was grouped with others by similarity using hierarchical clustering; groupings were visualized using TreeView 3.0 (ver. 3, beta 1) [25] (Figure S1A). Spikes of itraconazole, voriconazole, and micafungin showed YCG profiles similar to pure compounds (Figure 2.2 and Figure S1). Short lists of the most hypersensitive (or occasionally, most resistant) strains, which we term "YCG Profiles", were diagnostic of compounds or compound classes (Figure 2.2A–C, y-axes). YCG profiles sometimes reflected MoA, for example that of micafungin, which targets the cell wall, included the knockouts of the cell wall assembly and maintenance genes *SSD1*, *SKT5*, *CHS7*, and *CWH41* (Figure 2.2B). Additionally, the positive control MMS (methyl methanesulfonate), a DNA-damaging agent, included the knockouts of the DNA repair and maintenance genes *MMS1*, *MUS81*, *RTT101*, *RAD5*, *SGS1*, *CTF4*, *MMS4*, and *RAD55* (Figure 2.2C). For the positive control benomyl, which targets tubulin, the YCG profile included the knockouts of the tubulin and tubulin-folding genes *TUB3*, *CIN1*, and *CIN4* (Figure 2.2C). Conversely, itraconazole and voriconazole did not induce hypersensitivity in any gene knockout

directly related to their target, ergosterol synthesis (Figure 2.2A); we attributed this lack of connectivity between raw chemical genomic data and MoA, which is not infrequent in our experience to the complexity and interconnectivity of biological pathways.

Unlike the azoles and micafungin, the spiked caspofungin profile only partially matched that of the pure compound, and the spiked amphotericin B and natamycin did not cluster with their
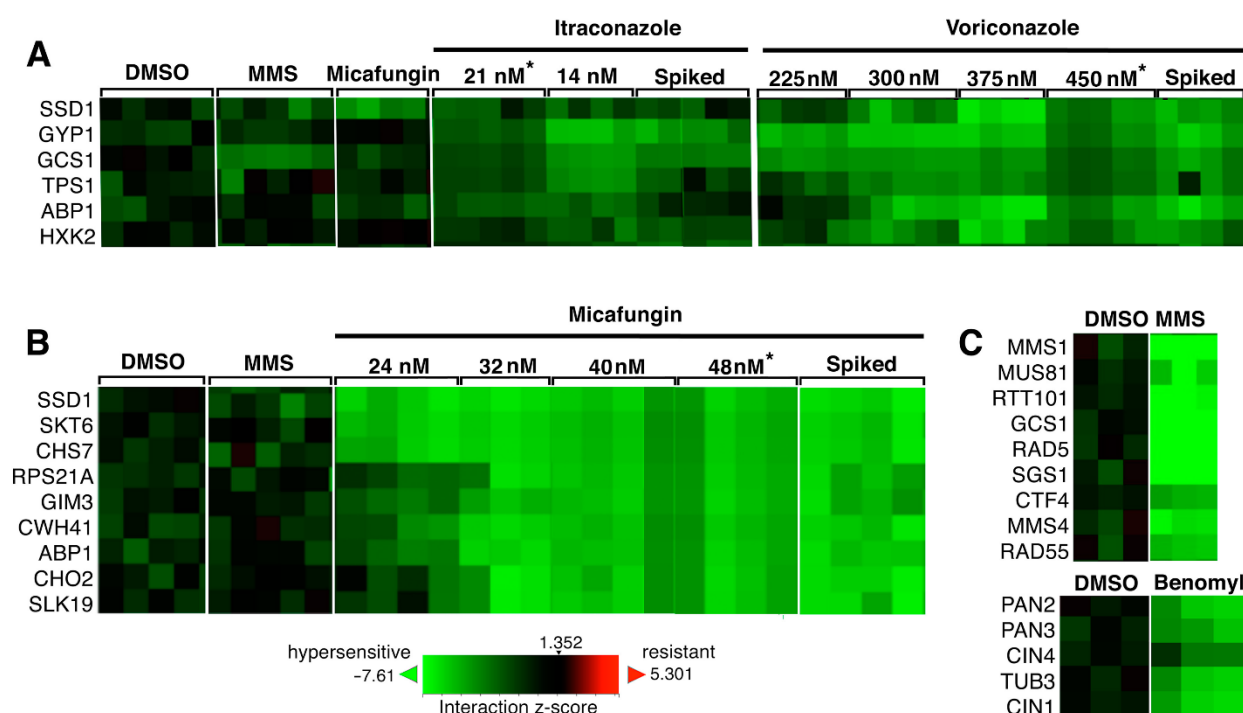
Figure 2.2: Detection by YCG of known antifungals in spiked cultures. Heatmaps indicate a spectrum of knockout strains (y-axes), from those with greatly diminished abundance following growth in the presence of antifungal compared to the DMSO control (bright green) to those that were unaffected (black); each column on the x-axes indicates a replicate of the indicated compound; 4–5 replicates were performed per condition, and occasionally, a replicate failed to produce data, resulting in fewer columns. Pearson correlation for the complete dataset consisting of four replicates of 287 strains is 0.69 (SD = 0.05) supporting YCG reproducibility. The strains shown are those most obviously correlated to antifungal activity out of 310 possible targets (see **Supplementary Material Dataset S5**). (**A**) Itraconazole-, voriconazole-, and (**B**) micafungin-spiked fraction heatmaps closely resembled those of each pure compound. Shown for comparison are heatmaps for negative control DMSO and positive control MMS (methyl methanesulfonate), benomyl, and micafungin, for the same genes. (**C**) The positive controls MMS and benomyl showed characteristic signals elsewhere in the heatmap. * YCG signal intensities that diminish with increasing antifungal concentration are likely attributable to changes in cell populations associated with the total library growth inhibition exceeding 50%.
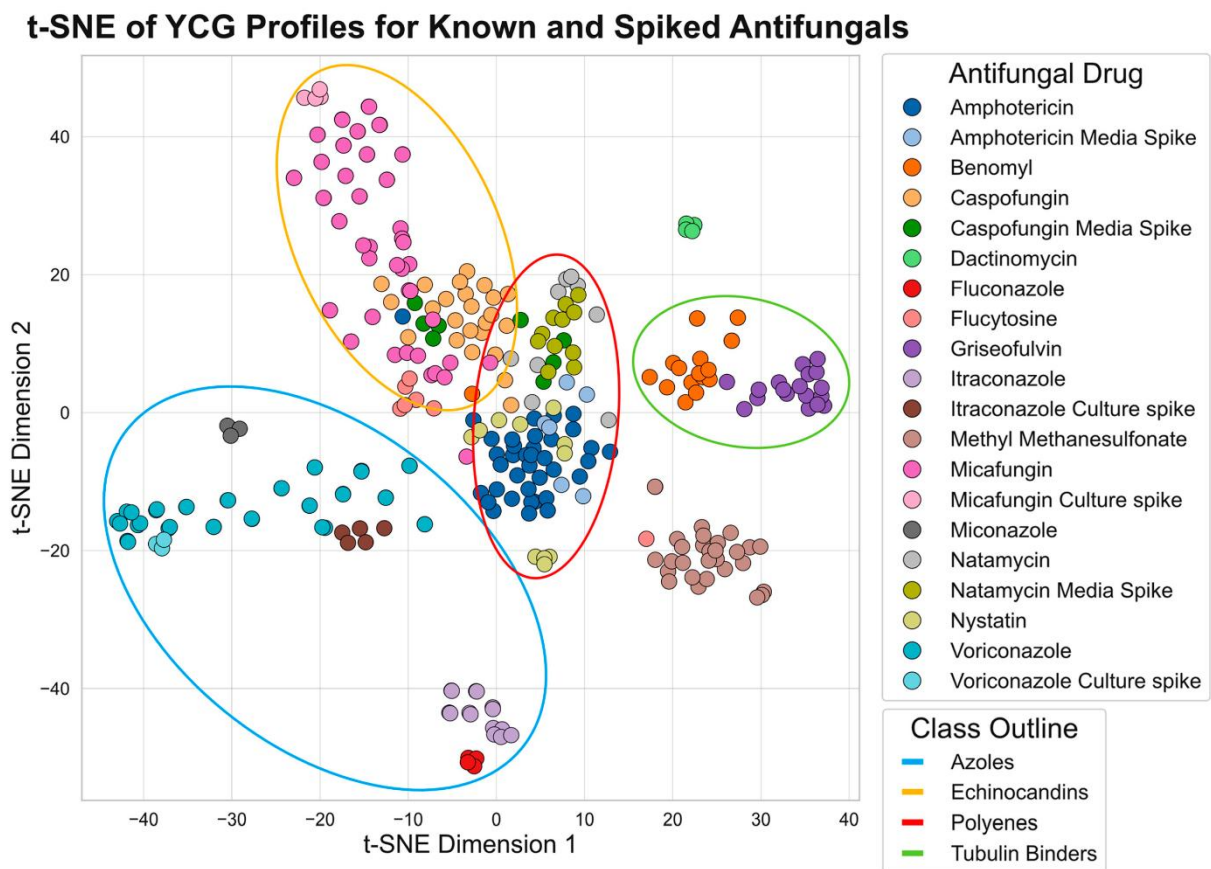
Figure 2.3: The t-distributed Stochastic Neighbor Embedding (t-SNE) analysis of full YCG datasets (313 strains) for the pure and spiked antifungals included in this experiment (amphotericin B, natamycin, caspofungin, micafungin, itraconazole, and voriconazole) as well as a range of additional antifungals; the benzamidazole benomyl; the actinomycin dactinomycin; the pyrimidine analog flucytosine; the methanosulfonate ester MMS; the azole miconazole; and the polyene nystatin.

respective pure compounds or show similar heatmap profiles (**Figures S2 and S3**); in fact, new heatmap signatures appeared for spiked caspofungin and amphotericin B. We speculated that caspofungin, amphotericin B, and natamycin were modified by the bacterial culture, or that they stimulated the production of new compounds by the bacterial culture, or both. To test this hypothesis, each of the three compounds was spiked into uninoculated fractionated media; such spiked samples were devoid of bacterial exposure. Importantly, the spiked compound YCG profiles clustered with the respective pure compounds (**Figure S3**), thus validating the notion of microbial involvement in generating the unexpected new heatmap signatures. Dimensional

reduction using t-SNE enabled the visualization of bioactivity clusters within the BEAN counter data [**24**]. The visualization of these distinct clusters of drug classes underscored YCG's utility in differentiating among the MoAs characteristic of various antifungal classes. Notably, antifungals spiked in either the culture or media grouped closely with their corresponding pure compounds (**Figure 2.2**D), effectively clustering within their respective drug families.

Next, we assessed the YCG profiles for known compounds that were confidently identified by LC-MS/MS to ensure that the YCG profiles were consistent with compound class. As an example of the complementary nature of LC-MS/MS and YCG, we employed the macrotetrolide family of compounds. Macrotetrolides (monactin, dinactin, trinactin, tetranactin, or nonactin) were identified by LC-MS/MS in fractions H5 and H7 of an extract from the insect microbiome-derived strain SID7958 (**Figure 2.3**). By HCA analyses of the YCG profiles, five other fractions were found to have similar profiles (**Figure S4A**). All seven YCG profiles showed a similar signature with the following knockouts being diagnostic: *SMY1*, *SUR1*, *SEC72*, *BST1*, *GIM3*, *GUP1*, and *SCS2* (**Figure S4B,C**). Importantly, targeted LC-MS/MS analyses enabled the identification of macrotetrolides in each of the five fractions where YCG indicated macrotetrolides (**Figure 2.3**A, **Supplementary Material Datasets S2.1–S2.5**).

Even though all seven fractions displayed similar YCG profiles, there was no convergence on one specific singular function or pathway. Therefore, we used the CG-Target software (ver. 0.6.1) to associate the YCG profiles to yeast bioprocesses by leveraging known genetic interactions across the whole yeast genome. CG-Target [**26**] uses a genome-wide interaction network (established via pairwise gene knockouts) to link chemical genomics data to likely impacted bioprocesses. The input of YCG data for the seven macrotetrolide fractions found a concentrated overlap of genes linked to mitochondrial function (**Figure 2.3**B), including the *MDM38*
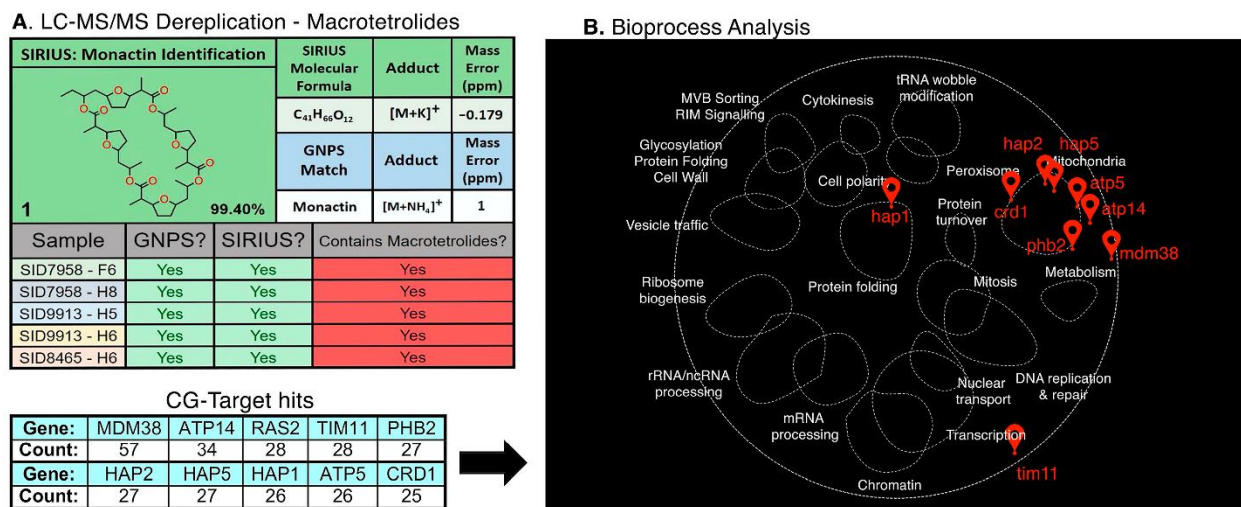
Figure 2.4: Identification of macrotetrolides by YCG and LC-MS/MS. (A) SIRIUS 5- and GNPS-driven analysis of the LC-MS/MS of the seven clustered fractions identified some combination of the macrotetrolides monactin, dinactin, trinactin, and nonactin in each. (B) CG-Target analysis of all 7 fractions and compilation of genes in the top 30 "Driver Scores" resulted in a grand total hit list of genes with largely mitochondrial/ion gradient functions, as illustrated in TheCellMap analysis.

mitochondrial K^+/H^+ exchange protein, the *ATP14*/*TIM11*/*ATP5* subunits of mitochondrial F1F0 ATP synthase, and the *HAP1/2/5* activators of respiratory gene expression. This network was visualized using TheCellMap [27], matching the profiles to mitochondrial bioprocesses (**Figure 2.3**B) in agreement with the known MoA of macrotetrolides, which are known to disrupt ion transport across the mitochondrial membrane.

Next, we evaluated the LC-MS/MS-YCG platform in characterizing polyenes. Polyenes such as amphotericin B, nystatin, and natamycin act primarily through their interaction with ergosterol and the subsequent formation of pores in fungal cell membranes [28,29,30]. Polyenes, in general, are associated with high toxicity and were consequently targeted for exclusion in our search for novel antifungals. The LC-MS/MS identification of polyenes is challenging as polyenes are prone to in-source fragmentation, vastly complicating the spectral analysis. Moreover, polyenes have exceptional potency that, in our experience, leads to bioactivity hits even when the concentrations of compounds are near or beyond LC-MS/MS detection limits. Additionally,
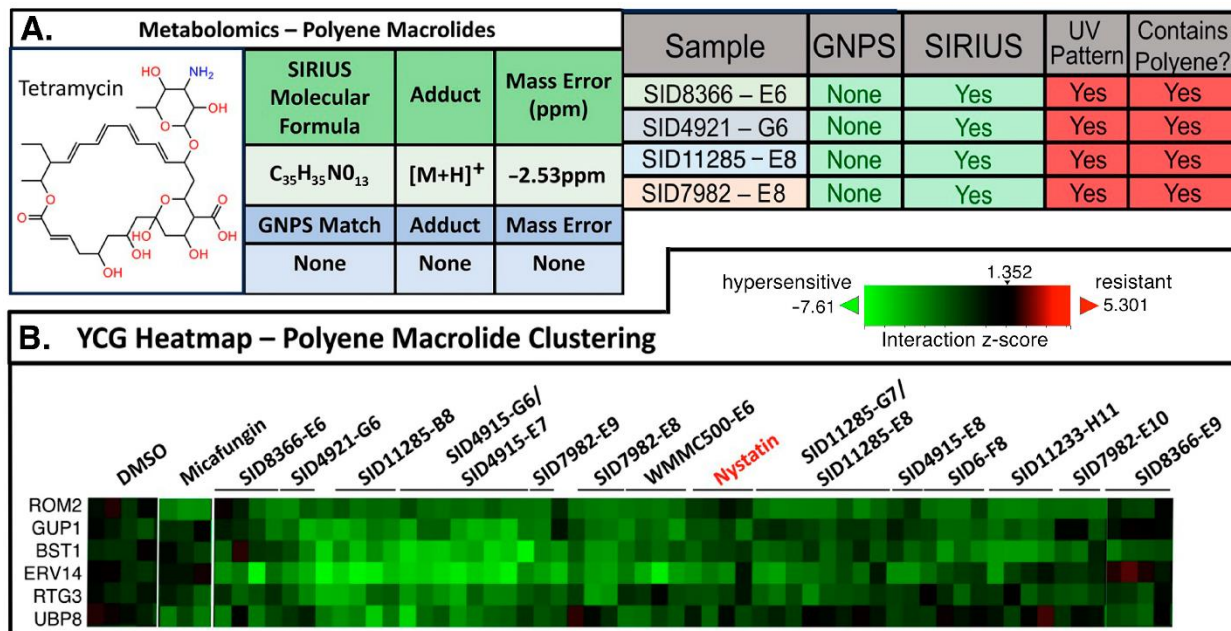
Figure 2.5: Identification of polyenes by LC-MS/MS, YCG, and UV spectroscopy. (**A**) Identification of the polyene antifungals by LC-MS/MS metabolomics processed using GNPS and SIRIUS 5, and by UV-Vis spectroscopy (**Supplementary Material Figure S5, Dataset S3** for UV-Vis data). (**B**) Heatmap of the fractions clustering with the polyene nystatin in a YCG experiment.
See **Supplementary Material Dataset S5** for comprehensive heatmap data.

MS/MS-based methods are convoluted by noise at low intensities and signal loss from in-source fragmentation. Although polyenes display distinctive UV-Vis spectra (**Figure S5**), this method is not sufficiently sensitive to identify them given their low concentrations in the majority of our working samples. Thus, considering these challenges and limitations, we evaluated the YCG datasets to help identify polyenes in hit wells based on their potent activity rather than physical detection methods.

Using the distinct UV spectral pattern characteristic of polyenes, we confirmed the presence of polyene antifungals, and the LC-MS/MS metabolomics tool SIRIUS 5 was employed to predict the structure of the antifungal agent; in this fashion, polyenes were found in seven wells. On the basis of the HCA of the YCG profiles, those seven wells and eight others grouped with the polyene nystatin (**Figure 2.4** and **Supplementary Material Datasets S3 and S5.3**). While the

clustering of the YCG data helped identify other wells containing polyenes, the hypersensitive YCG profiles for polyenes were only moderately diagnostic. In general, polyenes show profiles consistent with an impact on vesicle-mediated trafficking; for example, the most sensitive mutant is Δ-ERV14, which is involved with vesicle formation. We have two hypotheses supporting the observed YCG profile for polyenes: 1. cell stress due to polyenes requires increased vesicle-mediated trafficking to help stabilize the cell wall; or 2. ergosterol plays a role in vesicle formation and amphotericin could be interacting with ergosterol in vesicles. Regardless, the *ERV14* knockout was found to be consistently hypersensitive. Although the YCG signature is not as prominently defined as was the case for micafungin (**Figure 2.2**B), the combination of LC-MS/MS and YCG proved valuable for de-prioritizing wells containing polyenes.

2.3   MATERIALS & METHODS

2.3.1   *Microbiome strain isolation and in vitro activity testing*

Microbiome bacterial strains were collected, prioritized, fermented for production of compounds of interest, extracted, and fraction libraries generated as described from marine invertebrates [15], insects [11] and humans [31]. High throughput screening for growth inhibition of *Candida albicans* K1 was as previously described [15]. Wells with activity toward *C. albicans* were tested against multiply drug-resistant *Candida auris* B11211 and *Candida glabrata* 4720 as well as for hemolysis and human cell line cytotoxicity [15].

2.3.2   *Fermentation for Library Generation*

For each prioritized strain, 10 mL seed cultures (25 × 150 mm tubes) in medium DSC (5 g soluble starch, 10 g glucose, 5 g peptone, 5 g yeast extract per liter made with 50% artificial seawater) were inoculated and shaken (200 RPM, 28 °C) for seven days. Seed cultures (2.5 mL) were used to inoculate three 100 mL of media in 500 mL baffled flasks using two distinct media

(2 X 100 mL ASW-A and 100 mL RAM2) containing Diaion HP20 (7% by weight). ASW-A was made using 20 g soluble starch, 10 g glucose, 5 g peptone, 5 g yeast extract, 5 g $CaCO_3$ per liter of artificial seawater; RAM2 was made using 4 g corn meal, 10 g glucose, 15 g maltose, 7.5 g pharmamedia, 5 g yeast per liter of 50% artificial seawater. After fermentation for 7 days, the cells and HP20 were filtered using Miracloth, and the cells and HP20 were extracted with acetone (100 mL for 30 min).

2.3.3 Bacterial Extract Plate Library Generation

The crude extract was dried and then dissolved using the following solvent mixture: 1 mL dimethyl sulfoxide (DMSO), 1 mL methanol, and 10 mL H_2O. Subsequently, the mixture was fractionated on an Isolute ENV+ (500 g cartridge) using a modified Gilson GX-271 liquid handler with 100% H_2O (10 mL), 25% CH_3OH/H_2O [fraction 1], 50% CH_3OH/H_2O [fraction 2], 75% CH_3OH/H_2O [fraction 3], 100% CH_3OH [fraction 4] (8 mL of each solvent). The 100% water fraction went directly to waste while the remaining four fractions were collected and subsequently dried in a speedvac. Each fraction was dissolved in DMSO and subjected to HPLC using a Gilson HPLC integrated with a Gilson 215 fitted with a 96-well plate deck capable of holding ten plates. For HPLC, a Phenomenex Monolithic C18 (3 mm ID X 100 mm) was used. The following HPLC gradients were used: Fraction 1 (F1) t = 0 to 2 min, 90% H_2O/10% CH_3CN t = 14.5 to 14.51 min, 50% H_2O/50% CH_3CN t = 19 to 21.5 min, 100% CH_3CN t = 22 to 27 min, 90% H_2O/10% CH_3CN Fraction 2 (F2) and Fraction 3 (F3) t = 0 to 2 min, 90% H_2O/10% CH_3CN t = 19 to 21.5 min, 100% CH_3CN t = 22 to 27 min, 90% H_2O/10% CH_3CN Fraction 4 (F4) t = 0 to 2 min, 90% H_2O/10% CH_3CN t = 5 to 5.01 min, 70% H_2O/30% CH_3CN t = 19 to 32 min, 100% CH_3CN t = 32.5 to 37.5 min 90% H_2O/10% CH_3CN.

For each fraction above (F1–F4), 20 fractions were collected in 96-deepwell plates such

that for each extract, metabolites were arrayed in 80 wells. Each plate was quantified using ELSD using previously published methods [15, 32]. The plates were dried in a speedvac. DMSO (20 μL) was added to each well to dissolve the material. The contents were then transferred to Labcyte Echo plates prior to high-throughput screening.

2.3.4 Bioactivity High Throughput Screening

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

2.3.5 LC-MS/MS

Liquid chromatography tandem mass spectrometry (LC-MS/MS) data were acquired using a Bruker maXis II Ultra-High-Resolution LC-QTOF mass spectrometer (Bruker Scientific LLC., Billerica, MA, USA) coupled to a Waters Acquity H-Class UPLC system (Waters, Milford, MA,

USA) and operated by the Bruker Hystar 3.2 software. Chromatographic gradients were performed with a mixture of methanol and water (containing 0.1% formic acid) on an RP C-18 column (Phenomenex Kinetex 2.6 μm, 2.1 mm × 100 mm; Phenomenex, Torrance, CA, USA) at 0.3 mL/min. The method was as follows: 0-1 min (10%-10% MeOH in H_2O), 1–12 min (10%–97% MeOH in H_2O), and 12–15.5 min (97% MeOH in H_2O). A mass range of m/z 50–1550 was measured in positive ESI mode for all spectra. The mass spectrometer was operated with the following parameters: capillary voltage of 4.5 kV, nebulizer pressure of 1.2 bar, dry gas flow of 4.0 L/min, dry gas temperature of 205 °C, and scan rate of 2 Hz. Tune mix (ESI-L low concentration; Agilent, Santa Clara, CA, USA) was introduced through a divert valve at the end of each chromatographic run for automated internal calibration. MS/MS spectra were acquired at scan speeds of 2 Hz for signals above 1 x 104 counts and 6 Hz for signals above 1 x 106 counts. MS/MS spectra were collected using a stepping collision energy (CE) where CE increased linearly during MS/MS spectra collection. From time 0 to 32, the collision RF was 600, transfer time was 80, and CE was 70eV. From time 33-66, the collision RF was 600, transfer time was 72, and CE was 100eV. From time 67-100, the collision RF was 600, transfer time was 65, and CE was 130eV. The precursor list was set to exclude precursor ions for 0.2 min after two spectra with the same precursor ion were acquired. Additionally, if the intensity of an excluded precursor ion rose five-fold from the initial spectra, it was recollected.

2.3.6 *Metaboscape*

MetaboScape, a software developed by Bruker (ver. 5.0.0, Build 683), was utilized to process raw LC-MS/MS data obtained from the Bruker maXis II LC-QTOF instrument (Bruker Scientific LLC., Billerica, MA, USA). To extract comprehensive information, T-ReX 3D (LC-QTOF) algorithm was employed with specific parameters. Initially, all samples underwent filtering

with a value of 1 for both "Minimum # Features for Extraction" and "Presence of features in minimum # of analyses" settings. This filtering approach ensured the extraction of all features, including sample-specific features. The resulting filtered features were then subjected to further processing using customized T-ReX 3D parameters.

The customized T-ReX 3D processing parameters included an intensity threshold of 5000 counts, a minimum peak length of 7 spectra, feature signal based on intensity, enabled recursive feature extraction with a minimum peak length of 6 spectra for recursive processing, retention time range of 0.5 min to 16 min, and mass range of 50 m/z to 1550 m/z. Moreover, the software enabled the import and averaging of MS/MS data, and collision energy grouping was activated. Ion deconvolution was performed using the T-ReX Default Metabolomics Positive Ions settings. Additionally, mass recalibration was conducted using a customized T-ReX Positive Recalibration approach. The recalibration involved setting the retention time (RT) window to 17.5-20 min and utilizing the "Tuning Mix ES-TOF (ESI) Pos" as the list of calibrant signals. Upon successful bucketing of all LC-MS/MS data, data were exported using the "Export to GNPS" option.

2.3.7 LC-MS/MS Dereplication

From Metaboscape, the data were exported using the "Export to GNPS" option which provided the files needed for GNPS and SIRIUS 5 as mascot generic format (MGF) files. For GNPS, the file was processed using their METABOLOMICS-SNETS-V2 (v. release_28). On the GNPS servers, the data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.005 Da and a MS/MS fragment ion tolerance of 0.005 Da. A network was then created where edges were filtered to have a cosine score above 0.7 and more than 6 matched peaks. Further, edges

between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then queried against GNPS' spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks. For SIRIUS, the data was processed using SIRIUS 5 (ver. 5.8.0).

2.3.8 *Yeast Chemical Genomics (YCG)*

The 310-strain *S. cerevisiae* haploid non-essential gene "Diagnostic" knockout library was generously provided by Charles Boone and prepared by Jeff Piotrowski [23]. For each antifungal compound to be tested, we aimed for 20–50% growth inhibition versus DMSO control after static growth for 17–24 h at 30°C. Importantly, optimum heatmap signatures generally correlated to growth inhibition levels < 50% [23]. Inhibition levels exceeding 50%, in many cases, led to muted signal intensities [23]. Three concentrations including and bracketing the one determined above were used in the experiment; each condition was carried out in 4–5 replicates. In some cases, replicate experiments failed to produce data thus affording heatmap lane variances observed in Figures 2.2 and 2.4. DMSO-dissolved compounds were dispensed into 384- or 96-well flat-bottom plates using an Echo 650 Liquid Handler (Beckman-Coulter, Indianapolis, IN, USA). Library pool stocks containing approximately 250 cells/strain/µl were thawed, diluted 100-fold into YPD Broth (Thermo Fisher Scientific, Waltham, MA, USA), and 50 or 200 µl of diluted library was dispensed for 384-well or 96- well plates, respectively. OD600 was measured at 17–24 hours growth, and growth was continued to a final 48 hours before harvesting. Total genomic DNA was purified using a PureLink Pro 96 Genomic DNA Purification kit (Life Technologies, Carlsbad, CA, USA),

AcroPrep 384-Well Filter Plates (VWR, Radnor, PA, USA), and Zymolase (Thermo Fisher Scientific, Hampton, NH, USA). Barcodes were amplified using indexed forward primers and a universal reverse primer as previously described [33]. Ten µl of each 25 µl reaction were pooled, loaded onto a 2% agarose gel, and the 267 bp product band was purified using a QIAquick Gel Extraction Kit (Qiagen, Germantown, MD, USA). A sample was submitted to the University of Wisconsin-Madison Biotechnology Center for sequencing on an Illumina (San Diego, CA, USA) MiSeq. Data was processed using BEAN-counter (ver. 2.6.1)[24], TreeView3 (ver. 3.0, beta-1)[25], CG-Target (ver. 0.6.1)[26], and TheCellMap. An alternative barcode amplification/sequencing workflow with improved flexibility and enhanced cost effectiveness was developed and used in supporting experiments. Barcodes in the Diagnostic yeast library were amplified with one pair of non-indexed primers containing the same annealing sequences (underlined) as those used in the original method but with tails containing priming sites for a second round of PCR (bold) (Forward: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGTCCACGAGGTCTCT-3', Reverse: 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGG CACGTCAAGACTGT CAAGG-3'). In 15 µl reactions, 4.8 µl genomic DNA template was amplified (2 min 95°C, 30x[30s 95°C/30s 56°C/45s 68°C], 10 min 68° C), resulting in 238 bp products. A second round of PCR used 2 µl of this product as template in 20 µl reactions with pairs of indexed primers (Forward: 5'-AATGATACGGCGACC ACCGAGATCTACACNNNNNNNNNNTCGTCGGCAGCGTC-3', Reverse 5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTCTCGTGGGCTCGG-3') that added two unique 10 bp index tags and Illumina P5 and P7 stems (5 min 95°C, 12x[60s 95°C/30s 57.5°C/45s 68°C], 10 min 68°C). A portion (5-10 µl/well) of the resulting 311 bp products were pooled and gel-purified as before. The dual unique index tags permitted inexpensive

sequencing on a Illumina NovaSeq 6000 shared lane, and initial amplification with one pair of dedicated primers allows universal applicability of index primer pairs.

2.3.9 CG-Target

CG-Target (ver. 0.6.1) was used essentially as described [26] resulting in, for each compound identity/concentration condition, multiple possible functional assignments as GO (Gene Ontology) terms; these are ranked by p-value, False Discovery Rate, z-score, Driver (gene or genes driving that assignment) Score, and Driver Name. The best condition for each compound, ideally resulting in 20-50% inhibition of the culture and highly negative CG scores indicating hypersensitive strains, was chosen for further analysis. The data associated with each condition was sorted by the first score in the Driver Score column and, for the thirty best (highest scoring) rows, the most commonly occurring genes were extracted from the Driver Name column for a grand total count list, and the most common Gene Ontology terms were noted.

2.3.10 TheCellMap.org

TheCellMap (version [at: https://thecellmap.org/?q=bni1] [27] was used by entering, into the main search field, lists of genes from genetic interaction networks that CG-TARGET found to resemble the hypersensitive/resistant profiles resulting from exposure to antifungal compounds. Concentrations of "hits' in particular functional areas were taken as evidence for compound mechanism of action.

2.3.11 Spiked Bacterial Library Plates for YCG Screening

Spiked library plates were prepared by culturing the bacterial strain WMMC1424 with the same methods as described in the "Microbiome strain isolation and in vitro activity testing". WMMC1424 was identified as a *Micromonospora* sp. SG15 with a 99.5% (1361/1368) 16S alignment score, and the WMMC1424 extract was previously shown to have no antifungal activity

and no antifungal compounds were found using LC-MS/MS based metabolomics. To spike the extracts, 5 mg of each antifungal was added to the culture (each 1L in volume) 4 hours before extraction and left to shake to adhere to the extraction resin. The antifungals spiked into each 1L culture were natamycin, amphotericin B, voriconazole, itraconazole, caspofungin, and micafungin. Once the active wells were identified, they were analyzed by YCG and LC-MS/MS based metabolomics to confirm the presence of each spiked antifungal.

2.3.12 t-SNE Analysis of BEAN-Counter Output

To perform t-SNE analysis, we first prepared a matrix of the YCG Profiles from BEAN-Counter, with rows representing experimental conditions or compound treatments (e.g., pure compounds, spiked samples, etc..) and columns representing responses across selected features (e.g., growth differentials in specific knockout strains). Data were normalized to standardize feature scales using the Robust Scaler within the scikit-learn preprocessing package. We applied t-distributed Stochastic Neighbor Embedding (t-SNE) using key parameters: perplexity (set to 10), iterations (750), and early exaggeration (18). The t-SNE was implemented using the scikit-learn library in Python, with parameter adjustments based on pilot runs to optimize visualization. The results were plotted with different colors to distinguish compound types (e.g., pure vs. spiked samples). The drugs of each antifungal class are outlined according to class to demonstrate the observable difference in the observed YCG profiles.

2.4 DATA SUMMARY

All experimentals (wet chemistry/biology and computation) [31,32,33]; Figure S1: HCA and expanded heatmap analysis of itraconazole, voriconizole, and micafungin and spiked fractions; Figure S2: YCG heatmap analysis of pure and culture-spiked caspofungin, amphotericin B, and natamycin; Figure S3: YCG heatmap and HCA analysis of pure and media-spiked caspofungin,

amphotericin B, and natamycin; Figure S4: Identification of macrotetrolides by YCG; Figure S5: Detection of polyene macrolide antifungals by UV/Vis spectroscopy in pure compound stocks and complex bacterial extracts; Figure S6: Detection of polyenes by YCG and LC-MS/MS metabolomics using the SIRIUS 5 software suite; Dataset S1: Datasets for known antifungals (LC-HR-MS/MS, Sirius); Dataset S2: Dataset for macrotetrolide dereplication in complex bacterial extract fractions; Dataset S3: Dataset for polyenes identification in antifungal active extract fractions; Dataset S4: Dataset containing dereplication data (mirror plots) for samples from strains SID7958, SID8465 & SID9913

2.5 REFERENCES

(1) Fisher, M.C.; Alastruey-Izquierdo, A.; Berman, J.; Bicanic, T.; Bignell, E.M.; Bowyer, P.; Bromley, M.; Brüggemann, R.; Garber, G.; Cornely, O.A.; et al. Tackling the Emerging Threat of Antifungal Resistance to Human Health. Nat. Rev. Microbiol. 2022, 20, 557–571.

(2) Brown, G.D.; Denning, D.W.; Gow, N.A.R.; Levitz, S.M.; Netea, M.G.; White, T.C. Hidden Killers: Human Fungal Infections. Sci. Transl. Med. 2012, 4, 165rv13.

(3) Lee, Y.; Puumala, E.; Robbins, N.; Cowen, L.E. Antifungal Drug Resistance: Molecular Mechanisms in Candida albicans and Beyond. Chem. Rev. 2021, 121, 3390–3411.

(4) Gnat, S.; Łagowski, D.; Nowakiewicz, A.; Dyląg, M. A Global View on Fungal Infections in Humans and Animals: Opportunistic Infections and Microsporidioses. J. Appl. Microbiol. 2021, 131, 2095–2113.

(5) WHO. WHO Fungal Priority Pathogens List to Guide Research, Development and Public Health; World Health Organization: Geneva, Switzerland, 2022.

(6) De Oliveira, H.C.; Bezerra, B.T.; Rodrigues, M.L. Antifungal Development and the Urgency of Minimizing the Impact of Fungal Diseases on Public Health. ACS Bio Med Chem Au 2023, 3, 137–146.

(7) Lamoth, F.; Lockhart, S.R.; Berkow, E.L.; Calandra, T. Changes in the Epidemiological Landscape of Invasive Candidiasis. J. Antimicrob. Chemother. 2018, 73, i4–i13.

(8) Rabaan, A.A.; Sulaiman, T.; Al-Ahmed, S.H.; Buhaliqah, Z.A.; Buhaliqah, A.A.; AlYuosof, B.; Alfaresi, M.; Al Fares, M.A.; Alwarthan, S.; Alkathlan, M.S.; et al. Potential

Strategies to Control the Risk of Antifungal Resistance in Humans: A Comprehensive Review. Antibiotics 2023, 12, 608.

(9)   Hou, Y.; Braun, D.R.; Michel, C.R.; Klassen, J.L.; Adnani, N.; Wyche, T.P.; Bugni, T.S. Microbial Strain Prioritization Using Metabolomics Tools for the Discovery of Natural Products. Anal. Chem. 2012, 84, 4277–4283.

(10)   Chanana, S.; Thomas, C.S.; Zhang, F.; Rajski, S.R.; Bugni, T.S. HCAPCA: Automated Hierarchical Clustering and Principal Component Analysis of Large Metabolomic Datasets in R. Metabolites 2020, 10, 297.

(11)   Chevrette, M.G.; Carlson, C.M.; Ortega, H.E.; Thomas, C.; Ananiev, G.E.; Barns, K.J.; Book, A.J.; Cagnazzo, J.; Carlos, C.; Flanigan, W.; et al. The Antimicrobial Potential of Streptomyces from Insect Microbiomes. Nat. Commun. 2019, 10, 516.

(12)   Nielsen, K.F.; Månsson, M.; Rank, C.; Frisvad, J.C.; Larsen, T.O. Dereplication of Microbial Natural Products by LC-DAD-TOFMS. J. Nat. Prod. 2011, 74, 2338–2348.

(13)   Jouaneh, T.M.M.; Rosario, M.E.; Li, Y.; Leibovitz, E.; Bertin, M.J. Incorporating LC-MS/MS Analysis and the Dereplication of Natural Product Samples into an Upper-Division Undergraduate Laboratory Course. J. Chem. Ed. 2022, 99, 2636–2642.

(14)   Van Moll, L.; De Smet, J.; Cos, P.; Van Campenhout, L. Microbial Symbionts of Insects as a Source of New Antimicrobials: A Review. Crit. Rev. Microbiol. 2021, 47, 562–579.

(15)   Zhang, F.; Zhao, M.; Braun, D.R.; Ericksen, S.S.; Piotrowski, J.S.; Nelson, J.; Peng, J.; Ananiev, G.E.; Chanana, S.; Barns, K.; et al. A Marine Microbiome Antifungal Targets Urgent-Threat 1 Drug-Resistant Fungi. Science 2020, 370, 974–978.

(16)   Aron, A.T.; Gentry, E.C.; McPhail, K.L.; Nothias, L.F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J.M.; Sikora, N.; Vargas, F.; et al. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. Nat. Protoc. 2020, 15, 1954–1991.

(17)   Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. J. Cheminform. 2016, 8, 61.

(18)   Dührkop, K.; Nothias, L.F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M.A.; Petras, D.; Gerwick, W.H.; Rousu, J.; Dorrestein, P.C.; et al. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. Nat. Biotechnol. 2021, 39, 462–471.

(19)   Hoffmann, M.A.; Nothias, L.F.; Ludwig, M.; Fleischauer, M.; Gentry, E.C.; Witting, M.; Dorrestein, P.C.; Dührkop, K.; Böcker, S. High-Confidence Structural Annotation of Metabolites Absent from Spectral Libraries. Nat. Biotechnol. 2022, 40, 411–421.

(20)   Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A.A.; Melnik, A.V.; Meusel, M.; Dorrestein, P.C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. Nat. Methods 2019, 16, 299–302.

(21)   Giaever, G.; Chu, A.M.; Ni, L.; Connelly, C.; Riles, L.; Veronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; Andre, B.; et al. Functional Profiling of the Saccharomyces cerevisiae Genome. Nature 2002, 418, 387–391.

(22)   Parsons, A.B.; Lopez, A.; Givoni, I.E.; Williams, D.E.; Gray, C.A.; Porter, J.; Chua, G.; Sopko, R.; Brost, R.L.; Ho, C.H.; et al. Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast. Cell 2006, 126, 611–625.

(23)   Piotrowski, J.S.; Li, S.C.; Deshpande, R.; Simpkins, S.W.; Nelson, J.; Yashiroda, Y.; Barber, J.M.; Safizadeh, H.; Wilson, E.; Okada, H.; et al. Functional Annotation of Chemical Libraries across Diverse Biologic Processes. Nat. Chem. Biol. 2017, 13, 982–993.

(24)   Simpkins, S.W.; Deshpande, R.; Nelson, J.; Li, S.C.; Piotrowski, J.S.; Ward, H.N.; Yashiroda, Y.; Osada, H.; Yoshida, M.; Boone, C.; et al. Using BEAN-Counter to Quantify Genetic Interactions from Multiplexed Barcode Sequencing Experiments. Nat. Protoc. 2019, 14, 415–440.

(25)   Keil, C.; Leach, R.W.; Faizaan, S.M.; Bezawada, S.; Parsons, L. Treeview 3.0 (beta 1) — Visualization and analysis of large data matrices. Zenodo 2016, 10, 5281.

(26)   Simpkins, S.W.; Nelson, J.; Deshpande, R.; Li, S.C.; Piotrowski, J.S.; Wilson, E.H.; Gebre, A.A.; Safizadeh, H.; Okamoto, R.; Yoshimura, M.; et al. Predicting Bioprocess Targets of Chemical Compounds through Integration of Chemical-Genetic and Genetic Interactions. PLoS Comput. Biol. 2018, 14, e1006532.

(27)   Usaj, M.; Tan, Y.; Wang, W.; VanderSluis, B.; Zou, A.; Myers, C.L.; Costanzo, M.; Andrews, B.; Boone, C. TheCellMap.Org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. G3 Genes Genomes Genet. 2017, 7, 1539–1549.

(28)   Anderson, T.M.; Clay, M.C.; Cioffi, A.G.; Diaz, K.A.; Hisao, G.S.; Tuttle, M.D.; Nieuwkoop, A.J.; Comellas, G.; Maryum, N.; Wang, S.; et al. Amphotericin Forms an

Extramembranous and Fungicidal Sterol Sponge. Nat. Chem. Biol. 2014, 10, 400–406.

(29)   Guo, X.; Zhang, J.; Li, X.; Xiao, E.; Lange, J.D.; Rienstra, C.M.; Burke, M.D.; Mitchell, D.A. Sterol Sponge Mechanism Is Conserved for Glycosylated Polyene Macrolides. ACS Cent. Sci. 2021, 7, 781–791.

(30)   Szomek, M.; Reinholdt, P.; Petersen, D.; Caci, A.; Kongsted, J.; Wüstner, D. Direct Observation of Nystatin Binding to the Plasma Membrane of Living Cells. Biochim. Biophys. Acta Biomembr. 2021, 1863, 183528.

(31)   Salamzade, R.; Cheong, J.Z.A.; Sandstrom, S.; Swaney, M.H.; Stubbendieck, R.M.; Starr, N.L.; Currie, C.R.; Singh, A.M.; Kalan, L.R. Evolutionary Investigations of the Biosynthetic Diversity in the Skin Microbiome Using lsaBGC. Microb. Genom. 2023, 9, 000988.

(32)   Adnani, N.; Michel, C.R.; Bugni, T.S. Universal Quantification of Structurally Diverse Natural Products Using an Evaporative Light Scattering Detector. J. Nat. Prod. 2012, 75, 802–806.

(33)   Piotrowski, J.S.; Simpkins, S.W.; Li, S.C.; Deshpande, R.; McIlwain, S.J.; Ong, I.M.; Myers, C.L.; Boone, C.; Andersen, R.J. Chemical Genomic Profiling via Barcode Sequencing to Predict Compound Mode of Action. In Chemical Biology: Methods and Protocols; Hempel, J.E., Williams, C.H., Hong, C.C., Eds.; Humana Press: Totowa, NJ, USA, 2015; pp. 299–318.

Chapter 3: Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics

3.1   INTRODUCTION

The global surge in multidrug-resistant pathogens has intensified the need for new antimicrobial therapies [1−3]. Natural products (NP) have historically been a rich source of bioactive compounds, with unique organisms from diverse ecosystems providing novel mechanisms and scaffolds to explore [4−6]. NPs have been the best source of drug leads for infectious disease with ~73% of antibiotic pharmacophores having a NP scaffold [7]. However, navigating the vast chemical space of NPs presents significant challenges, primarily the issue of rediscovery [8]. With hundreds of thousands of NPs already reported, reisolating known compounds leads to duplicated efforts and wasted resources, stalling novel discoveries [8]. Dereplication, the process of identifying previously characterized natural products, has become crucial in NP discovery [9]. Although various mass spectrometry (MS) and machine learning (ML) methods have been developed for dereplicating bacterial metabolites, limitations persist, necessitating improved strategies to facilitate dereplication and accelerate the prioritization of novel bioactive scaffolds [8,10].

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

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

Machine learning has emerged as a powerful tool for analysis of complex NP metabolomics data and the acceleration of NP-based drug discovery campaigns. ML algorithms can extract meaningful insights from the vast chemical space of NPs, surpassing traditional analytical methods in speed and accuracy [20]. By integrating diverse data types, including MS/MS spectra, molecular structures, and biological activities, ML approaches create predictive models that guide the prioritization of promising lead compounds [21]. Significant advances have been made in predicting molecular structures from spectral data, automating compound family classification, and enhancing screening of large chemical libraries for potential drug candidates [22,23].
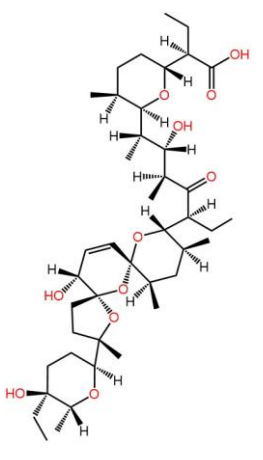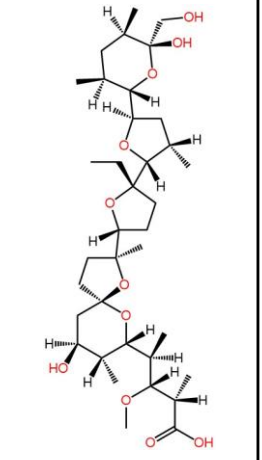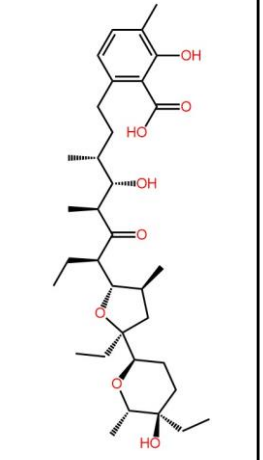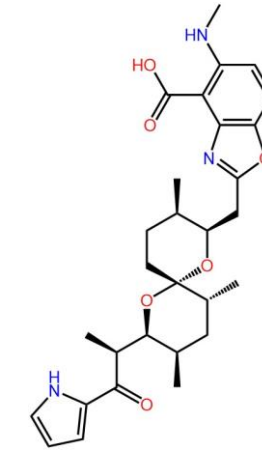
Although machine learning models for natural product analysis would ideally be trained on experimentally collected MS data, current spectral libraries are limited in both size and chemical diversity. For instance, GNPS, the largest repository, contains over 1,011,644 MS/MS spectra representing approximately 56,626 compounds as of December 17th, 2024, but this covers only a fraction of the 400,000 known NPs [11,28] Relying solely on these data risks poor generalization to experimental data not represented in the training set. To overcome this limitation, we explored the generation of *in-silico* MS2 spectra as an alternative approach to produce sufficient training data while maintaining relevance to real-world applications.

New approaches in metabolomics focus on methods independent of MS/MS databases to overcome the limitations of finite experimental data. Tools like Mass Frontier and CFM-ID4 generate *in-silico* mass spectra and expand the availability of fragmentation spectra where

experimental references are not available [29,30]. Among newer tools, SIRIUS 5 computes molecular formulas, generates molecular fingerprints (MFPs) of likely substructures, and classifies compounds based on MS/MS spectra [14,22,31,32. These *in-silico* fragmentation approaches, combined with SIRIUS's ability to generate MFPs, provide a solution to data scarcity, expanding training sets for machine learning models while maintaining relevance to NP analysis and dereplication. By leveraging these computational tools, we generate a vast array of simulated spectral data and corresponding MFPs, significantly expanding the training set for ML models beyond experimentally available spectra.

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

Current classification methods for untargeted MS data, such as CANOPUS within SIRIUS 5, utilize specific chemical or biosynthetic features to group compounds based on their structural framework or biosynthetic origins. Although valuable, these methods can be limited by their focus on all aspects of the molecular architecture, which may not strongly correlate with biological activity. As illustrated in Figure 3.1, CANOPUS classifications can vary significantly with structural modifications outside a compound's core structure, changing the predicted chemical and natural product classifications while the bioactive pharmacophore remains the same. This variability is evident in classifying polyether ionophore antifungals, where the bioactive core structure is not always the final basis for classification. The variations in structure outside the polyether core are the features prioritized such as, for instance, the "Diterpene Glycosides" and "Amino acids and derivatives" in the classification of salinomycin and calcimycin, respectively. This stands in contrast to monensin, which retains its classification as a polyether ionophore. We theorize this can be accounted for by utilizing a pharmacophore-based approach for classification, zeroing in on the functional groups and spatial arrangements essential for a compound's biological activity. With a focus on the pharmacophore, dereplication of known compounds can be expedited, prioritizing the discovery of new drugs with unknown biological effects; such an approach overcomes the limitations of current classification strategies that may overlook crucial bioactive features in favor of broader chemical or biosynthetic considerations. This paper emphasizes features driving bioactivity, providing a more direct link to therapeutic potential. We hypothesize that pharmacophore classification is less susceptible to variations in noncritical parts of the

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

Figure 3.1. CANOPUS classifications of GNPS spectra of compounds within the polyether ionophore antifungal class.

molecule, making it more robust in predicting bioactivity across structurally diverse compounds.

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

3.2   RESULTS & DISCUSSION

3.2.1 Comparison of in-silico and Experimentally Derived Fingerprints

To ensure a comprehensive evaluation of machine learning approaches, we selected 10 diverse algorithms for this study, encompassing linear models (e.g., Logistic Regression), the historically important perceptron model, tree-based models (e.g., Random Forest), support vector machines, neural networks, and ensemble methods [24−27]. These algorithms were chosen based on their proven success in similar applications, their ability to effectively process high-dimensional data, and their balance between computational efficiency and predictive performance. As the field continues to evolve, the synergy between ML and metabolomics shows great promise in overcoming traditional bottlenecks in NP drug discovery.

To train our ML models, we generated MFPs from *in-silico* MS2 spectra. This process, outlined in Figure 3.2, involved simulating compound fragmentation using Mass Frontier software, then using SIRIUS 5 to create MFPs from these simulated spectra. The classification in
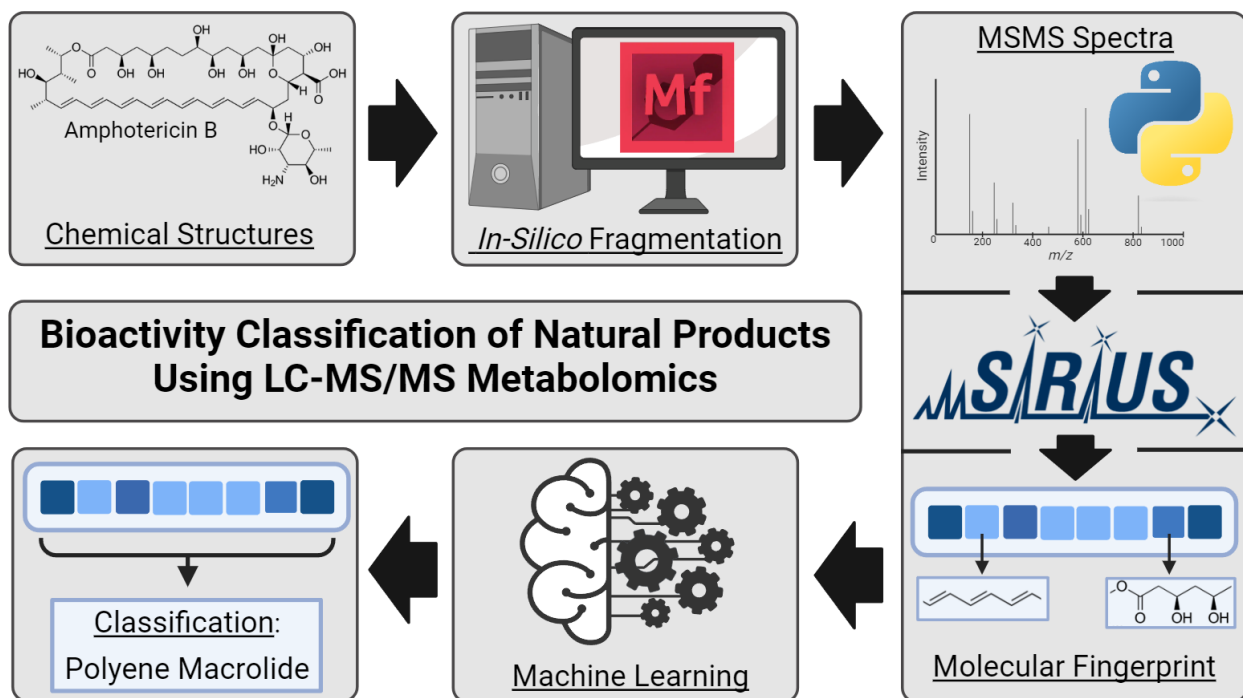
Figure 3.2. Workflow for generating *in-silico* fragmentation MS from input structures, building their MFPs, and using them to train machine learning classifiers of NP bioactivity

this study is based on pharmacophore groupings rather than hierarchical structure-based methods like ClassyFire or NP Classifier. Each class was built using 5–8 representative compounds and expanded through PubChem similarity searches (Tanimoto score ≥98), ensuring a biologically relevant classification focused on pharmacophore features rather than structural taxonomy. The resulting data set was used to train our models. To assess the similarity of the MFPs produced from *in-silico* generated mass spectra and those derived from experimental spectra, MFPs were generated with SIRIUS for 8,521 *in-silico* spectra generated using Mass Frontier and 1,256 GNPS spectra. To properly summarize the information within the high dimensional data set, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to project the high-dimensional fingerprint data onto two dimensions for cluster visualization [39]. The t-SNE projection, Figure 3.3, revealed distinct and well-separated clusters based on drug classes. The separate, cohesive clusters demonstrate that the MFPs effectively encode the unique structural features and patterns characteristic of each pharmacophore scaffold. Some MFPs, especially experimental fingerprints from GNPS spectra, were outside of their respective cluster, but by and large MFPs clustered by structural class.

Importantly, fingerprints derived from the *in-silico* fragmentation data, *in-silico* MFPs, clustered cohesively with those generated from real experimental GNPS spectral data for the same class. This colocalization highlights the high degree of integrity and accuracy with which *in-silico* MFPs effectively summarized key structural features from each of the 21 classes. This t-SNE visualization provided a powerful confirmation that MFP patterns did, indeed, distinguish between the diverse drug pharmacophores.

3.2.2 Generating Training and Testing Sets of in-silico MFPs for Polyene Macrolide Classification Model

To demonstrate our proof-of-concept on a single bioactive antifungal class, polyene macrolides were selected; the problematic characteristics of polyenes during isolation, characterization and application are well established [40−44]. Polyenes tend to be highly bioactive at low concentrations and degrade during LC-MS/MS data collection, leading to difficulties in dereplication. Polyene antifungals also present a difficult target for dereplication due to low spectra diversity and abundance in publicly available databases with only 37 spectra covering 4 unique compounds available in GNPS to date. To compile a diverse training set, eight polyene antifungals known for their antifungal activity (SI, Table S3) were expanded via PubChem similarity searches
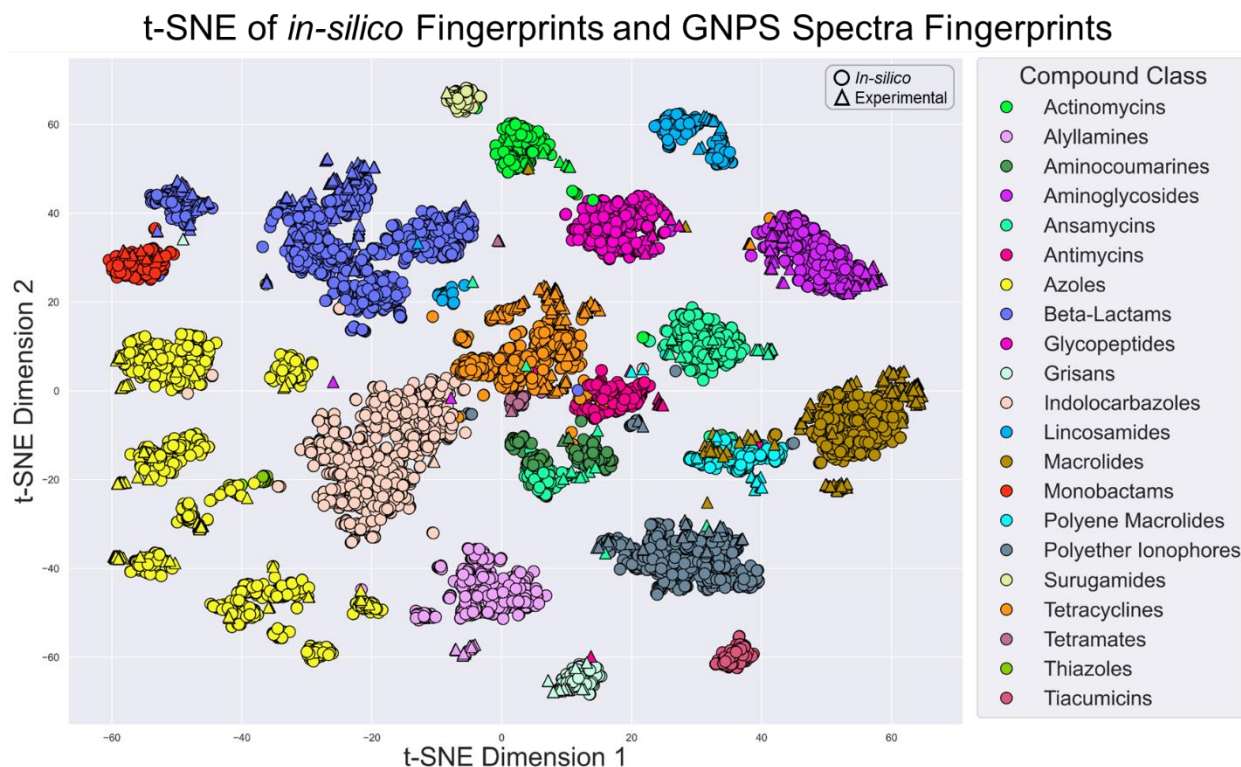
Figure 3.3. t-distributed Stochastic Neighbor Embedding (t-SNE) of *in-silico* and GNPS spectra derived molecular fingerprints. The 21 classes of bioactive compounds cluster distinctly, demonstrating that the MFPs can provide distinction based on pharmacophore. Additionally, *in-silico* and experimentally based MFPs cluster closely indicating similar structural information encoded in them.

to retrieve all similar analogues above a Tanimoto similarity score of 0.98, resulting in a set of 366 unique structures (SI, Table S7). *In-silico* fragmentation spectra were generated using Mass Frontier 8.1 and processed with SIRIUS 5 to create molecular fingerprints, forming the positive training set. For the negative data set, 2,778 compounds were selected from the RIKEN Natural Products Depository. Notably, this repository contains both natural and synthetic derivatives, characterized mechanisms of action within the MOSAIC database, and coverage of a large chemical space (SI, Figure S7); [45,46] special care was taken to avoid any overlap with polyene antifungals. The combined data set was split into an 80−20 ratio for training and testing and ten ML models were evaluated using this data set.

As seen in Table 1, each model displayed high metrics for learning the training data. The models were evaluated using accuracy, precision, recall, and F1. Accuracy measures the number of correctly identified samples over the total samples. Precision measures true positives over the total true and false positives. Recall measures true positives over the total true positive and false negatives. F1 is the harmonic mean of the precision and recall values and allows for prioritization of a model that has a balance between false positive (FP) and false negatives (FN). The top four models, the Passive Aggressive Classifier, the MLP Regressor, the Ridge Classifier, and Random Forest Classifier, achieved nearly identical accuracies and F1 scores, yet differed by recall and precision. The Passive Aggressive Classifier excelled at recall and predicting with no FN predictions. Conversely the Random Forest Classifier demonstrated high precision with no FP predictions. The Ridge Classifier and MLP Regressor both demonstrated a more balanced profile with a high F1 score despite having some FP and FN. Overall, all models displayed high levels of

Table 3.1. Results of Polyene Macrolide Trained ML Classifier Models Evaluated on Test Data (20% held-out set).

Model	Accuracy	Precision	Recall	F1	FPR (%)
Passive Aggressive Classifier	0.9968	0.9733	1.0000	0.9865	1.556
MLP Regressor	0.9968	0.9863	0.9863	0.9863	2.572
Ridge Classifier	0.9968	0.9863	0.9863	0.9863	1.707
Random Forest Classifier	0.9968	1.0000	0.9726	0.9861	0.540
Support Vector Classifier	0.9952	0.9730	0.9863	0.9796	1.129
Logistic Regression	0.9936	0.9726	0.9726	0.9726	1.669
Perceptron	0.9936	0.9859	0.9589	0.9722	0.508
K-Neighbors Regressor	0.9921	0.9474	0.9863	0.9664	2.861
Gaussian Naïve Bayes	0.9825	0.8875	0.9726	0.9281	5.484
Decision Tree Classifier	0.9825	0.9559	0.8904	0.9220	2.290

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

learning on the training set and high evaluation metrics on the 20% held-out testing data.

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

Each model was performed well on the same *in-silico* data it was trained on and generated classifications specific to polyenes, but it was most important that each model generalized to experimental spectra. Each model's performance on experimental spectra was evaluated using 37 polyene antifungal spectra from GNPS. The spectra were processed using SIRIUS 5 with either the known molecular formula, five predicted molecular formulas, or 10 predicted molecular formulas. Evaluation of spectra from known formulas provided an ideal scenario. Use of predicted molecular formulas emulated the results from untargeted metabolomics methods where the compound identity was unknown. The K-Neighbors Regressor performed best, identifying 100% of the GNPS spectra MFPs correctly as polyenes and identifying 71% and 67% of MFPs from 5-10 predicted formula, respectively. Most models showed accuracies between 78 and 95%, dropping to 45–70% when tested on MFPs from 5-10 predicted formulas (Table 3.2). These studies showed that high accuracy in MFPs based on predicted formulas is crucial for applications in

Table 3.2. Results of Polyene Macrolide Trained ML Classifier Models Tested on 37 Polyene Macrolide GNPS Spectra

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

*Results of polyene macrolide trained ML classifier models tested on 37 polyene macrolide MFPS from GNPS spectra using the known molecular formula, five predicted molecular formula, and 10 predicted molecular formula. Additionally, the performance for SIRIUS 5's tool, CANOPUS, on the same datasets.

untargeted identification since the error in mass accuracy can result in the correct molecular formula not being the one with the lowest error. Therefore, despite only one correct formula among multiple predictions, the models demonstrated flexibility in correctly identifying the pharmacophore class. As shown in Table S5 (SI), the number of fingerprints identified as polyene macrolides increased with more predicted fingerprints, surpassing the 37 fingerprints from the known formula.

3.2.3 *Generating a Data Set of in-silico Molecular Fingerprints for Training and Testing Multiclass Classification Model for 21 Bioactive Drug Classes*

Given the high performance of the binary classifier for polyenes, we next compiled a multiclass data set for 21 different classes of bioactive NPs. The classes created represented monobactams, actinomycins, beta-lactams, indolocarbazoles, cyclic peptides, azoles, tetracyclines, aminocoumarins, allylamines, tiacumicins, aminoglycosides, polyether ionophores, polyene macrolides, lincosamides, macrolides, grisans, [47,48] antimycins, ansamycins, surugamides, thiazoles, and tetramates.

To establish each drug class around a bioactive pharmacophore, between five to ten representative structures were selected for each class, which were then expanded with compounds obtained from PubChem similarity searches with a Tanimoto score of 98. Using structural

Table 3.3. Results of Bioactive Drug Class Multiclass Classification ML Models on the 20% Held Out Testing Data.

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

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

similarity increased the number of pharmacophore containing compounds. After removing duplicates, a total of 8,521 unique structures were retrieved representing the 21 drug classes (SI, Table S1). For these structures, *in-silico* MS/MS spectra were created using Mass Frontier 8.1, processed through SIRIUS 5 to generate molecular fingerprints, and split into 80/20 training/testing sets. Nine different multiclass classifiers were trained and evaluated on the held-out test set (Table 3).

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

indicating the model was easily capable of distinguishing between classes. Most other

models also demonstrated metrics over 96%, highlighting their robust multiclass predictive abilities.

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

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

Table 3.4. Results of Bioactive Drug Class Multiclassification ML Models Testing on 1,256 GNPS Spectra.

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

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

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

3.2.4   *Multiclassification Model Performance on Experimental MS/MS Data*

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

The SVC model achieved 93.58% accuracy, precision of 94.69% (which indicates a low rate of false positives), and an MCC score of 0.879 indicating high performance for all basic rates. The Ridge Classifier and SGD Classifier also exhibited strong performance, with accuracies over 89%, F1 scores around 90%, and MCC scores above 0.8. In contrast, the Decision Tree Classifier struggled considerably on this diverse GNPS data set, achieving only 73.8% accuracy; this reduced

MCC of 0.6543 indicated relatively poor predictive performance compared to the top models. Relative to CANOPUS's ability to distinguish the correct drug class, the SVC model exhibited a 30.5% increase in accuracy. Because CANOPUS evaluates compound class using the hierarchical classifications of ClassyFire and NP Classifier, we needed to map its classifications to the pharmacophore groupings used in this study. To determine what constitutes a correct classification by CANOPUS, we included all relevant hierarchical classes that could align with the assigned bioactivity class, even though some of these classes do not specifically describe a bioactive pharmacophore (SI, Table S8).

3.2.5   *Evaluating Multiclassification Models on Complex Bacterial Extracts*

To evaluate the accuracy of models on complex LC-MS/MS samples, 25 bacterial extract fractions displaying activity against *C. albicans* and suspected to contain polyenes, were selected.
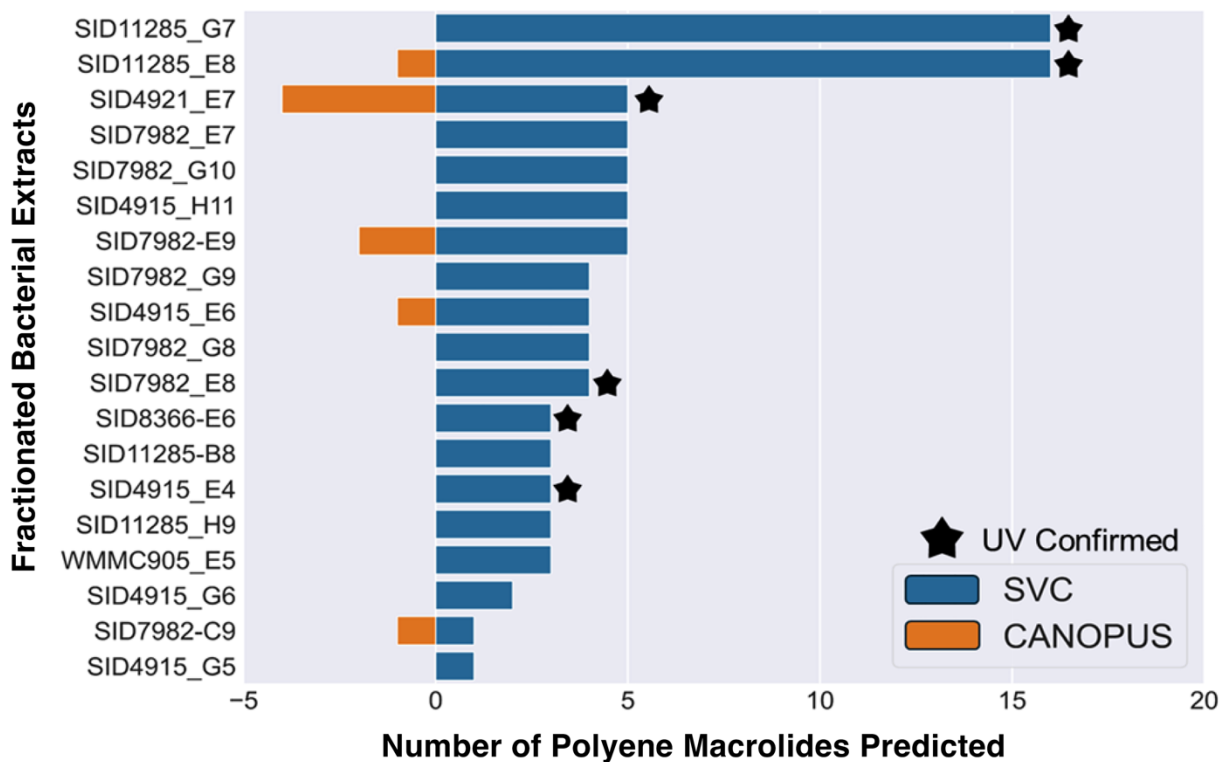
Figure 3.4. Results of the multiclassification SVC model predictions for complex bacterial extract LC-MS/MS data of bioactive antifungal samples. Samples labelled with a star are confirmed to contain polyenes based on their unique UV absorbance patterns (SI, Figure S2).

After fermentation, a two-step chromatographic approach was employed to array molecules into 96-well plates over a total of 80 fractions. The moderately purified fractions demonstrating strong activity against *C. albicans* were prepared, and then analyzed by LC-MS/MS. MFPs were compiled from a SIRIUS 5 analysis for predictions as well as the polyene classifications from SIRIUS 5's CANOPUS tool. The SVC model was chosen to evaluate all fingerprints within the LC-MS/MS data due to the highest metrics on GNPS spectra and lowest FPR. Overall, the model identified polyenes in 19 of the 25 samples (Figure 3.4). These active fractions were simultaneously analyzed by UV–vis spectroscopy to confirm the presence of polyenes using their unique UV absorbance pattern (Figure S2). The SVC model revealed a significant number of MFPs classified as polyenes; this approach revealed nearly 4 times as many "polyene samples" as the CANOPUS classifier with nearly twice as many MFPs classified on average per sample.

Discussion of Polyene Antifungal Binary Classifier Results

The results derived from the polyene classifier showed high metrics in distinguishing polyene antifungals from a diverse array of compounds. Specifically, the MLP Regressor model displayed exceptional performance in generalizing to experimental spectra and correctly identified 97.3% of polyene MFPs from known molecular formulas. This level of accuracy underscored the model's robust capability to recognize polyene molecular fingerprints. Notably, the model maintained high classification rates despite predicting from multiple predicted molecular formulas of which only one for each spectrum was correct, achieving 70% and 66% accuracy for 5 and 10 predicted formulas, respectively. The ability to correctly identify fingerprints, despite the underlying molecular formula being slightly incorrect, was considered an important feature; this enabled the model to function with MFPs from spectra with lower mass accuracies if necessary. This level of performance, especially in scenarios mimicking real-world untargeted metabolomics

workflows, highlighted the model's adaptability and its potential to streamline the identification of polyenes in complex biological matrices.

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

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

Multiclassification Model Considerations

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

The success of the multiclassification approach was largely credited to the comprehensive data set of molecular fingerprints employed during its training phase. This data set enabled the model to distinguish between subtle differences among classes effectively as reflected by low rates

of misclassification and minimized false positives. Specifically, analyses of confusion matrices highlight the fact that classes with fewer training examples—such as polyenes, monobactams, and aminocoumarins—were more susceptible to misclassification, indicating the impact of training diversity on model performance. The SVC model maintained high precision, recall, and F1 scores, suggesting that, even with a 0–1.3% occurrence of false negatives and a similar range for false positives, the model's overall classification accuracy was robust.

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

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

When the SVC model was applied to LC-MS/MS data of complex fractionated bacterial extracts, which were antifungal and suspected of containing polyenes, it was able to identify 4 times as many bacterial extracts containing polyenes as CANOPUS and nearly twice as many

MFPs as polyenes within the same samples. Validating this model's amenability to complex samples could drastically improve detection of these classes.

3.3 CONCLUSION

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

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

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

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

By facilitating rapid dereplication of known bioactive scaffolds directly from

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

3.4 MATERIALS & METHODS

3.4.1 *Utilization of PubChem for Expansion of Structure Set Using Similarity Scoring*

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

3.4.2 *Selection of GNPS Data Sets for Model Evaluation*

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

Random data sets were used for broader evaluations, specifically false positive testing and to ensure diverse representation of the GNPS chemical space. To evaluate the binary classifier's ability to identify polyene macrolides 5,000 randomly selected non-polyene spectra were used to calculate the false positive rate. While a larger data set of 9,443 random spectra was compiled to assess the specificity of the multiclassification models. Random selection was performed programmatically using Python, with a random seed to ensure reproducibility. At the time of data

set generation in April 2024, the GNPS database contained approximately 40,000 unique compounds. A non-exhaustive approach was adopted to avoid processing the entire database of around 600,000 spectra, which would have imposed significant computational burdens and potential redundancy in the data. Instead, the random selection was designed to encompass sufficient chemical diversity, capturing a broad representation of the GNPS library while avoiding unnecessary overlap.

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

3.4.3   Generating Negative Training Examples from RIKEN NP Depo

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

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

3.4.4 Mass Frontier for Batch Fragmentation of Compound Structures

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

3.4.5 Generating the MS Files for SIRIUS Using Python

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

3.4.6   Analyzing in-silico Mass Spectra with the Known Molecular Formula of Each Compound

SIRIUS 5 was employed to analyze each mass spectrum and fragmentation pattern, returning a fingerprint for each compound. The settings used for processing all the *in-silico* files were the default settings for each module. When only one formula (the known formula) was desired for prediction, the formula was specified in the MS files and the SIRIUS settings were set to only predict 1 formula.

3.4.7   Compiling in-silico Fingerprints into a Training Matrix

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

3.4.8   Utilizing the Scikit-Learn Python Package for Training and Testing Set Generation

Scikit-learn was used for creating the training and testing sets of the fingerprint data as well as for the initialization and testing of each model type. The primary metrics used in the evaluation of each machine learning model were precision, recall, F1, and accuracy. The polyene fingerprints were combined with the negative example fingerprints from the RIKEN NP Depo set and binarily labeled. The data set was split using an 80% and 20% ratio for the training and testing sets, respectively. For the multiclassification data set the MFPs were combined with the diverse negative set and labeled with their respective drug class or "negative". Each multiclassification

model was run with controlled random state and a decision function of "One-versus-Rest" or "One-versus-All".

3.4.9 LC-MS/MS Method for Bacterial Extracts

Liquid chromatography tandem mass spectrometry (LC-MS/MS) data were acquired using a Bruker maXis II Ultra-High-Resolution LC-QTOF mass spectrometer coupled to a Waters Acquity H-Class UPLC system and operated by the Bruker Hystar 3.2 software. Chromatographic gradients were performed with a mixture of methanol and water (containing 0.1% formic acid) on an RP C-18 column (Phenomenex Kinetex 2.6 μm, 2.1 mm × 100 mm) at 0.3 mL/min. The method was as follows: 0–1 min (10%–10% MeOH in H_2O), 1–12 min (10%–97% MeOH in H_2O), and 12–15.5 min (97% MeOH in H_2O). A mass range of m/z 50–1550 was measured in positive ESI mode for all spectra. The mass spectrometer was operated with the following parameters: capillary voltage of 4.5 kV, nebulizer pressure of 1.2 bar, dry gas flow of 4.0 L/min, dry gas temperature of 205 °C, and scan rate of 2 Hz. Tune mix (ESI-L low concentration; Agilent) was introduced through a divert valve at the end of each chromatographic run for automated internal calibration. MS/MS spectra were acquired at scan speeds of 2 Hz for signals above 1×104 counts and 6 Hz for signals above 1×106 counts. MS/MS spectra were collected using a stepping collision energy (CE) where CE increased linearly during MS/MS spectra collection. From time 0 to 32, the collision RF was 600, transfer time was 80, and CE was 70 eV. From time 33–66, the collision RF was 600, transfer time was 72, and CE was 100 eV. From time 67–100, the collision RF was 600, transfer time was 65, and CE was 130 eV. The precursor list was set to exclude precursor ions for 0.2 min after two spectra with the same precursor ion were acquired. Additionally, if the intensity of an excluded precursor ion rose 5-fold from the initial spectrum, it would be recollected.

3.4.10 t-SNE Visualization of Molecular Fingerprints

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

3.4.11 Fermentation for Library Generation

For each prioritized strain, 10 mL seed cultures (25×150 mm tubes) in medium DSC (5 g soluble starch, 10 g glucose, 5 g peptone, 5 g yeast extract per liter made with 50% artificial seawater) were inoculated and shaken (200 rpm, 28 °C) for 7 days. Seed cultures (2.5 mL) were used to inoculate 3×100 mL of media in 500 mL baffled flasks using two distinct media (2×100 mL ASW-A and 100 mL RAM2) containing Diaion HP20 (7% by weight). ASW-A was made using 20 g soluble starch, 10 g glucose, 5 g peptone, 5 g yeast extract, 5 g $CaCO_3$ per liter of artificial seawater; RAM2 was made using 4 g corn meal, 10 g glucose, 15 g maltose, 7.5 g Pharmamedia, 5 g yeast per liter of 50% artificial seawater. After fermentation for 7 days, the cells and HP20 were filtered using Miracloth, and the cells and HP20 were extracted with acetone (100 mL for 30 min).

3.4.12 Library Generation

The crude extract was dried and then dissolved using the following solvent mixture: 1 mL dimethyl sulfoxide (DMSO), 1 mL methanol, and 10 mL H_2O. Subsequently, the mixture was fractionated on an Isolute ENV+ (500 g cartridge) using a modified Gilson GX-271 liquid handler with 100% H_2O (10 mL), 25% CH_3OH/H_2O [fraction 1], 50% CH_3OH/H_2O [fraction 2], 75% CH_3OH/H_2O [fraction 3], 100% CH_3OH [fraction 4] (8 mL of each solvent). The 100% water fraction went directly to waste while the remaining four fractions were collected and subsequently

dried in a speedvac. Each fraction was dissolved in DMSO and subjected to HPLC using a Gilson

HPLC integrated with a Gilson 215 fitted with a 96-well plate deck capable of holding ten plates.

For HPLC, a Phenomenex Monolithic C18 column (3 mm ID X 100 mm) was used. The following

HPLC gradients were used:

Fraction 1 (F1)

0–2 min, hold at 90% H_2O/10% CH_3CN

2–14.5 min, ramp to 50% H_2O/50% CH_3CN

14.5–19 min, ramp to 100% CH_3CN

19–22 min, hold at 100% CH_3CN

22–27 min ramp to 90% H_2O/10% CH_3CN

Fraction 2 (F2) and Fraction 3 (F3)

0–2 min, hold at 90% H_2O/10% CH_3CN

2–19 min, ramp to 100% CH_3CN

19–21.5 min, hold to 100% CH_3CN

21.5–22 min, ramp to 90% H_2O/10% CH_3CN

22–27 min, hold at 90% H_2O/10% CH_3CN

Fraction 4 (F4)

0–2 min, hold at 90% H_2O/10% CH_3CN

2–5 min, ramp to 70% H_2O/30% CH_3CN

5–19 min, ramp to 100% CH_3CN

19–32 min, hold at 100% CH_3CN

32–32.5 min, ramp to 90% H_2O/10% CH_3CN

32.5–37.5 min, hold at 90% H_2O/10% CH_3CN

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

3.4.14 Computation of the Training Compound Chemical Properties

Compound properties were computationally generated using RDKit (version 2024.03.5). SMILES strings of compounds were processed to calculate a diverse set of molecular descriptors. For each compound, descriptors such as total atom count, exact molecular weight, molecular formula, clogP, topological polar surface area (TPSA), number of rotatable bonds, and Lipinski's hydrogen bond donors and acceptors were computed. Additional properties included the number of aromatic rings, fraction of sp3 carbons, QED drug-likeness score, formal charge, minimal ring count, and the Murcko scaffold framework. Input data sets, comprising positive and negative classes of compounds, were merged and preprocessed to ensure valid SMILES representations. The calculated descriptors were stored in a consolidated pandas dataframe for further analysis. These computations provided a detailed molecular profile to facilitate the assessment of compound characteristics

3.4.15 Computation of "Drug-Like" Property using Computed Chemical Properties and Lipinski's Rules

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

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

3.4.16 Evaluation of the Chemical Diversity within Each Training Class Using Tanimoto Similarity and Bemis-Murcko Frameworks

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

3.4.13 High Throughput Screening

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

3.4   DATA SUMMARY

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

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

The distribution of classes in the training data set used for the multiclassification model (Table S1); The distribution and number of GNPS spectra per class used for model evaluation (Table S2); The initial list of polyene macrolide antifungals selected for binary classifier training (Table S3); Taxonomic information on bacterial strains analyzed for polyene macrolide presence

(Table S4); Identification counts for polyene identification of GNPS spectra (Table S5); Classification equivalents between pharmacophore classifications and ClassyFire or NP classification systems (Table S8); Demonstration that utilizing mass spectra directly as machine learning input does not generalize well to experimental data (Table S9); A demonstration that changing the intensity of *in-silico* MSMS peaks does not impact molecular fingerprints derived from SIRIUS 5 (Figure S1); UV–vis confirmation of polyene UV patterns within bacterial extract LC-MS/MS data (Figure S2); Details on the bioactivity evaluation of fractionated bacterial library plates against *C. albicans* (Figure S3); Confusion matrices for ML model performance on 20% held-back training data and GNPS spectra are provided (Figure S4 and S5); A comparison of *in-silico* and GNPS spectra for polyene representative compounds using mirror matching and cosine similarity of spectra and molecular fingerprints (Figure S6); An evaluation of the diversity within all drug classes using Bemis-Murcko frameworks and average Tanimoto similarity (Figure S7). The list of GNPS spectrum IDs used for model performance evaluation in the form of a Microsoft word document (Table S6). Structure strings used to generate training and testing data sets in the form of a Microsoft excel file (Table S7). Computed chemical properties for all training compounds in a Microsoft excel workbook (Table S10). Computed chemical properties for representative structures of each class in a Microsoft excel notebook (Table S11).

3.5 REFERENCES

(1) Antibiotic Resistance Threats in the United States, 2019; Atlanta, Georgia, 2019. https://doi.org/10.15620/cdc:82532.

(2) COVID-19: U.S. Impact on Antimicrobial Resistance, Special Report 2022; Atlanta, Georgia, 2022. https://doi.org/10.15620/cdc:117915.

(3) NIAID's Antibiotic Resistance Research Framework: Current Status and Future Directions 2019. 2019.

(4) Genilloud, O. Actinomycetes: Still a Source of Novel Antibiotics. Nat Prod Rep 2017, 34 (10), 1203–1232. https://doi.org/10.1039/C7NP00026J.

(5) De Simeis, D.; Serra, S. Actinomycetes: A Never-Ending Source of Bioactive Compounds—An Overview on Antibiotics Production. Antibiotics 2021, 10 (5), 483. https://doi.org/10.3390/antibiotics10050483.

(6) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod 2020, 83 (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

(7) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod 2016, 79 (3), 629–661. https://doi.org/10.1021/acs.jnatprod.5b01055.

(8) Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. Nat Rev Drug Discov 2021, 20 (3), 200–216. https://doi.org/10.1038/s41573-020-00114-z.

(9) Gaudêncio, S. P.; Bayram, E.; Lukić Bilela, L.; Cueto, M.; Díaz-Marrero, A. R.; Haznedaroglu, B. Z.; Jimenez, C.; Mandalakis, M.; Pereira, F.; Reyes, F.; Tasdemir, D. Advanced Methods for Natural Products Discovery: Bioactivity Screening, Dereplication, Metabolomics Profiling, Genomic Sequencing, Databases and Informatic Tools, and Structure Elucidation. Mar Drugs 2023, 21 (5), 308. https://doi.org/10.3390/md21050308.

(10) Lima, N. M.; dos Santos, G. F.; da Silva Lima, G.; Vaz, B. G. Advances in Mass Spectrometry-Metabolomics Based Approaches; 2023; pp 101–122. https://doi.org/10.1007/978-3-031-41741-2_5.

(11) Wang, M.; Carver, J. J.; Phelan, V. V; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.;

Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. Nat Biotechnol 2016, 34 (8), 828–837. https://doi.org/10.1038/nbt.3597.

(12)   Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. Journal of Mass Spectrometry 2010, 45 (7), 703–714. https://doi.org/10.1002/jms.1777.

(13)   Montenegro-Burke, J. R.; Guijas, C.; Siuzdak, G. METLIN: A Tandem Mass Spectral Library of Standards; 2020; pp 149–163. https://doi.org/10.1007/978-1-0716-0239-3_9.

(14)   Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. Nat Methods 2019, 16 (4), 299–302. https://doi.org/10.1038/s41592-019-0344-8.

(15)   Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A. Dereplication of Microbial Metabolites through Database Search of Mass Spectra. Nat Commun 2018, 9 (1), 4035. https://doi.org/10.1038/s41467-018-06082-8.

(16)   van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. Topic Modeling for Untargeted Substructure Exploration in Metabolomics. Proceedings of the National Academy of Sciences 2016, 113 (48), 13738–13743. https://doi.org/10.1073/pnas.1608041113.

(17)   Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. Nat Methods 2022, 19 (7), 865–870. https://doi.org/10.1038/s41592-022-01486-3.

(18)   Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. Nucleic Acids Res 2023, 51 (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956.

(19)   Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. J Chem Educ 2010, 87 (11), 1123–1124. https://doi.org/10.1021/ed100697w.

(20)   Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Product Drug Discovery. Nat Rev Drug Discov 2023, 22 (11), 895–916. https://doi.org/10.1038/s41573-023-00774-7.

(21)   Zhang, R.; Li, X.; Zhang, X.; Qin, H.; Xiao, W. Machine Learning Approaches for Elucidating the Biological Effects of Natural Products. Nat Prod Rep 2021, 38 (2), 346–361. https://doi.org/10.1039/D0NP00043D.

(22)   Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. Bin; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural

Network-Based Structural Classification Tool for Natural Products. J Nat Prod 2021, 84 (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.

(23)   Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. Nat Biotechnol 2021, 39 (4), 462–471. https://doi.org/10.1038/s41587-020-0740-8.

(24)   Dara, S.; Dhamercherla, S.; Jadav, S. S.; Babu, C. H. M.; Ahsan, M. J. Machine Learning in Drug Discovery: A Review. Artif Intell Rev 2022, 55 (3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4.

(25)   Obaido, G.; Mienye, I. D.; Egbelowo, O. F.; Emmanuel, I. D.; Ogunleye, A.; Ogbuokiri, B.; Mienye, P.; Aruleba, K. Supervised Machine Learning in Drug Discovery and Development: Algorithms, Applications, Challenges, and Prospects. Machine Learning with Applications 2024, 17, 100576. https://doi.org/https://doi.org/10.1016/j.mlwa.2024.100576.

(26)   Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021, 2 (3), 160. https://doi.org/10.1007/s42979-021-00592-x.

(27)   Freund, Y.; Schapire, R. E. Large Margin Classification Using the Perceptron Algorithm. Mach Learn 1999, 37 (3), 277–296. https://doi.org/10.1023/A:1007662407062.

(28)   Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. J Cheminform 2020, 12 (1), 20. https://doi.org/10.1186/s13321-020-00424-9.

(29)   Thermo Fisher Scientific. Mass Frontier Spectral Interpretation Software. https://www.thermofisher.com/us/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/mass-frontier-spectral-interpretation-software.html.

(30)   Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. Anal Chem 2021, 93 (34), 11692–11700. https://doi.org/10.1021/acs.analchem.1c01465.

(31)   Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. Proceedings of the National Academy of Sciences 2015, 112 (41), 12580–12585. https://doi.org/10.1073/pnas.1509788112.

(32)   Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S. Database-

Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC. Nat Mach Intell 2020, 2 (10), 629–641. https://doi.org/10.1038/s42256-020-00234-6.

(33)   Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. J Cheminform 2020, 12 (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(34)   Huber, F.; Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Diblen, F.; Rogers, S.; Rogers, S.; Hooft, J. J. J. van der. Spec2Vec: Improved Mass Spectral Similarity Scoring through Learning of Structural Relationships. bioRxiv 2021. https://doi.org/10.1101/2020.08.11.245928.

(35)   Huber, F.; van der Burg, S.; van der Hooft, J. J. J.; Ridder, L. MS2DeepScore: A Novel Deep Learning Similarity Measure to Compare Tandem Mass Spectra. J Cheminform 2021, 13 (1), 84. https://doi.org/10.1186/s13321-021-00558-4.

(36)   Baygi, S. F.; Barupal, D. K. IDSL_MINT: A Deep Learning Framework to Predict Molecular Fingerprints from Mass Spectra. J Cheminform 2024, 16 (1), 8. https://doi.org/10.1186/s13321-024-00804-5.

(37)   Goldman, S.; Wohlwend, J.; Stražar, M.; Haroush, G.; Xavier, R. J.; Coley, C. W. Annotating Metabolite Mass Spectra with Domain-Inspired Chemical Formula Transformers. Nat Mach Intell 2023, 5 (9), 965–979. https://doi.org/10.1038/s42256-023-00708-3.

(38)   da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. Illuminating the Dark Matter in Metabolomics. Proceedings of the National Academy of Sciences 2015, 112 (41), 12549–12550. https://doi.org/10.1073/pnas.1516878112.

(39)   Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. Journal of machine learning research 2008, 9 (11).

(40)   Hamilton-Miller, J. M. Chemistry and Biology of the Polyene Macrolide Antibiotics. Bacteriol Rev 1973, 37 (2), 166–196. https://doi.org/10.1128/br.37.2.166-196.1973.

(41)   Eble, T. E.; Garrett, E. R. Studies on the Stability of Furnagillin. Journal of the American Pharmaceutical Association (Scientific ed.) 1954, 43 (9), 536–538. https://doi.org/10.1002/jps.3030430906.

(42)   Garrett, E. R. Studies on the Stability of Furnagillin. Journal of the American Pharmaceutical Association (Scientific ed.) 1954, 43 (9), 539–543. https://doi.org/10.1002/jps.3030430907.

(43)   Guruceaga, X.; Perez-Cuesta, U.; Abad-Diaz de Cerio, A.; Gonzalez, O.; Alonso, R. M.; Hernando, F. L.; Ramirez-Garcia, A.; Rementeria, A. Fumagillin, a Mycotoxin of Aspergillus Fumigatus: Biosynthesis, Biological Activities, Detection, and Applications. Toxins (Basel) 2019, 12 (1), 7. https://doi.org/10.3390/toxins12010007.

(44)   DEKKER, J.; ARK, P. A. Protection of Antibiotic Pimaricin from Oxidation and Ultraviolet Light by Chlorophyllin and Other Compounds. Antibiot Chemother (Northfield) 1959, 9 (6), 327–332.

(45)   Kato, N.; Takahashi, S.; Nogawa, T.; Saito, T.; Osada, H. Construction of a Microbial Natural Product Library for Chemical Biology Studies. Curr Opin Chem Biol 2012, 16 (1–2), 101–108. https://doi.org/10.1016/j.cbpa.2012.02.016.

(46)   Nelson, J.; Simpkins, S. W.; Safizadeh, H.; Li, S. C.; Piotrowski, J. S.; Hirano, H.; Yashiroda, Y.; Osada, H.; Yoshida, M.; Boone, C.; Myers, C. L. MOSAIC: A Chemical-Genetic Interaction Data Repository and Web Resource for Exploring Chemical Modes of Action. Bioinformatics 2018, 34 (7), 1251–1252. https://doi.org/10.1093/bioinformatics/btx732.

(47)   Quintana, R. P.; Lasslo, A.; Boggs, P. P. Grisan Derivatives in Pure and Mixed Monomolecular Films. J Colloid Interface Sci 1968, 26 (2), 166–174. https://doi.org/https://doi.org/10.1016/0021-9797(68)90309-3.

(48)   Cacho, R. A.; Chooi, Y.-H.; Zhou, H.; Tang, Y. Complexity Generation in Fungal Polyketide Biosynthesis: A Spirocycle-Forming P450 in the Concise Pathway to the Antifungal Drug Griseofulvin. ACS Chem Biol 2013, 8 (10), 2322–2330. https://doi.org/10.1021/cb400541z.

(49)   Chicco, D.; Jurman, G. The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. BioData Min 2023, 16 (1), 4. https://doi.org/10.1186/s13040-023-00322-4.

Chapter 4: Pharmacophore-Based Machine Learning for Rapid Opioid Detection in High-Resolution LC-MS/MS Metabolomics

4.1 INTRODUCTION

The opioid crisis is among the most severe public-health emergencies of the 21st century, driving a sharp increase in overdose fatalities worldwide. In the United States alone, opioid-related deaths reached 114,000 in 2023 and 87,000 in 2024, making opioids the leading cause of mortality for individuals aged 18–44 [1]. Additionally, the trafficking of both natural and synthetic opioids has climbed steadily since 2014, with fentanyl seizures nearly tripling between 2021 and 2023 [2]. The widespread misuse of these illicit opioids, including heroin, fentanyl, and nitazenes, has strained healthcare systems and public safety initiatives [3-7]. A major challenge in addressing the crisis is the proliferation of novel synthetic opioids, which often evade conventional detection methods [8-10]. Illicit drug manufacturers continually modify chemical structures to bypass legal restrictions and standard drug screening techniques, which underscores the pressing need for innovative approaches to improve opioid detection and characterization [11-12].

Traditional drug testing methodologies, including immunoassays, gas chromatography-mass spectrometry (GC-MS), and liquid chromatography-mass spectrometry (LC-MS), remain the backbone of drug screening in forensic and clinical toxicology [13-17]. While these techniques provide robust analytical capabilities, each modality has notable limitations when applied to the detection of novel or structurally diverse synthetic opioids. Immunoassays, for instance, are highly specific to their target drug and often fail to cross-react with newer or less common substances [15]. Although GC-MS and LC-MS/MS deliver high accuracy, they rely on matching to existing spectral libraries and thus fail when presented with unreferenced analogs [14,16].

One of the primary limitations of these conventional methods is their inability to identify unknown or novel synthetic opioids. Database-dependent approaches inherently lag behind the pace of emerging illicit synthetic opioids [18,19]. Many emerging fentanyl and nitazene analogs lack reference fragmentation spectra, resulting in undetected or misidentified substances during routine screening [18–21]. Large-scale spectral searches also incur heavy computational costs and can yield ambiguous hits for isomeric compounds. Together, these limitations motivate the development of complementary, database-independent detection methodologies to supplement these conventional methods.

Machine learning (ML) has emerged as a transformative tool in toxicology, metabolomics, and mass spectrometry-based chemical analysis [22-27]. ML algorithms excel at detecting patterns in high-dimensional datasets, making them particularly well-suited for analyzing complex LC-MS/MS data. In recent years, ML-driven approaches have been applied to a wide range of applications, from disease biomarker discovery and toxicological quantitative structure-activity relationship (QSAR) modeling to untargeted metabolomics profiling [24,26]. More recently, we demonstrated the ability of ML models to predict structural classes of antimicrobial natural products [28]. Given our success, we hypothesized that a similar approach would be applicable to opioids and provide a route to identifying novel synthetic opioids not present in databases.

Our approach leveraged ML to develop a pharmacophore-based opioid detection framework that integrated high-resolution LC-MS/MS data with computationally generated molecular fingerprints. We generated an *in-silico* library of fragmentation spectra for morphinan, fentanyl, and nitazene cores. Using SIRIUS 5, each spectrum was converted into a MACCS style molecular fingerprint encoding key substructures; these fingerprints then served as features for the classification models [29-32]. These ML models were trained on both *in-silico* and experimental

data, allowing them to generalize effectively to real-world opioid detection scenarios. The performance of the trained models was evaluated using experimental datasets from the Global Natural Products Social Molecular Networking (GNPS) database and the Centers for Disease Control and Prevention (CDC) Fentanyl Analog Screening (FAS) kit [33-36].

In this study, we present a robust, ML-driven framework which improved opioid detection in high-resolution LC-MS/MS clinical metabolomics data, irrespective of current spectral libraries. We demonstrate that ML classifiers can reliably detect morphinan, fentanyl, and nitazene opioids and even analogs and metabolites in complex clinical metabolomic samples. This approach not only enhances current forensic and clinical toxicology workflows but also establishes a generalizable and scalable blueprint for future ML applications in computation toxicology and metabolomics.

4.2 RESULTS & DISCUSSION

4.2.1 Compilation of Opioid Class Datasets

To construct robust machine learning (ML) classifiers for morphinan, fentanyl, and nitazene opioids, we first compiled core compound sets for each class with verified biological relevance, drawing from literature and curated sources (Supp Table A.1, Supp Table A.2, Supp Table A.3). The morphinan core compounds included canonical opioids such as morphine, codeine, and hydrocodone, while the core fentanyl and nitazene compounds were composed of 119 and 61 compounds, respectively, representing the breadth of structural analogs documented in public health and forensic literature.

Using PubChem, each opioid compound dataset was expanded, resulting in 1,651 fentanyl analogs and 650 morphinan compounds (Supp Table A.4, Supp Table A.5). Due to structural divergence and a limited number of suitable analogs, the nitazene class was retained at the 61 core compounds (Supp Table A.3).

To train classifiers capable of distinguishing opioids from other bioactive compounds, we generated a comprehensive negative set using the RIKEN Natural Products Depository (NP Depo), which encompasses 10,959 structurally diverse natural products [37]. This diversity was essential for promoting generalizability and reducing model bias.

4.2.2 In-silico Spectral Generation and Fingerprinting

In-silico MSMS spectra were generated for all compounds using Mass Frontier 8.1 software and converted to MSMS spectra. Subsequently, SIRIUS 5 was employed to generate molecular fingerprints (MFPs) from the *in-silico* spectra, capturing substructural features crucial for the ML-based classification. We previously showed that MSMS spectra were not suitable for ML due to the high variability in fragmentation patterns from small structural deviations between related compounds [28].

4.2.3   Visualizing Chemical Space and Classification Disparities

To visualize the information within the high dimensional data set and verify the similarity of the samples within each opioid class, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to project the high-dimensional fingerprint data onto two dimensions for cluster visualization. The t-SNE projection in Figure 4.1 revealed distinct and well-separated clustering of the fingerprints, within the chemical space of the dataset, according to the three opioid classes.
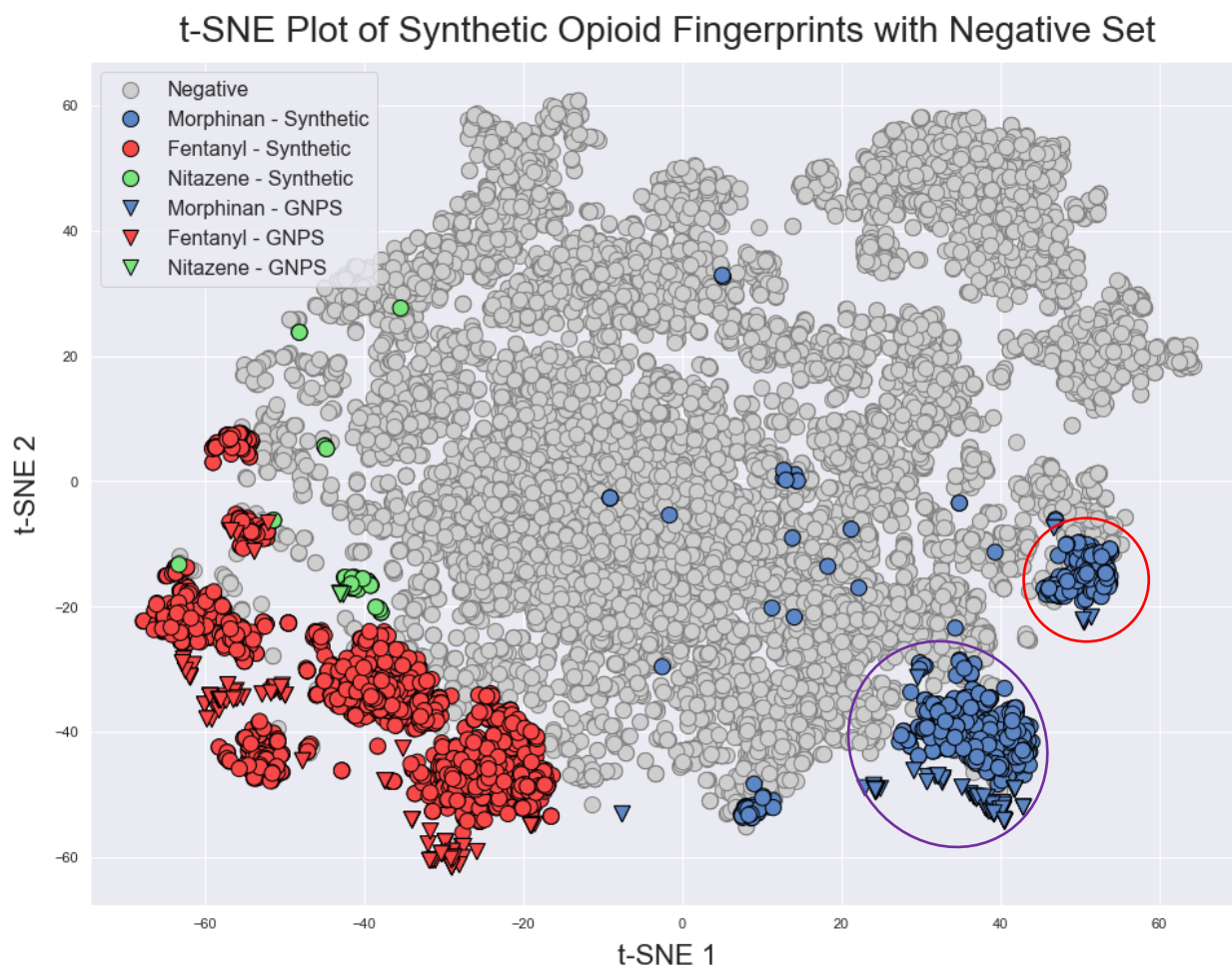
Figure 4.1. *t-Distributed Stochastic Neighbor Embedding (t-SNE) projection of molecular fingerprints derived from in-silico and GNPS MS/MS spectra.* The three opioid classes—morphinan, fentanyl, and nitazene—form distinct, non-overlapping clusters, indicating that the molecular fingerprints capture pharmacophore-specific structural information. Notably, fingerprints generated from in-silico spectra closely align with those derived from experimental GNPS data, demonstrating the consistency and representational accuracy of the in-silico fingerprinting approach.

The separate, cohesive clusters proved that the MFPs effectively encoded the unique structural features and patterns characteristic of each opioid pharmacophore scaffold. Additionally, clusters of compounds within each group represented unique pockets of chemical diversity relating to substructural features. For example, the cluster of morphinan opioids outlined in red in Figure 4.1 compose a cluster of compounds related to buprenorphine, with cyclopropane to cyclopentane terminated alkyl tails (Supp. Figure B.7). The group of morphinan opioids outline in purple in Figure 4.1 corresponds to a more prototypical morphinan opioid with different derivatives of morphine with small alkyl tails and different substitutions on the 3 and 6 position hydroxyl groups (Supp. Figure B.7). Finally, an important association in this clustering is that the experimental GNPS molecular fingerprints, shown as triangles in Figure 4.1, cluster closely to their *in-silico* generated molecular fingerprint counterparts, shown as circles. This demonstrates that the chemical structure information encoded in the molecular fingerprints for both the experimental and *in-silico* data is comparable despite the orthogonal sources.

4.2.4   *Model Training and Evaluation on In-silico Data*

A suite of ten ML model types—including logistic regression, support vector machines, random forests, and multilayer perceptron models—were trained using the compiled *in-silico* MFPs. Performance was benchmarked using accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC). Accuracy measures the number of correctly identified samples over the total samples. Precision measures true positives over the total true and false positives. Recall measures true positives over the total true positive and false negatives. F1 is the harmonic mean of the precision and recall values and allows for prioritization of a model that has a balance between false positive (FP) and false negatives (FN). Once evaluated, the top performing model was selected and optimized for performance to maximize the accuracy and minimize the false

positive rate of each model. Each model was trained on 80% of the respective dataset and tested on the held-back, previously unseen, 20%. As can be seen by the top performing models for each opioid type in Table 1, each model demonstrated accuracy on the held-back set above 99% with F1 and MCC score above 95%. This shows the models are balanced with both high precision and high recall, representing low rates of both false positive and false negative. Overall, all models displayed high levels of learning on the training set and high evaluation metrics on the 20% held-out testing data.

To test the specificity of classification and ensure that the models can accurately distinguish between opioid and non-opioid MFPs, each model was tested against 9443 randomly selected spectra from GNPS. Each model was evaluated to determine how likely they falsely predicted a compound as an opioid; the false positive rate (FPR). The FPR of the models ranged from 0% to 0.73%, with the morphinan opioid classifier showing the highest 0.73% FPR and the nitazene demonstrating an impressive 0% false positive rate (Table 1).

Table 4.1. Results of Opioid Trained Binary ML Classifier Models Evaluated on Test Data (20% Held-out Set) *

Model	Accuracy	Precision	Recall	F1	MCC	FPR (%)
Morphinan Opioids: K-Neighbors Classifier	0.995279	0.944056	0.978261	0.960854	0.958515	0.731552
Nitazene Opioids: Random Forest Classifier	0.998058	1	0.909091	0.952381	0.952518	0
Fentanyl Opioids: Support Vector Regressor	0.998476	0.996671	0.998823	0.997746	0.996596	0.084782

*Results of binary opioid ML models on test set data using accuracy, precision, recall, F1 score, and MCC. The false positive rate of each model when tested on 9443 random GNPS spectra.

4.2.5 Experimental Validation with GNPS and FAS Kit Libraries

To evaluate the generalizability of the opioid classifiers beyond computationally generated data, their performance was assessed on experimental MS/MS spectra obtained from two distinct sources: the Global Natural Products Social Molecular Networking (GNPS) database and the CDC's Fentanyl Analog Screening (FAS) kit. For the GNPS dataset, each model was tested on a curated subset of experimental spectra specific to its target opioid class—204 morphinan spectra, 207 fentanyl spectra, and 6 nitazene spectra—each derived from spectra containing known molecular formulas (Supp Table A.6, Supp Table A.7, Supp Table A.8). The models were tested on the MFPs from the GNPS spectra to determine the accuracy on experimental opioid spectra (Table 4.2). Additionally, to ensure rigorous evaluation and prevent model overprediction, each GNPS set was combined with MFPs from an additional 190 randomly selected GNPS spectra representing non-opioid compounds (Supp Table A.9) and evaluated using accuracy, precision, recall, F1, and MCC.

Against the experimental MFPs, the models displayed robust predictive accuracy and specificity, as seen in Table 4.2. All three classifiers demonstrated outstanding performance on experimental MFP data. Across morphinan, fentanyl, and nitazene classes, the models consistently achieved >97% accuracy on GNPS spectra and >98% overall accuracy. Aggregate F1 and Matthews correlation coefficients exceeded 0.95 for each classifier, with the nitazene model attaining perfect scores.

These results were further supported by evaluations using the CDC FAS Kit Library, which contains 13 morphinan, 416 fentanyl, and 6 nitazene experimental spectra. The classifiers maintained consistent high performance on this independent dataset, with accuracy values exceeding 95% for all three opioid classes.

Overall, to assess real-world applicability, we evaluated the classifiers on two independent

Table 4.2. Results of Opioid Trained Binary ML Classifier Models Evaluated on GNPS Experimental Spectra Derived Molecular Fingerprints *

Model	Only GNPS Accuracy	Accuracy	Precision	Recall	F1	MCC	FAS Kit Library
Morphinan Opioids: K-Neighbors	0.98170	0.99527	0.94405	0.97826	0.96085	0.95851	1
Nitazene Opioids: Random Forest	1	1	1	1	1	1	1
Fentanyl Opioids: Support Vector Regressor	0.97525	0.98557	0.99463	0.95854	0.97623	0.9662	0.96778

*Results of opioid trained ML classifier models tested on GNPS spectra using the known molecular formula. The morphinan, fentanyl, and nitazene models were evaluated on 204, 207, and 6 molecular fingerprint respectively, all derived from GNPS spectra of that class. Additionally, the models were evaluated using the FAS Kit Library which contains 13 morphinan opioid spectra, 416 fentanyl opioid spectra, and 6 nitazene opioid spectra.

experimental MS/MS collections: curated GNPS spectra (opioid vs. non-opioid subset) and the CDC FAS kit. In both cases, the models sustained high accuracy (>95%), F1 scores, and MCC values, confirming that they generalize beyond *in-silico* training data. Inclusion of unrelated GNPS compounds demonstrated that performance gains arise from learning class-specific fragmentation features rather than overfitting to generic spectral patterns.

4.2.6   *Identification of Known Drugs and Metabolites*

The classifiers were then tested on the GNPS Drugs and Metabolites Library to determine if models trained on *in-silico* and canonical opioid spectra could recognize structurally diverse analogs and metabolic byproducts. This benchmark is especially valuable because many of these compounds are under-represented in existing spectral libraries and frequently evade traditional, database-reliant identification methods.
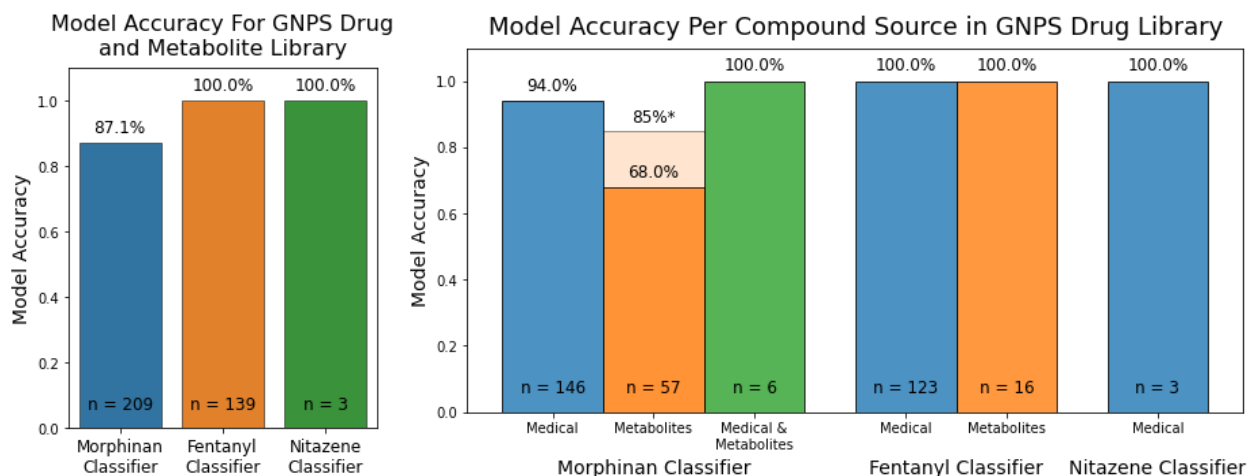
Figure 4.2. *Classification performance of opioid-specific machine learning models on the GNPS Drugs and Metabolites Library.* Each model was evaluated on experimental MS/MS spectra representing known pharmaceutical opioids and their metabolites. Bars indicate classification accuracy for compounds grouped by annotation: medical drugs, drug metabolites, or compounds classified as both.

*This accuracy is excluding spectra with putative, unconfirmed structures.

Each classifier performed strongly on its target class in the GNPS Drugs and Metabolites Library (Figure 4.2). The overall accuracy for the morphinan model was 87%. Breaking this down by compound type, it correctly identified 94% of parent pharmaceuticals (n = 146), 68% of known metabolites (n = 57), and 100% of spectra labeled as both drug and metabolite (n = 6) (Supp. Table A.10). The only missed metabolite identifications were for nalfurafine, hydroxy-butorphanol, naltrexone, and naloxone—several of which are putative metabolites without confirmed structures. Excluding these putative entries raises metabolite accuracy to 85% (n = 46) and overall model accuracy to 92.4%. This adjustment highlights the classifier's robust performance on validated metabolite spectra. Collectively, these results confirm the model's reliable detection of both parent opioids and their metabolic derivatives.

The fentanyl classifier achieved 100% overall accuracy on the GNPS Drugs and Metabolites Library. When broken down by compound type, it achieved 100% accuracy on medical fentanyl analogs (n = 123) and 100% accuracy on known fentanyl metabolites (n = 16)

(Supp Table A.11). This highlighted the model's strong generalization capabilities, particularly its robustness in identifying metabolites that may differ from the original compounds.

The nitazene classifier achieved perfect accuracy (100%) on the GNPS Drugs and Metabolites nitazene spectra. Although the nitazene subset was small (n = 3), these compounds represented medical opioids, and the classifier correctly identified all instances without error (Supp Table A.12). Furthermore, this lack of representative nitazene opioids in GNPS, one of the largest spectral repositories, shows the scarcity of available reference data for this emerging drug threat.

Overall, the high performance of each classifier established the capability to identify both parent pharmaceuticals and metabolites. Since metabolites of opioids are consistently excluded from spectral libraries, these models can supplement identifications.

4.2.7   *Application of Opioid Classifiers to Clinical Blood and Urine Samples for Drug Testing*

To evaluate the clinical applicability and performance of the developed opioid classifiers, we tested the models on 332 anonymized clinical blood and urine samples provided by the Wisconsin State Laboratory of Hygiene through collaboration with Dr. Heather Barkholtz. Samples were obtained from routine drug testing and were analyzed using data-independent acquisition (DIA) LC-MS/MS.

Across the analyzed clinical samples, a total of 14,775 features were extracted, and their corresponding fragmentation spectra were processed using SIRIUS 5, resulting in 35,622 predicted MFPs. Each classifier (morphinan, fentanyl, and nitazene) was then applied to the complete set of 35,622 MFPs to identify candidate opioid features specific to each opioid class. Features predicted as positive were further examined to confirm or identify potential opioid identities using a multi-tiered spectral analysis strategy based on modified cosine similarity scoring (See experimental).
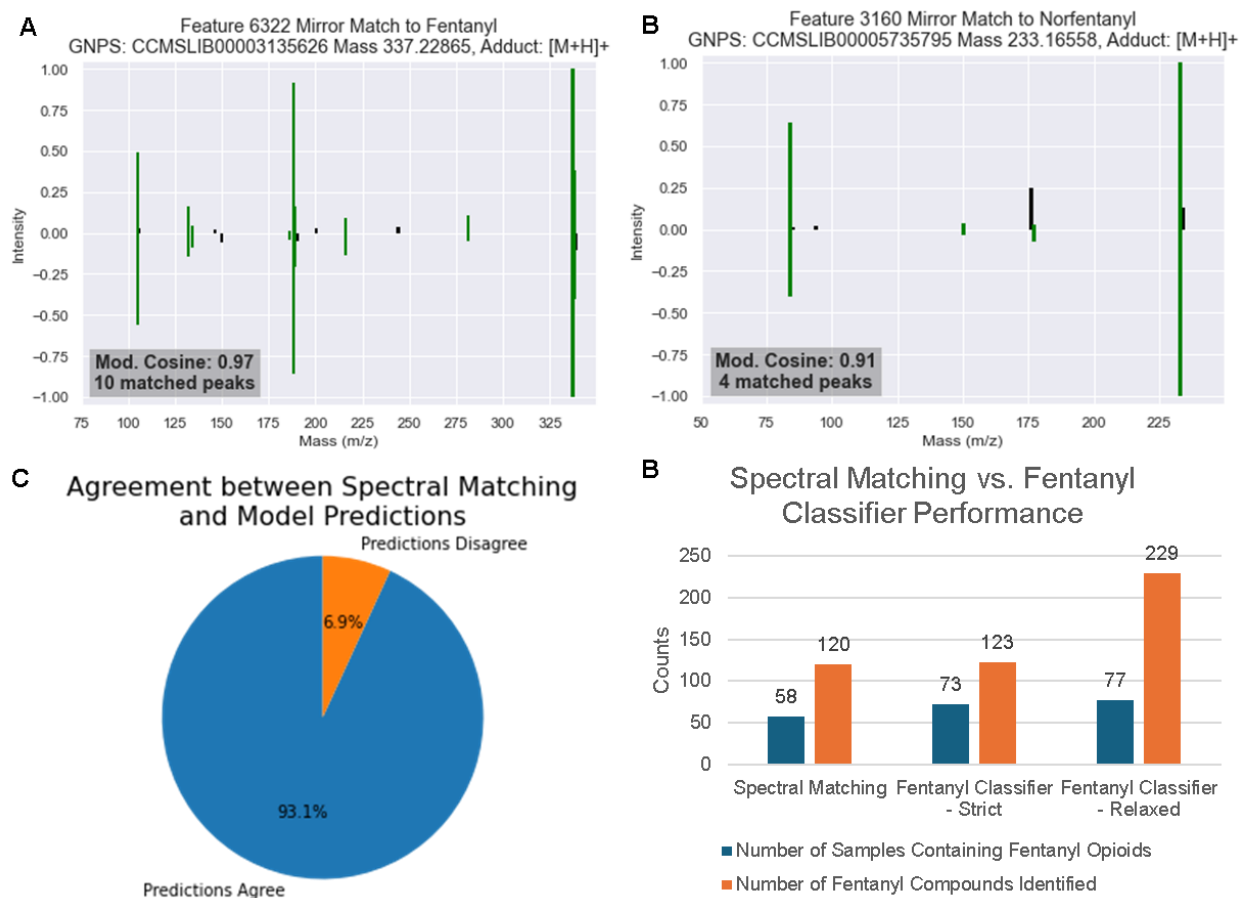
Figure 4.3. *Identification performance of the fentanyl opioid classifier on the 332 clinical samples.* (A/B). Demonstrating the identified features match to known fentanyl opioids, fentanyl and norfentanyl, with high spectral matching. (C). The agreement of identifications between the spectral database searching and fentanyl classifier. (D) A bar graph demonstrating that the sample and compound identification of the fentanyl opioid model exceed those done by spectral database searching.

4.2.8 Fentanyl Opioid Identification in Clinical Samples

The fentanyl classifier flagged 64 MFPs corresponding to 38 unique MS/MS features. SIRIUS 5 structural annotation of these features suggested the presence of fentanyl, norfentanyl, β-hydroxyfentanyl, and 4-hydroxyfentanyl, supporting the classifier's initial predictions. To validate these assignments, we applied a modified cosine similarity search against GNPS reference spectra. Two features matched confidently to fentanyl (cosine 0.97) and norfentanyl (cosine 0.91) (Figure 4.3), yielding 123 fentanyl identifications across 73 clinical samples.

Expanding the cosine threshold to include close analogs increased identifications to 229 feature-compound pairs across 77 samples (Supp Table A.14). In parallel, a conventional spectral-library search detected 120 fentanyl compounds in 58 samples (Supp Table A.13). Thus, our classifier revealed fentanyl in 25 % more samples (73 vs. 58) and—when analogs were counted—delivered a 33 % sample-level increase (77 vs. 58) and nearly 90% more compound identifications (229 vs. 123) compared to database matching (Supp Fig B.3).

Comparing both methods on overlapping samples revealed 93.1% concordance (Figure 4.3). The few discrepancies—four fentanyl and one fluorofentanyl missed by the classifier—stemmed from an intensity threshold filter and an erroneous SIRIUS 5 formula prediction, respectively. Overall, the strong agreement with spectral matching and the additional detections uniquely captured by our ML approach underscore its robustness and its ability to extend beyond the limits of traditional library-based workflows.

4.2.9 *Morphinan Opioid Identification in Clinical Samples:*

The morphinan classifier predicted 157 molecular fingerprints corresponding to morphinan-class opioids, representing 127 unique MS/MS features. SIRIUS 5 structural annotations for these features predicted several clinically relevant morphinan compounds and metabolites, including Levorphanol, Dextromethorphan, Norlevorphanol, Naloxone, Hydrocodone, and Codeine. To validate the morphinan classifier predictions, we again used modified cosine similarity scoring to compare against the GNPS database. Five features were exact matches to known morphinan opioids: Dextromethorphan, Levorphanol/Dextrorphan, 3-hydroxymorphinan, Hydrocodone/Codeine, and Naloxone. Notably, Dextromethorphan and Hydrocodone/Codeine demonstrated exceptionally high cosine scores of 1.0 and 0.99, respectively (Figure 4.4, Supp Table A.15). These exact matches accounted for 96 morphinan-class

identifications across 67 clinical samples (Supp Fig B.4). By contrast, traditional spectral-library matching identified only 40 morphinan compounds in 33 samples. Thus, the classifier doubled sample coverage (67 vs. 33) and increased confirmed compounds by 2.4-fold (96 vs. 40).

When expanding the model identification to high-scoring structurally related compounds, the number of detected compounds rose to 282 morphinan-related compounds in 131 samples (Supp. Fig B.5; Supp. Table A.15). When compared to the spectral matching results, with all high-scoring matches, the model achieved a six-fold increase in compound identifications (282 vs. 40)
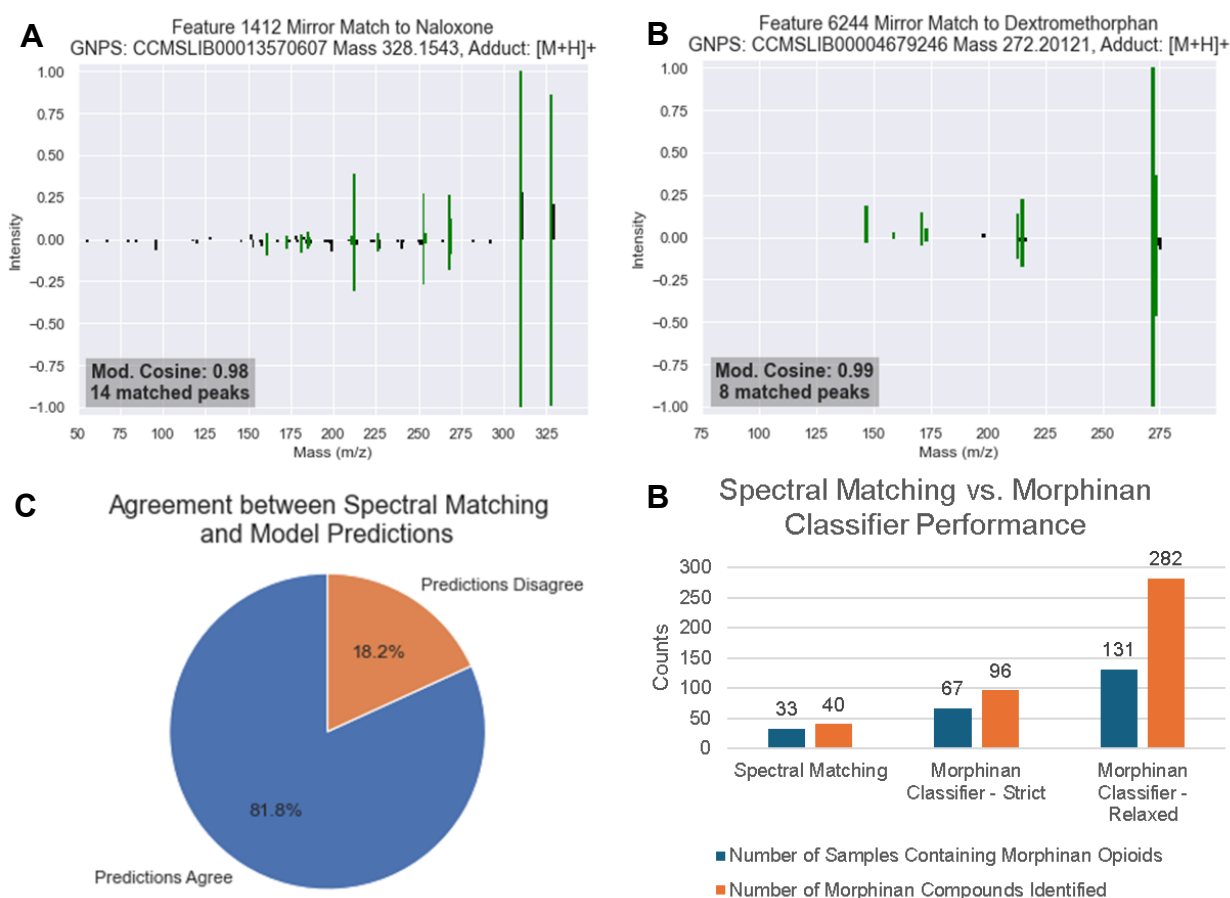
Figure 4.4. *Identification performance of the morphinan opioid classifier on the 332 clinical samples.* (A/B). Demonstrating the identified features match to known morphinan opioids, Naloxone and Dextromethorphan, with high spectral matching. (C). The agreement of identifications between the spectral database searching and morphinan classifier. (D) A bar graph demonstrating that the sample and compound identification of the morphinan opioid model exceed those done by spectral database searching.

and a four-fold increase in sample coverage (131 vs. 33). These gains reflect the classifier's ability to leverage averaged DIA-MS spectra—capturing low-intensity features—and to detect metabolites absent from standard libraries (e.g., dextrorphan sulfate, dextrorphan glucuronide, 6-oxycodol) (Supp Table A.15).

The classifier reproduced 81.8 % (27 of 33) of the morphinan-positive samples originally detected by spectral matching, demonstrating substantial overlap in true-positive identifications (Figure 4.4). The six missed identifications comprised two buprenorphine/norbuprenorphine pairs, two morphine spectra, and one naltrexone spectrum; these features fell below our preset intensity threshold and therefore were not submitted for SIRIUS 5 MFP prediction. Importantly, the classifier also flagged 131 samples beyond the database searching method—many containing low-abundance compounds or metabolites such as dextrorphan sulfate and glucuronide conjugates—that spectral matching failed to detect (Supp. Table A.15). The high degree of consistency between our model-based identifications and those independently obtained by the spectral matching provides strong supporting evidence for the validity of the model's broader set of opioid identifications.

4.2.10 Nitazene Opioid Identification in Clinical Samples:

The nitazene classifier predicted 78 MFPs to be associated with nitazene opioids across 62 MS/MS spectra. SIRIUS 5 annotations of the MS/MS spectra led to no known nitazene structures, although some predictions did contain a benzimidazole core indicating a similar core chemistry as nitazenes.

However, when the MS/MS spectra were subjected to a modified cosine score search; no relevant matches were found within the GNPS database, indicating that the predicted compounds did not correspond to any known nitazene opioids. The classifier predictions were also compared

to database matching results, revealing a significant disparity, with 90% of the nitazene spectral matches being missed by the model. Only a single feature identified by our model was also detected in the library identifications, however it does not match to any nitazenes.

Although the nitazene classifier did not generalize as expected in clinical sample analysis, the results emphasize the need for further refinement. The model's performance on the GNPS and FAS Kit data show the promise of the model on curated data. However, the model may require additional training data, considering the current positive training examples of only 61, or additional tuning to better capture the spectral features that are unique to nitazene compounds. Furthermore, the nature of DIA results in MSMS spectra with unrelated or missing peaks when compared to those collected using DDA, therefore the spectral data for this uncommon opioid class may be insufficient for SIRIUS 5 to encode a consistent molecular fingerprint for our model to identify. Further adjustments to the dataset and utilizing data-dependent acquisition (DDA) LC-MS/MS might improve the specificity and sensitivity of the classifier for this opioid class.

4.3 CONCLUSION

In this study, we developed a machine learning (ML)-based framework for identifying morphinan, fentanyl, and nitazene opioids from high-resolution LC-MS/MS data. By leveraging *in-silico* fragmentation data and computationally generated molecular fingerprints (MFPs), we built robust classification models that demonstrated excellent performance across a variety of datasets, including *in-silico* data, experimental GNPS spectra, and clinical samples.

The morphinan and fentanyl classifiers consistently exhibited high accuracy, with precision, recall, and Matthew's correlation coefficient (MCC) scores exceeding 95%. Notably, the performance of these classifiers on clinical blood and urine samples surpassed traditional spectral database search results, with significant increases in compound identification and sample coverage ranging from 90-600% and 33-400%, respectively. These results underscore the potential of ML-based approaches to enhance opioid detection, even within the complex and diverse matrices of clinical samples.

While the nitazene classifier showed promising potential in experimental settings, it underperformed when applied to clinical samples. This performance indicaties a need for improved training data, advanced tuning techniques, and possibly a shift to data-dependent acquisition (DDA) LC-MS/MS to enhance spectral data quality and model specificity.

Overall, these findings demonstrate the substantial benefits of ML-based opioid detection models in practical settings, delivering greater sensitivity and specificity than conventional spectral-library searches. As synthetic opioids continue to evolve, our data-driven classifiers can flag both established molecules and emergent analogs that evade standard workflows. To facilitate broad adoption, we are packaging these models into a user-friendly software suite—complete with a graphical interface—so that forensic and clinical laboratories can deploy the classifiers without extensive computational expertise. Further work will focus on refining the nitazene classifier,

expanding the model's ability to generalize across a broader spectrum of opioid analogs and metabolites, and identifying new data sources to optimize training sets for computational toxicology applications.

4.3 MATERIALS & METHODS

4.3.1 Blood and Urine Sample Preparation

The procedure for blood and urine sample preparation has been described in detail by Bates et al. (2024) [16]. A 100 µL aliquot of each authentic specimen was added to a 96-well Agilent Captiva Enhanced Matrix Removal-lipid cleanup plate, which was positioned over a 96-well collection plate containing glass inserts. To each well, 10 µL of internal standard was added, followed by a 5-minute equilibration period. A crashing solvent consisting of ice-cold 15:85 (v/v) methanol:acetonitrile (400 µL) was then added to each well, and the plate was vortexed. Samples were processed using a Waters 96-well positive-pressure manifold, initiating the elution process at 0.5–1 psi and gradually increasing to 15 psi. Following this, 200 µL of a 1:4 (v/v) methanol:water solution was added, and the samples were vortexed and eluted under the same pressure conditions until the extraction plate was dry. The collected eluates were dried in glass inserts using an Organomation microplate evaporator (N_2, 30 °C), reconstituted in 150 µL of a 1:1 (v/v) methanol:water solution, and vortexed for thorough mixing. The samples were then centrifuged at 3000 rpm for 5 minutes before being transferred to the instrument's sample manager for analysis.

4.3.2 UHPLC and LC-MS/MS Instrument Parameters

Ultrahigh-performance liquid chromatography (UHPLC) separation was carried out on a Waters Acquity HSS C18 column (2.1 mm × 150 mm, 1.8 µm particles), maintained at 50 °C. All gases were set according to the manufacturer's recommendations, and the sample temperature was maintained at 15 °C. A 5 µL injection volume was employed with a flow rate of 0.400 mL/min. The mobile phases consisted of A1 (5 mM ammonium formate, pH 3.0), B1 (acetonitrile with 0.1% formic acid), A2 (water with 0.001% formic acid), and B2 (acetonitrile with 0.001% formic acid). A gradient was applied for the positive-ion acquisition mode: 13% B1 for 10 minutes,

followed by 50% each of A1 and B1 for 0.75 minutes, then 95% B1 for 1.5 minutes, and finally returning to initial conditions. For negative-ion acquisition, a gradient with A2 and B2 was employed, beginning with 13% B2 for 4.5 minutes, followed by 95% B2 for 1 minute, and returning to initial conditions.

Mass spectrometry was conducted using a Waters Xevo G2-XS QToF system with an electrospray ionization (ESI) source, operating in both positive and negative ionization modes. The ESI conditions for positive-ion mode were set as follows: capillary voltage 0.80 kV, sample cone 25 V, and cone gas flow 20 L/h. For negative-ion mode, the capillary voltage was set to 1.50 kV, the sample cone to 40 V, and cone gas flow to 50 L/h. Both source and desolvation temperatures were 150 °C and 400 °C, respectively, with desolvation gas flow at 800 L/h.

Data-independent acquisition mode (MSE) was used, employing a low collision energy of 6 eV to minimize fragmentation, followed by ramped high collision energies (10–40 eV) for optimal fragment ion generation. Precursor and fragment ion data were collected from 40 to 1000 m/z. Lock mass calibration was performed using leucine enkephalin as the reference. Peak detection and identification were carried out using the Waters UNIFI 3D peak algorithm, with matching against an in-house spectral library.

4.3.3 Authentic Sample Collection and Sourcing

The collection and sourcing of authentic blood samples from suspected impaired drivers and postmortem specimens, as submitted to the Wisconsin State Laboratory of Hygiene (WSLH), have been described by Bates et al. (2024) [16]. Due to the anonymization process, demographic data and submission details were not retained. Following submission, the samples underwent the extraction and LC-QToF-MS data collection procedures outlined above.

4.3.4 Utilization of PubChem for Expansion of Structure Set Using Similarity Scoring

Opioid compounds representing three classes (morphinan, fentanyl, and nitazene) were initially identified through literature reviews, and their structures were represented as SMILES strings (Supp. Table A.1, Supp. Table A.2, Supp. Table A.3). These SMILES strings were input into PubChem's similarity search to identify analogous structures with a Tanimoto score ≥ 98. Python, utilizing RDKit, was employed to compile, deduplicate, and remove stereochemical variations, ensuring structural uniqueness. Compounds listed as charged salts, acids, or ions were simplified by removing these components prior to final deduplication.

4.3.5 Selection of GNPS Data Sets for Model Evaluation

MS/MS spectra were retrieved from the GNPS database to evaluate the machine learning models. Both targeted and random selection methods were employed to ensure broad chemical space representation while maintaining computational feasibility. The targeted approach involved searching for opioid compounds within the GNPS library, yielding 204 morphinan, 207 fentanyl, and 6 nitazene spectra. The random selection aimed to capture diverse spectral data, particularly for false positive testing, resulting in a set of 9,443 spectra from approximately 50,000 unique compounds in the GNPS database. All spectra were processed through SIRIUS 5 to generate molecular fingerprints, with spectra failing to process due to insufficient fragmentation being excluded.

4.3.6 Generating Negative Training Examples from RIKEN NP Depo

To balance the positive data set, negative training examples were sourced from the RIKEN Natural Products Depository (NP Depo), which offers a diverse collection of natural products and synthetic derivatives. Compounds from the opioid classes were excluded, resulting in a final selection of 10,959 unique compounds, ensuring broad structural diversity. Future studies may explore alternative databases such as ZINC, DrugBank, COCONUT, and The NP Atlas for

additional diversity.

4.3.7 Mass Frontier for Batch Fragmentation of Compound Structures

Mass Frontier 8.1 (Thermo Fisher Scientific) was used for batch fragmentation of compound structures to predict fragment ions for *in-silico* mass spectra. SMILES strings were converted into Structure-Data Format (SDF) files using RDKit, and the fragmentation settings were configured to reflect protonation under electrospray ionization conditions. The Batch Fragment Generation method was used with extensive rule databases for accurate fragmentation predictions, specifying ion types, charge states, and radical ions. The fragmentation was limited to 60 unique fragments to conform to SIRIUS 5's input limitations.

4.3.8 Generating the MS Files for SIRIUS Using Python

SDF files containing the computed fragments from Mass Frontier were processed using Python and RDKit to extract the exact mass for each fragment. These masses were deduplicated and formatted into MS-Format files compatible with SIRIUS 5. These files served as input for subsequent analysis.

4.3.9 Analyzing In-silico Mass Spectra with the Known Molecular Formula of Each Compound

SIRIUS 5 was used to analyze each *in-silico* mass spectrum and fragmentation pattern, generating molecular fingerprints for each compound. The analysis was configured to predict only the known molecular formula, with default settings applied for all modules.

4.3.10 Compiling In-silico Fingerprints into a Training Matrix

Molecular fingerprints generated by SIRIUS 5 were compiled into a matrix. Each fingerprint, representing 3,878 unique substructures, was associated with metadata, including the sample name. The final matrix was used as input for machine learning models, with each row representing a unique fingerprint.

4.3.11 Utilizing the Scikit-Learn Python Package for Training and Testing

Scikit-learn was used to create training and testing sets for the molecular fingerprint data. The data set was split 80% for training and 20% for testing. Performance metrics such as precision, recall, F1 score, and accuracy were used to evaluate the models. Training involved exhaustive parameter optimization using Scikit-learn and Tensorflow.

4.3.12 t-SNE Visualization of Molecular Fingerprints

To visualize the molecular fingerprints, t-SNE, a dimensionality reduction technique in Scikit-learn, was used to map the fingerprints to a 2D space. The perplexity was set to 100, and the number of iterations was set to 1000. The plot was generated using Matplotlib, with each opioid class labeled with a different color for clarity.

4.3.13 Authentic Sample Mass Spectrometry Data Processing

Raw LC-MS/MS files were processed in MZmine 4.0 to get MS/MS features for subsequent ML analysis. Centroid-based mass detection (noise level 5E3 and 1E3) was performed separately on MS^1 and MS^2 scans, the masses were calibrated using a warfarin-d5 internal standard, and extracted-ion chromatograms were built using the ADAP algorithm (minimum group size = 3 scans, intensity threshold = 1E4, tolerance = 20ppm), then smoothed with a Savitzky–Golay filter (5-point window). A C-13 isotope filter (±3 ppm tolerance, maximum charge = 2). Pseudo-MS^2 spectra were reconstructed via the DIA Pseudo MS2 Builder using a Pearson correlation cutoff of 0.95 to associate fragment ions with their precursors, a feature intensity of 1E4 and fragment intensity of 1E3. Features were then aligned across samples (RT tolerance = 0.1 min, m/z tolerance = 10 ppm), blank-subtracted (Minimum 1 detection, 300% folder increase), and filtered to eliminate duplicates. The curated MS^2 feature list was exported in MGF format—including retention time, precursor m/z, and fragment spectra—and imported into SIRIUS 5 for molecular-formula assignment and fingerprint generation. This workflow ensured that downstream

classification models received deconvoluted, high-fidelity spectral inputs..

4.3.14 Modified Cosine Scoring of Mass Spectra with Different Precursor Masses

The modified cosine score is calculated using the ModifiedCosine function from the MatchMS Python package, which compares two mass spectrometry spectra based on their m/z (mass-to-charge ratio) and intensity values. Initially, both spectra are normalized to the same intensity scale, and any peaks with m/z values exceeding 4 Da above the precursor mass are removed to prevent interference with the scoring process. Subsequently, a noise filter is applied to eliminate peaks with intensities below 0.5% of the maximum intensity. After preprocessing, the ModifiedCosine function compares the two spectra by calculating their similarity. The parameters used for the comparison are a mass tolerance of 0.05 Da, an *m/z* power of 0, and an intensity power of 1. The function then returns a similarity score that quantifies the degree of match between the two spectra.

4.3.15 Multi-tiered Spectral Analysis using Modified Cosine Score

A multi-tiered spectral analysis approach was employed to facilitate the identification and comparison of opioid compounds. Initially, the predicted positive features and their corresponding mass spectra were retrieved and processed into Spectrum objects using the MatchMS Python package. Next, the top structural predictions for each feature from SIRIUS 5 were retrieved, and their corresponding SMILES strings were used to perform structure searches in the GNPS library. This search aimed to identify exact structural matches by querying the extensive compound database within GNPS.

The spectra corresponding to exact structural matches were collected, converted into Spectrum objects using MatchMS, and compared to the feature's mass spectrum using the Modified Cosine Score from MatchMS. The parameters for the modified cosine score were a mass

tolerance of 0.05 Da, an '*m/z* power' of 0, and an intensity power of 1.

In addition, for further analysis, an ECFP4 fingerprint was generated from the structure predicted by SIRIUS 5. This fingerprint was then used in a Tanimoto similarity search against all compounds in the GNPS database. Compounds with a Tanimoto score of 0.7 or higher were selected, and their corresponding spectra were retrieved from the GNPS spectral library. These spectra were then compared to the feature's mass spectrum to assess its relationship to structurally similar known compounds.

The results from each search were filtered using two sets of criteria. The first set of criteria applied stringent conditions: only matches with a modified cosine score greater than 0.7, at least three matching peaks, and a precursor mass difference of less than 0.1 m/z were retained. These criteria were used to identify direct matches to known compounds. In the second tier, a more relaxed filtering approach was used, considering only the modified cosine score greater than 0.7 and a minimum of three matching peaks. This allowed the identification of compounds that were structurally related to opioids, differing by a single modification but maintaining a similar fragmentation pattern.

4.5 REFERENCES

(1)   Centers for Disease Control and Prevention. CDC Reports Nearly 24% Decline in U.S. Drug Overdose Deaths. February 25, 2025. https://www.cdc.gov/media/releases/2025/2025-cdc-reports-decline-in-us-drug-overdose-deaths.html (accessed April 28, 2025).

(2)   U.S. Drug Enforcement Administration. DEA Releases 2024 National Drug Threat Assessment; May 09, 2024. https://www.dea.gov/press-releases/2024/05/09/dea-releases-2024-national-drug-threat-assessment (accessed April 28, 2025).

(3)   Ringuette AE, Spock M, Lindsley CW, Bender AM. DARK Classics in Chemical Neuroscience: Carfentanil. ACS Chem Neurosci. 2020 Dec 2;11(23):3955-3967.

(4)   Devereaux AL, Mercer SL, Cunningham CW. DARK Classics in Chemical Neuroscience: Morphine. ACS Chem Neurosci. 2018 Oct 17;9(10):2395-2407.

(5)   Burns SM, Cunningham CW, Mercer SL. DARK Classics in Chemical Neuroscience: Fentanyl. ACS Chem Neurosci. 2018 Oct 17;9(10):2428-2437.

(6)   Mella-Raipán J, Romero-Parra J, Recabarren-Gajardo G. DARK Classics in Chemical Neuroscience: Heroin and Desomorphine. ACS Chem Neurosci. 2020 Dec 2;11(23):3905-3927.

(7)   Gardner EA, McGrath SA, Dowling D, Bai D. The Opioid Crisis: Prevalence and Markets of Opioids. Forensic Sci Rev. 2022 Jan;34(1):43-70.

(8)   Jones CM, Bekheet F, Park JN, Alexander GC. The Evolving Overdose Epidemic: Synthetic Opioids and Rising Stimulant-Related Harms. Epidemiol Rev. 2020 Jan 31;42(1):154-166.

(9)   Prekupec MP, Mansky PA, Baumann MH. Misuse of Novel Synthetic Opioids: A Deadly New Trend. J Addict Med. 2017 Jul/Aug;11(4):256-265.

(10)   Albores-García D, Cruz SL. Fentanyl and other New Psychoactive Synthetic Opioids. Challenges to Prevention and Treatment. Rev Invest Clin. 2023;75(3):93-104.

(11)   Papsun DM, Krotulski AJ, Logan BK. Proliferation of Novel Synthetic Opioids in Postmortem Investigations After Core-Structure Scheduling for Fentanyl-Related Substances. Am J Forensic Med Pathol. 2022 Dec 1;43(4):315-327.

(12)   Clinton HA, Thangada S, Gill JR, Mirizzi A, Logan SB. Improvements in Toxicology Testing to Identify Fentanyl Analogs and Other Novel Synthetic Opioids in Fatal Drug Overdoses, Connecticut, January 2016-June 2019. Public Health Rep. 2021 Nov-Dec;136(1_suppl):80S-86S.

(13)   Uljon S. Advances in fentanyl testing. Adv Clin Chem. 2023;116:1-30.

(14)   Palmquist KB, Truver MT, Shoff EN, Krotulski AJ, Swortwood MJ. Review of analytical methods for screening and quantification of fentanyl analogs and novel synthetic opioids in biological specimens. J Forensic Sci. 2023 Sep;68(5):1643-1661.

(15)   Broussard LA. Choosing a Fentanyl Immunoassay. J Appl Lab Med. 2024 Sep 3;9(5):871-873.

(16)   Bates MN, Helm AE, Barkholtz HM. Screening for Forensically Relevant Drugs Using Data-Independent High-Resolution Mass Spectrometry. Chem Res Toxicol. 2024 Apr 15;37(4):571-579.

(17)   Tabarra I, Soares S, Rosado T, Gonçalves J, Luís Â, Malaca S, Barroso M, Keller T, Restolho J, Gallardo E. Novel synthetic opioids - toxicological aspects and analysis. Forensic Sci Res. 2019 Jul 3;4(2):111-140.

(18)   Bittremieux W, Wang M, Dorrestein PC. The critical role that spectral libraries play in capturing the metabolomics community knowledge. Metabolomics. 2022 Nov 19;18(12):94.

(19)   Fabresse, N. LC-HRMS: Benefits and limitations in clinical and forensic toxicology. Toxicologie Analytique et Clinique 37, S63–S64 (2025).

(20)   Aldy K, Krotulski A, Brent J, Campleman S, Culbreth R, Logan B, Wax P, Amaducci A, Judge B, Levine M, Schwarz E, Calello DP, Meaden CW, Shulman J, Hughes A, Hendrickson R, Carpenter J, Buchanan J, Manini AF; Toxicology Investigators Consortium (ToxIC) Fentalog Study Group. Emergency Department Patients with Para-Fluorofentanyl Overdose. J Emerg Med. 2024 Nov 20:S0736-4679(24)00366-4. doi: 10.1016/j.jemermed.2024.11.020.

(21)   Vandeputte MM, Stove CP. Navigating nitazenes: A pharmacological and toxicological overview of new synthetic opioids with a 2-benzylbenzimidazole core. Neuropharmacology. 2025 Apr 17:110470. doi: 10.1016/j.neuropharm.2025.110470.

(22)   Lin Z, Chou WC. Machine Learning and Artificial Intelligence in Toxicological Sciences. Toxicol Sci. 2022 Aug 25;189(1):7-19. doi: 10.1093/toxsci/kfac075.

(23)   Jia X, Wang T, Zhu H. Advancing Computational Toxicology by Interpretable Machine Learning. Environ Sci Technol. 2023 Nov 21;57(46):17690-17706. doi: 10.1021/acs.est.3c00653.

(24)   Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.;

Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Product Drug Discovery. Nat Rev Drug Discov 2023, 22 (11), 895–916. https://doi.org/10.1038/s41573-023-00774-7.

(25) Xu Y, Cao L, Chen Y, Zhang Z, Liu W, Li H, Ding C, Pu J, Qian K, Xu W. Integrating Machine Learning in Metabolomics: A Path to Enhanced Diagnostics and Data Interpretation. Small Methods. 2024 Dec;8(12):e2400305.

(26) Zhang JD, Xue C, Kolachalama VB, Donald WA. Interpretable Machine Learning on Metabolomics Data Reveals Biomarkers for Parkinson's Disease. ACS Cent Sci. 2023 May 9;9(5):1035-1045.

(27) Chen Y, Wang B, Zhao Y, Shao X, Wang M, Ma F, Yang L, Nie M, Jin P, Yao K, Song H, Lou S, Wang H, Yang T, Tian Y, Han P, Hu Z. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. Nat Commun. 2024 Feb 23;15(1):1657.

(28) Brittin NJ, Anderson JM, Braun DR, Rajski SR, Currie CR, Bugni TS. Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics. J Nat Prod. 2025 Feb 28;88(2):361-372.

(29) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. Nat Methods 2019, 16 (4), 299–302.

(30) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. Bin; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J Nat Prod 2021, 84 (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.

(31) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. Proceedings of the National Academy of Sciences 2015, 112 (41), 12580–12585.

(32) Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S. Database-

Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC. Nat Mach Intell 2020, 2 (10), 629–641.

(33)   Centers for Disease Control and Prevention. Fentanyl Analog Screening Kits (FAS) and Emergent Panels; U.S. Department of Health and Human Services. https://www.cdc.gov/chemical-threats-and-toxins-laboratory/php/opioids-laboratory/fas-kits.html (accessed April 22, 2025).

(34)   Krajewski LC, Swanson KD, Bragg WA, Shaner RL, Seymour C, Carter MD, Hamelin EI, Johnson RC. Application of the fentanyl analog screening kit toward the identification of emerging synthetic opioids in human plasma and urine by LC-QTOF. Toxicol Lett. 2020 Mar 1;320:87-94.

(35)   Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016 Aug 9;34(8):828-837.

(36)   Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard PM, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M,

Martin H C, McCall LI, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira N, Wang M, Dorrestein PC. Feature-based molecular networking in the GNPS analysis environment. Nat Methods. 2020 Sep;17(9):905-908.

(37)   Kato, N.; Takahashi, S.; Nogawa, T.; Saito, T.; Osada, H. Construction of a Microbial Natural Product Library for Chemical Biology Studies. Curr Opin Chem Biol 2012, 16 (1–2), 101–108.

Chapter 5: Conclusions and Future Work

5.1 CONCLUDING REMARKS

This dissertation presents a cohesive and multidisciplinary framework that addresses two of the most urgent challenges in public health and pharmaceutical sciences: the discovery of novel therapeutics against multidrug-resistant (MDR) pathogens and the accurate detection of synthetic opioids in complex clinical samples. Through the integration of high-resolution LC-MS/MS, yeast chemical genomics (YCG), and machine learning (ML), this body of work illustrates how modern analytical and computational approaches can be combined to accelerate discovery and improve detection workflows in both drug development and forensic toxicology.

In the first project, we established a powerful high-throughput pipeline that integrates YCG with LC-MS/MS-based metabolomics for antifungal discovery. This approach allows for the simultaneous structural and functional characterization of thousands of bacterial natural product fractions. The innovative use of YCG enables mechanism of action (MoA) profiling at scale, while LC-MS/MS facilitates dereplication and structural elucidation. By applying this integrated workflow to over 40,000 bacterial fractions, we can not only prioritize extracts with likely novel antifungal agents but also distinguished those with known MoAs—ensuring that only the most promising candidates advance. This work addresses longstanding bottlenecks in natural product discovery, where prioritization often relies solely on structural novelty without functional validation, frequently resulting in the rediscovery of known compounds. The high throughput nature of this platform enables large scale screening efforts that can efficiently identify and prioritize novel anti-infective agents from diverse and previously untapped microbial sources.

The second project leveraged the complex data generated from metabolomics workflows to train a highly capable pharmacophore-based machine learning multi-classifier. Using *in-silico*

fragmentation spectra and molecular fingerprints, we trained models to categorize natural products into 21 pharmacologically relevant drug classes, achieving >93% accuracy even when applied to noisy or incomplete experimental data. By focusing on pharmacophores rather than full chemical structure or biosynthetic classification, these models were able to extract functionally meaningful patterns that correlate with bioactivity across structurally diverse compounds. This strategy improves dereplication by rapidly identifying known pharmacophores within natural extracts, allowing researchers to triage rediscovery events early while prioritizing compounds that may possess novel mechanisms. Moreover, the use of synthetic training data enables continual model expansion to new classes without requiring exhaustive experimental spectra, making this a scalable solution for accelerating antimicrobial discovery campaigns.

In the third project, we applied our pharmacophore-based ML framework to clinical metabolomics data for opioid detection. Here, we developed robust classifiers for the morphinan, fentanyl, and nitazene opioid classes using molecular fingerprints derived from *in-silico* spectra, which were then validated against experimental GNPS data, the CDC's FAS reference library, and anonymized blood and urine samples collected in collaboration with the Wisconsin State Laboratory of Hygiene. The morphinan and fentanyl classifiers achieved precision and recall values above 95%, significantly surpassing the detection capabilities of traditional spectral database search methods. These models improved compound identification by up to 600% and extended sample coverage by up to 400%, demonstrating their value in computational toxicology and clinical screening. Although the nitazene classifier showed diminished performance in clinical samples—likely due to limited training data and spectral quality—its accuracy on curated experimental data highlights its potential with further optimization. This project underscores the real-world impact of data-driven detection tools in responding to the dynamic and chemically

diverse landscape of synthetic opioids.

Taken together, these three projects demonstrate the transformative potential of YCG, LC-MS/MS, and ML and how they can be applied to multiple fields doing small molecule characterization. From streamlining natural product dereplication to enhancing the detection of synthetic drugs of abuse, this work lays a strong foundation for future applications in both drug discovery and public health surveillance. As chemical and biological datasets continue to expand, the frameworks developed here offer scalable, adaptable, and biologically meaningful tools for unlocking new therapeutic insights and addressing emergent chemical threats.

5.2   FUTURE DIRECTIONS

The research presented in this dissertation lays a strong foundation for scalable, high-throughput workflows in natural product discovery and computational toxicology. Moving forward, several key directions can further enhance the impact, accessibility, and applicability of these methods across a broader range of chemical and biological challenges.

First, the integrated LC-MS/MS and yeast chemical genomics (YCG) workflow offers numerous opportunities for expansion and refinement. Future efforts should focus on applying this platform to characterize more previously characterized small molecule natural product antifungals. By targeting additional antifungals beyond those currently analyzed—especially commonly encountered compounds with known mechanisms—the pipeline can be strengthened and improve its translational relevance and dereplication abilities. As the scale of these screening campaigns grows, automating the statistical and data analysis components of the workflow will be essential. Building reproducible pipelines for spectral processing, feature prioritization, and YCG profile interpretation will streamline the workflow, minimize analytical bottlenecks, and increase the throughput of functional and structural discovery [1].

In the spirit of open science, an important next step is contributing curated screening results to public repositories. Mass spectrometry identifications from LC-MS/MS experiments can be submitted to community-accessible spectral databases such as GNPS or MassBank, providing valuable reference spectra for future dereplication efforts [2,3]. Expanding these resources benefits the entire natural products community by enabling cross-validation, comparative analysis, and improved metabolomic annotation in future studies. By sharing both raw and processed data, this workflow can serve as a model for collaborative and transparent discovery efforts.

On the machine learning front, future work will focus on expanding the coverage and capabilities of natural product pharmacophore classifiers. While the current models achieve high

accuracy across a broad range of antibacterial and antifungal classes, many important sources of bioactive natural products—such as fungal and plant metabolites—remain underrepresented. Extending the model's scope to include these compound families will capture a more complete landscape of known pharmacophores and enhance predictive performance on truly novel structures. Additionally, emerging spectral representation techniques such as MIST, Spec2Vec, and DreaMS offer new ways to encode structural information directly from fragmentation data [4-6]. Incorporating these mass spectral embeddings into model training could improve the specificity and generalizability of pharmacophore classification, enabling access to chemical space not well captured by traditional fingerprints or substructure-based approaches.

In the context of computational toxicology, continued development of opioid classifiers will benefit from training datasets that better reflect the diversity of real-world biological matrices. Expanding to include spectra from blood plasma, vitreous humor, and other clinically relevant fluids will help reduce false positives and increase the selectivity of the models for forensic applications. Beyond opioids, there is a pressing need to develop models for other drug classes such as stimulants, benzodiazepines, synthetic cannabinoids, and hallucinogens—many of which are poorly represented in current spectral libraries and increasingly present in illicit drug markets. These models can be integrated into broader toxicological analysis pipelines, offering a unified ML-driven platform for identifying diverse classes of drugs of abuse.

Finally, democratizing access to these machine learning tools is a critical priority. Developing user-friendly graphical interfaces that wrap these models into intuitive applications will facilitate adoption by forensic toxicologists and analytical chemists without computational backgrounds. Such tools could enable real-time prediction and compound classification directly from experimental data, bridging the gap between advanced algorithmic development and real-

world forensic and public health applications.

Together, these future directions aim to enhance the robustness, reach, and utility of the analytical and computational frameworks developed in this work. By advancing both the scale and accessibility of small molecule discovery and detection, this research continues to push the boundaries of what is possible in modern metabolomics, drug discovery, and forensic science.

5.3   REFERENCES

(1)   Brittin NJ, Aceti DJ, Braun DR, Anderson JM, Ericksen SS, Rajski SR, Currie CR, Andes DR, Bugni TS. Dereplication of Natural Product Antifungals via Liquid Chromatography-Tandem Mass Spectrometry and Chemical Genomics. Molecules. 2024 Dec 28;30(1):77.

(2)   Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016 Aug 9;34(8):828-837. doi: 10.1038/nbt.3597. PMID: 27504778; PMCID: PMC5321674.

(3)   Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010 Jul;45(7):703-14. doi: 10.1002/jms.1777. PMID: 20623627.

(4)   Goldman, S., Wohlwend, J., Stražar, M. *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Mach Intell* **5**, 965–979 (2023).

https://doi.org/10.1038/s42256-023-00708-3

(5)　　Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, van der Hooft JJJ. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLoS Comput Biol. 2021 Feb 16;17(2):e1008724. doi: 10.1371/journal.pcbi.1008724. PMID: 33591968; PMCID: PMC7909622.

(6)　　Bushuiev, R.; Bushuiev, A.; Samusevich, R.; Brungs, C.; Sivic, J.; Pluskal, T. ChemRxiv 2025. This content is a preprint and has not been peer-reviewed.

Appendix A

Supplementary Data Tables for Chapter 4:

Table A.1. Core fentanyl structures used in similarity searching to create the fentanyl opioid class

Common name	CAS number
2-Fluoroacrylfentanyl	2309383-09-1
2-Fluorobutyrfentanyl (o-FBF)	2163847-76-3
2-Fluoroisobutyrfentanyl (o-FIBF)	2351142-33-9
2-Methylacetylfentanyl (o-MAF)	90736-11-1
2-Methylmethoxyacetylfentanyl	
2,5-Dimethylfentanyl	42045-97-6
2,2'-Difluorofentanyl	2748343-87-3
3-Allylfentanyl	82208-84-2
3-Fluorofentanyl (NFEPP)	1422952-84-8
3-Furanylfentanyl (3FUF)	101343-82-2
3-Methylbutyrfentanyl	97605-09-9
3-Methylcrotonylfentanyl	
p-Methylcrotonylfentanyl	
3-Methylfentanyl (3-MF)	42045-86-3
3-Methylfuranylfentanyl (3MFUF, TMFUF)	
3-Methylthiofentanyl	86052-04-2
3-Phenylpropanoylfentanyl	79279-02-0
4-Fluorobutyrfentanyl (4-FBF)	244195-31-1
4-Chloroisobutyrylfentanyl (4-CIBF)	244195-34-4
4-Fluoroisobutyrfentanyl (4-FIBF)	244195-32-2
4-Fluorofentanyl	90736-23-5
Paramethylfentanyl (p-MF)	1838-67-1
para-fluorofuranylfentanyl (p-F-Fu-F)	1802489-71-9
para-chlorofuranylfentanyl (p-Cl-Fu-F)	
ortho-methylfuranylfentanyl (o-Me-Fu-F)	2309383-07-9
ortho-methoxyfuranylfentanyl (o-MeO-Fu-F)	101343-50-4
ortho-isopropylfuranylfentanyl (o-iPr-Fu-F)	
4-Phenylfentanyl	120448-97-7
4-Methoxybutyrfentanyl	2088842-68-4
para-Hydroxy-butyrylfentanyl	
4"-Fluorofentanyl	2748343-99-7
4"-Fluoro-*o*-fluoro-3-methylfentanyl	
4-Methyl-methoxyacetylfentanyl (4-Me-MAF)	
4-Methylphenethylacetylfentanyl	1071703-95-1
Acrylfentanyl	82003-75-6
α-Methylacetylfentanyl	101860-00-8
α-Methylbutyrfentanyl	244195-36-6
α-Methylfentanyl (AMF)	79704-88-4
α-Methylthiofentanyl	103963-66-2
α-Methyl-β-hydroxyfentanyl	
Acetylfentanyl	3258-84-2

Alfentanyl	71195-58-9
Benzodioxolefentanyl	2306823-01-6
Benzoylfentanyl	2309383-15-9
Benzylfentanyl	1474-02-8
BDBM50223545 (Berger Fentanyl)	
	78995-10-5
β-Hydroxythiofentanyl	1474-34-6
β-Hydroxy-4-methylfentanyl	
β-Methylfentanyl	79146-56-8
Brorphine	2244737-98-0
Butyrfentanyl (Bu-F, BUF)	1169-70-6
Brifentanyl	101345-71-5
Carfentanyl	59708-52-0
Benzylcarfentanil	61085-72-1
Ethylcarfentanil	
Crotonylfentanyl	760930-59-4
Cyclopentylfentanyl	2088918-01-6
Cyclobutylfentanyl	2306827-55-2
Cyclopropylfentanyl	1169-68-2
EAZ-91-05, Psicofentanil	
Ethoxyacetylfentanyl	
Isobutyrylfentanyl	119618-70-1
Isofentanyl	79278-40-3
Homofentanyl (N-phenylpropylnorfentanyl)	59708-54-2
R-4173	2413-90-3
trans-phenylcyclopropyl-norfentanyl	102504-49-4
Fentanyl azepane	
Fentanyl carbamate	1465-20-9
Pyridylfentany	1443-41-0
Furanylbenzylfentanyl	497240-21-8
Furanylfentanyl (Fu-F, FUF)	101345-66-8
Furanylethylfentanyl (FUEF)	802544-02-1
2-Benzofuranylethyl-alpha-methylfentanyl	
Fentanyl 4-methylene analogue (WO 2007/093603)	947139-57-3
Indolylethylfentanyl	58399-46-5
IQMF-4	497100-48-8
Lofentanyl	61380-40-3
N-Methylnorcarfentanyl	59708-50-8
Methoxyacetylfentanyl (MAF)	101345-67-9
meta-fluorofentanyl	90736-22-4
Meta-fluoro-methoxyacetylfentanyl	2306825-32-9
4-(Methylthiazolyl)-pyrazolylethylfentanyl	120070-51-1
Mirfentanyl	117523-47-4

MP102	
MP135	2677687-49-7
N-(MDA)-fentanyl	
N-(2C-B)-fentanyl	
Ocfentanyl	101343-69-5
Ohmefentanyl	78995-14-9
Ohmecarfentanil	
4"-Fluoroohmefentanyl	
Orthofluorofentanyl	910616-29-4
p-Bromofentanyl	117994-23-7
p-Nitrofentanyl	
Parafluoroisobutyrylbenzylfentanyl	
4-Fluorocyclopropylbenzylfentanyl	2344231-47-4
Pivaloylfentanyl	
Pyrrole-fentanyl	
R-30490	60618-49-7
Remifentanyl	132875-61-7
Remifentanil bis ethyl ester	
RR49	2376328-79-7
SR-16412	16223-24-8
Secofentanyl	253342-66-4
Senecioylfentanyl	2630378-28-6
Sufentanil	56030-54-7
Tetrahydrofuranylfentanyl	2142571-01-3
Tetramethylcyclopropylfentanyl	2309383-11-5
Tetramethylfentanyl	
Thenylfentanyl	117332-93-1
Thiafentanil	101345-60-2
Thiofentanyl	1165-22-6
Thiophenylfentanyl (Thiofuranylfentanyl)	2306823-38-9
Trefentanyl	120656-93-1
Trifluorofentanyl	
Tropafentanyl (Fentanyl tropane)	
3,4-dichloro-4"-methoxyfentanyl	1161705-29-8
Ureafentanyl	1443-50-1
Valerylfentanyl (VF)	122882-90-0

Table A.2. Core morphinan structures used in similarity searching to create the morphinan opioid class.

Common Name	Pubchem ID
Buprenorphine	644073
Codeine	5284371
Hydrocodone	5284569
Hydromorphine	5284570
Morphine	5288826
Oxycodone	5284603
Oxymorphone	5284604
Levorphanol	5359272

Table A.3. Core nitazene structures used to create the nitazene opioid class.

Compound Name	pubchem ID
Desnitazene (1-diethylaminoethyl-2-benzyl-benzimidazole)	28787
Metodesnitazene (Metazene)	26412
Metodesnitazepyne	168310617
Etodesnitazene (Etazene)	149797386
Etodesnitazepyne	162623599
Etodesnitazepipne	162623611
Protodesnitazene	157010653
Isotodesnitazene	162623708
Nitazene	15327524
meta-Metonitazene	
Metonitazene	53316366
Metonitazepyne	168323127
Metonitazepipne	168323148
N-Desethylmetonitazene	168310587
Metomethazene	
Dimetonitazene	162623836
α-methyl-metonitazene	162625089
Metonitazene phenethyl homologue (Ethylene metonitazene)	
Etonitazene	13493
O-Desethyl-etonitazene	156588969
N-Desethyletonitazene (NDE)	162623580
Etonitazene 5-amino metabolite	13408927
Etomethazene	168310446
Etonitazene 5-trifluoromethyl analogue (Etotriflazene)	21815908
Etonitazene 5-cyano analogue (Etocyanazene)	27268
Etonitazene 5-acetyl analogue (Etoacetazene)	25957
Etonitazene 5,6-dichloro analogue (Etodicloazene)	
Etonitazene N,N-dimethyl analogue	67089584
Etonitazepyne	155804760
Etonitazepipne	162623834
Etonitazene morpholine analogue	162623685
Etonitazene 6-nitro isomer (iso-etonitazene)	59799752
Protonitazene	156589001
Protonitazepyne	168322728
Protonitazepipne	168323138
N-Desethylprotonitazene	168310594
Isotonitazene	145721979
Isotonitazepyne	168322631
Isotonitazepipne	168322735

N-Desethylisotonitazene	162623899
Butonitazene	156588955
Isobutonitazene	168322282
Secbutonitazene	168322285
Etoetonitazene	162623504
Flunitazene	156588967
Clonitazene	62528
Diclonitazene	
α-carboxamido-clonitazene	
Bronitazene	162623726
Nitronitazene	
Methylnitazene (Menitazene)	162623683
Ethylnitazene (Enitazene)	162623845
Propylnitazene (Pronitazene)	162623877
t-Butylnitazene	162623621
Acetoxynitazene	162623779
Methylthionitazene	162623790
Ethylthionitazene	162623931
Etodesnitazene phenylthio analogue	21045
Etodesnitazene phenylthio / pyrrolidine analogue	19846499
Methylenedioxynitazene	
Ethyleneoxynitazene	168310596

Table A.4. The 1,651 Pubchem IDs for all the fentanyl analog training structures.

171381222	171441041	173755	173761	1057432	17488205	17893384	19009442
19030772	19030983	10596596	19031103	19372908	19772622	19829407	10596700
19829427	19935657	20055412	20068771	20068779	20106542	20190687	20295516
20295526	20295531	20334067	20334069	20334071	20334078	20334080	20345523
10619456	20382195	20382239	20382252	20382277	20382279	20382330	20382345
20382381	20382383	20393086	20393093	20976596	20976597	20976598	21087476
21136023	21299416	21299612	21299613	21299614	21299615	21299628	21595393
21595394	21595395	21595396	21595397	21595399	21595400	21595401	10687182
21595402	21595404	21595405	21595406	21595407	217866	21855264	10690570
100943644	21926105	21926117	21926200	21926461	21926600	10714591	21926816
21927116	21927168	21927282	21951516	21951518	21951520	21951521	22105201
22294993	22558096	22558116	10739284	22558171	22809178	23083211	23083219
23083257	23109894	23324781	10761716	23339019	23339023	23339036	23339046
23339047	10785926	23339053	23339055	23339058	23339063	23339067	23339069
10809686	23339077	23339081	23398453	23398454	23539668	10810457	23623857
23623859	23623866	23802069	23887195	2396802	10810458	2396805	23988584
10832145	100970758	24731180	24731873	24731884	24843769	24843770	10833545
24898825	24898826	24898827	25110126	25191318	27282486	3017231	10834364
3024792	3025331	3025332	3025333	3031170	3038995	3042857	3042858
3044053	3046356	3058848	3073882	3074076	3074078	108507270	3074079
3074080	3078857	3078858	3078860	3078861	3078862	3081002	3081003
3276653	3345	3616435	3655995	37888366	37888371	398662	108507330
40453084	4105009	4135021	4167174	42167079	42247416	42628841	42853212
42853216	42853217	42853221	42853223	42853224	42853228	42853243	11025431
100974914	42853250	42853251	42853252	42853254	42853263	42853264	42853272
42853274	42853281	42853285	42853286	42853289	42853290	42853291	42853294
42853343	42853344	42853347	42853348	42853349	42853352	42853354	42853355
42853356	42853360	42853362	42853367	42853368	11047199	42853369	42853370
42853371	42853373	42853374	42853375	42853376	42853377	42853378	42853385
42853388	42853396	110558577	42853435	42853440	42853441	42853442	42853444
42853446	42853450	42856534	42856535	42856538	111538024	42856545	42856546
42856560	42856561	42856562	100974916	42856564	42856565	42856567	42856571
42856572	42856573	42856575	42856576	42856577	11165586	42856581	42856582
42856583	42856584	42856585	42856586	42856587	42856588	42856589	42856590
42856591	42856592	42856593	42856594	42856595	42856596	42856604	42856605
42856864	42858014	42858018	42858020	11326977	42858023	42858024	42858035
42858038	42858039	42858040	42858044	42858052	11349939	42858067	42858073
42858078	11382307	42858082	42858107	42858115	42858116	42858117	42858119
42858123	42858131	42858132	42858133	42858150	42858151	42858152	42858217
42858218	42858219	42858220	42858231	42858232	42858235	42858239	42858240
42858244	42858250	11495094	42858263	42858264	42858278	42858282	42858284
42858290	42858293	42858295	42858298	42858299	42858300	42858301	42858317
42858328	42858374	42858383	42858387	11552110	42858397	42858401	42858403
42858412	42858418	11552587	42858419	42858421	42858423	42858425	42858426
42858452	42858458	42858472	42858474	42858475	42858486	11640721	42858488
42858489	42858490	42858510	42858521	42858544	101015858	42858549	42858555
42884	43822366	44125152	44125153	44125154	44125155	44125156	44125157
44125283	44318928	44319006	44319008	44319009	44319010	44319339	44350683
44380587	44382166	44392095	44392148	44416393	44416410	44416411	44416658
44425369	44425374	44425378	44425380	44458919	44891003	451052	451056
451057	45901022	46047771	46047772	46047773	46047774	46047775	46047778
46047784	46047785	117857359	46047786	46047787	46047788	46047789	46047790

46047792	46047793	46047794	46047795	46047796	11792630	46047798	46047800
46047804	46047820	46047899	46047900	46047903	11793161	46047909	46047911
46047912	46047913	46047914	46047916	46047918	46047919	46047928	46047929
46047931	46047932	46047933	46047934	46047937	46047938	46047942	118604724
101015859	46047943	46047948	46047954	46047956	46047957	46047959	118612099
46047962	46047963	46047964	46047966	46047967	46047984	46047985	46047986
46047987	46047988	46047989	46047990	46047991	46047995	46047998	46047999
118718260	46048001	46048003	46048004	46048008	46048016	118796436	46048018
46048019	46048020	46048026	46048032	46048033	118796493	46048035	46048036
46048054	46048055	46048056	46048057	46048059	46048061	46048062	118796521
46048063	46048072	46048073	46048075	46048078	46048079	46048080	119026032
46048081	46048082	46048094	46048135	46048136	46048138	46048139	46048140
119064355	46048148	46048149	46048150	46048161	46048162	46048163	46048164
46048167	46048168	46048169	46048170	46048171	46048172	46048173	46048174
46048176	101015860	46048177	46048178	46048184	46048188	46048195	46048196
46048198	46048217	46048218	11950862	46048219	46048220	46048221	46048222
46048224	46048231	46048232	46048233	46048235	46048238	11951038	46048239
46048241	46048250	46048258	46048314	46048318	11951208	46048323	46048324
46048328	46048329	46048330	46048331	46048332	46048333	11951210	46048335
46048337	46051938	46051939	46051940	46051943	11951923	46051952	46051953
46051954	46051959	46051960	46051961	46051962	46051963	46051966	46051970
46051977	46051981	46051982	46051994	46051999	46052000	46052001	46052003
46052005	46052013	46052014	46052016	46052017	46052018	11955472	46052019
46052023	46052025	46052027	46052028	46052030	46052031	46052032	46052033
101015861	46052034	46052035	46052036	46052037	46052038	46052039	46052041
46052042	46052045	46052046	46052048	46052049	46052050	46052051	46052052
46052055	46052059	46052064	11956067	46052079	46052080	46052081	46052958
46052959	46052960	46054239	11956068	46054248	46054249	46054250	46054251
46054254	46054257	11956069	46054274	46054276	46054277	46054280	46054283
11956072	46054289	46054292	46054294	46054302	46054312	46054313	46054314
46054316	46054335	46054336	46054344	46054346	46054347	46054349	46054350
46054351	46054352	46054366	46054376	46054382	46054385	46054439	46054442
119878228	10042904	46054449	46054452	46054456	46054460	46054462	46054463
119878273	46054466	46054468	46054469	46054473	46054476	119883042	46054490
46054493	46054500	46054502	46054580	46054589	46054599	46054600	46054602
46054604	46054605	46054607	46054610	121350377	46054613	46054623	46054624
46054626	46054636	121350387	46054641	46054662	46054663	121358751	46054673
46054678	46054679	46054680	46054683	46054685	46054686	46054688	121358757
46054689	46054692	46054705	46054706	46054707	46054708	121543410	46054712
46054715	46054717	46054718	46054720	46054721	46054723	46054725	46054728
46054742	46054744	46054745	46054809	46054810	46054822	46054827	46054828
46054862	46054869	46054875	46054888	46054890	46054898	46054919	46054924
46054930	46054932	46054944	46054947	46054949	46054953	46054963	46054964
46054965	46055008	46055009	46055017	46055022	46055035	46055043	46055044
46055046	46055047	46055048	46055054	46055067	12297997	46055123	46115048
46158952	46158966	46158989	46158998	46158999	46159007	46159020	46159021
46159023	46159054	46159056	12298001	46159066	46159075	46159084	46159119
46160556	46160568	46160573	46160580	46160582	46161573	12298011	46161589
46161603	46161644	46161653	46161654	46161656	46161673	46161703	12298017
46161774	46161781	46161783	46161810	12298019	4692322	46950900	46981646
46986131	46992798	47080497	12298022	5257828	527009	527010	12298024
527011	527012	527013	527014	527015	527016	527017	527018
527020	12298025	527021	53299315	53316367	53322195	53352464	53352540
53352581	12298027	53547741	53549745	53609127	53610443	53618245	53687214

12298029	53753750	53878971	53900759	53905679	54206312	54210636	54238167
54264746	12298031	54287070	54439950	56364279	56364683	56365463	56365464
56365763	12298036	56401210	56523169	56605077	56672266	56789510	56845994
12298038	57096656	57217020	12298040	57557589	57567616	57567619	58036805
58036807	58636	12298042	58708472	58853398	59152757	59200765	59200780
59200806	59200813	59200818	12298044	59487289	59714849	12298047	59714889
59714927	59715004	12298049	59869520	60073349	12298056	60073420	60073424
60073465	60157785	60157786	60157787	60157788	60196318	60196323	12298058
60196324	60237731	60245120	60245151	60252389	60254076	12298060	60301653
60391797	60575	60609027	615722	6425761	6569989	66616016	66737933
66804485	66850145	66869050	67121034	67225123	67286282	67646546	67647581
67808932	67809723	67809800	67895970	67895986	67896121	67896165	67896166
67896242	67904225	67904546	67904685	12298079	67928137	68023435	68570081
68574852	68769029	68770097	68771350	12298082	68960896	12298085	69146406
69161601	69183375	69188140	69259738	12298088	69305096	69418487	6957584
12298090	69635904	69646947	69646971	12298092	69646978	69646983	69647005
69647057	69684196	12298094	10114933	69839659	69839938	69840730	12298096
69886729	69951432	70115110	70367747	70431646	70431864	12298098	70432524
70454709	70454825	70454882	70483312	70483400	70484862	70503051	12298101
70503138	70504027	70505115	70611034	70611057	70611214	70611824	70858071
12298103	70861517	70861530	70861550	70861554	70861604	70861630	70928095
70928107	12298106	71043768	71043769	71058	12298107	71198908	71312323
71337341	71345469	71354243	12298109	71624265	71624266	71624267	71624268
71624269	71624390	71624391	12298111	71624392	71624505	71625483	71625484
71748858	71750358	12298113	72017786	72025515	7271781	73345882	7365707
75020716	12298115	76320583	76335552	76900314	77626357	83281398	83287383
83288006	12298116	83288287	84008	85672931	85823519	86130529	86130531
86143852	12298122	86174165	86174168	86280430	86745720	86887423	86887442
86887456	12298127	86887458	86901383	86904208	86933952	86952725	12298129
87734020	88315627	88442519	88798729	88985342	89088770	89138156	12298133
89213376	89250041	89358772	89757168	12298166	90268755	90268870	90316480
90316604	90316606	90316607	90316608	90341316	90349949	90459841	90475918
12298183	90655976	90655977	12298214	90935544	90954902	91071409	91081492
91091176	12298216	91124050	91128650	91152256	91265526	91281675	123228931
91468186	91581350	91635140	91635143	91723091	91723093	91753683	123404734
91753685	91753687	91753688	91753689	91753690	91753692	91753693	91753694
91753696	933144	123437306	933299	95162283	968688	968691	97537765
98216073	98871253	9906422	9933543	9954409	123703550	9976044	9976187
123807110	124136863	124523933	101189688	124716	126635327	126669203	126669270
126669277	126669384	126669457	101193924	126669568	126669670	126669671	126682248
126682249	126691691	126713114	126803101	101193925	10044438	12741853	12741854
12741855	12741856	12741857	12741858	12741865	12741866	12741880	12782335
12782336	12782338	12782339	12782361	12782363	12782364	129051185	10137989
129076806	129377891	129522083	129522102	129597874	129670628	129697494	129844553
101454647	12991697	130286902	13065741	131633018	131699291	131700645	131864579
131888528	101454648	131917297	131992936	132176627	132176681	132176682	132176683
132176696	101454649	132182361	101586196	132608673	132988369	132992169	101587025
133865	13408923	13408924	13408925	134231897	13428511	101587028	134546659
134793530	134849518	134849519	101598717	136525433	13653605	13653608	13653610
13653612	13653614	13653630	13653632	13653647	13653667	13653678	13653679
13653698	101616955	13653723	13653726	13653734	13653738	13653749	13653754
101616956	13676753	101636664	13676761	13676767	13676769	13676771	137332272
137491030	137699862	137699889	137699891	137699934	137699946	137699951	137699957
137699965	137699970	137699972	137699999	101641408	137700005	137700008	137700009

137700015	137700016	137700022	137700024	137700028	137700031	101641409	137700034
137700036	137700038	137700040	137700042	137700044	137700046	137700048	137700050
101802436	137700052	137700055	137700057	137700059	137700061	137700065	137700066
137700068	137700073	101837733	137700076	137700078	137700080	137700082	137700085
137700087	137700090	137700094	137700097	101837734	137700099	137700105	137700108
137700116	137700118	137700119	137700121	137700124	137700126	137700139	137700144
137700202	137700284	137700315	137700357	137700400	137700414	137700422	137700436
137700441	137700454	137700474	137700476	137700482	137700484	137700488	137700501
137700503	137700509	137700511	137700512	137700525	137700529	138395499	138455085
139033121	139386870	139409514	139597318	140009604	10184281	140104955	140394781
140418226	140523878	140759152	102014812	14099644	141033145	141141931	141272149
14129713	14129715	14129718	141338049	141338051	141532072	141532078	141855167
102014814	142147338	142256314	142989407	143062549	143303318	10201888	143303334
143303341	143303352	143656184	143664755	143665555	143722343	143763066	143969943
144547810	144663136	14494438	14494454	14494456	14494466	10225693	14494470
14494472	14494476	14494478	14494480	14494486	10226815	145068329	145271286
14571011	145723951	10229844	146048044	146050387	146167149	146167150	146167151
146167152	102426087	146312425	146657814	14742901	148095834	148170726	14873773
14873777	14879856	14879857	14879883	14879888	149149474	149149929	10317423
149512913	14969510	14969565	10317660	15067960	15067962	15067965	15075638
15075651	15169882	15169883	15169884	15169885	152780568	152822668	153116157
153387665	153387666	153387667	153387669	153387670	153387671	153387674	10360675
153395448	153395456	153705111	153893982	154099681	154200904	154325429	154353132
154415657	154572819	10363138	154953055	155042263	155269371	155439887	155439920
155439921	10364256	155533331	155548401	155587430	155801564	155856567	156314366
156346344	156589019	156596447	156596512	156613767	156614134	156782577	156789215
157010657	157010706	157439981	158020199	158147541	158646172	158759141	158783774
159427142	160716717	10383666	100831594	161096650	161455950	161616706	161648445
161670897	162056	162394591	10384181	162643506	162663559	163203610	10385859
163603277	163811983	163884436	164061980	16413904	164188433	164851888	164851889
164851891	164851893	164851894	164851895	164851897	164851898	164851902	164851903
164851904	164851905	164851907	164851909	164851910	164851912	164851913	164851915
164851916	164851922	164851923	165023440	165351589	165361376	165361426	165361438
165361450	165361468	165361493	165361501	165361531	165361535	165361569	165361570
165361579	165361580	165361586	165361588	165361602	165361615	165361683	165361684
165361685	165361686	165361694	165361709	165361738	165361749	165362008	165362244
165362254	165362255	165362258	165362321	165362376	165365054	100943641	165365069
165365073	165365092	165365111	165365135	165365138	165365263	165365264	166057256
166129741	166129745	166129750	166129751	166129752	166129753	166129803	166129824
166129839	10499027	166653518	166653538	166653559	166654128	166654131	166654141
166654142	166659183	166659243	10522383	166659306	167185819	167222863	167264748
167270948	10527494	168084550	168084552	168084560	168084562	168084566	10545121
168268664	168268665	168312200	168312202	168313361	168313601	168313701	168313708
10545824	168313914	168314346	168322878	168325134	168796180	169026662	169026663
10548474	169090884	169090887	169436287	169437213	10568896	169443253	169443845
169446045	169764504	169765061	169765114	10570297	169792068	169865761	170223808
170361259	171327140	10570799					

Table A.5. The 650 Pubchem IDs for all the morphinan opioid training structures.

129640780	129640789	129643810	129645307	129678937	129678954	129684446	129686172
129715417	129734435	101602191	130207503	130207509	130211031	130211033	130211051
130222935	130223068	130223147	130223352	130223591	101666025	130223853	130226897
130293500	130324219	13065595	13065597	13065599	131967192	132184799	134186042
101737703	134226130	134226132	134227543	134251690	134275559	134308527	134602847
134836390	134925881	135203961	101774461	13559325	136368059	13751901	137541753
138605191	138720655	139351507	139352024	13946392	102061769	139468470	140325979
140337145	140520615	140970564	140976824	14099371	140995038	141022215	141080702
10215960	141080705	141106745	141215702	141334677	141341942	141345094	141347262
141362973	141397654	102384265	141460581	14190489	14190494	142026340	14263005
14263014	144294207	145783920	14582062	14582066	10291563	14589517	14589519
14589521	146036309	14683796	14698178	14698180	147581880	147954167	147977653
10322719	148827351	150055085	150664178	150865895	15102404	151056914	151747040
151994811	153470024	10041407	15386302	154290160	154318936	154395329	154733675
154733676	155607242	156057941	156092245	156234623	10379999	156518121	158460105
158613719	158970695	159452794	161460093	161636666	162098187	10382574	162443
162648676	162652544	163285541	163495659	163677873	163798186	10383718	164041403
164068086	16407254	164080277	164085905	164169612	165351373	166672819	166885961
10393632	166912465	167041651	167069303	167120603	167441348	17748249	18396871
197073	20054882	10480086	204640	20831729	21159580	21279955	21299523
21299643	21299645	21582261	21595536	21787156	10667557	21787157	21795964
22816303	22870688	23242433	23393293	25064535	107765	25123782	25123784
25178033	434374	44179702	44215615	44272199	44272297	44276697	44276752
10928736	44276797	44276809	44276916	44276930	44276960	44281959	44281960
44282000	44282086	44282112	10948296	44282145	44282158	44282413	44282452
44282458	44282460	44293036	44293037	44293051	44293052	10980702	100927606
44293481	44293586	44293610	44293611	44293612	44293616	44293693	44293694
44293864	44293935	10993301	44300656	443408	44372185	44372348	44380353
44380504	44385751	44386942	44401217	44413961	11001908	46844894	46844895
11056967	46844896	46877675	46877678	46877679	46877680	46877681	46877682
46877711	46878174	46878175	11163648	46878177	46878179	46878180	46878183
46878184	46878185	46878186	46878187	46878189	49835639	11164168	50915453
50915781	50915939	50922685	5284371	5284569	5284570	5284595	5284596
5284603	5284604	5288826	52946047	53327323	53327324	5357395	5359271
5359421	5360516	5361887	11301558	5361918	5362507	5362518	53741380
53756271	53774906	54005992	54063408	54183786	11427766	54357925	54372392
54411474	54497266	5462505	5462506	5464194	54756989	54756990	5484369
11743925	5486608	5486611	5486940	5487695	5488009	5491906	5492746
56653640	11754959	101070757	56681954	56843102	56843404	57021400	57046068
57055134	57170108	57284474	57296881	57321384	117617705	5745703	5745713

5745810	5748235	5748326	5748371	57590826	57745504	57745518	57745532
117654814	57764534	58105345	58174345	58174407	58202450	58259502	58296570
58308194	58380456	58380488	117704565	58480335	58618946	58618948	58798386
58840856	58840875	58840876	58840877	58874200	59027198	117733799	59164554
59439392	59456381	59549322	59765024	59765038	59866327	60073310	60073342
60073347	117733819	60073360	60073380	60073411	60073476	60107318	60125256
644073	66587870	66587875	66713919	117733822	66828855	67203232	67679952
67679962	67680000	67680106	67680108	68114362	68163323	68219712	117750458
68247320	68248577	68256023	68256025	68256043	68256047	68256050	68256052
68299962	68325065	117865334	68332528	68336453	68336479	68338189	68339357
68346092	68346094	68538994	68905457	68916847	117883153	69120976	69272927
69664197	70348103	70357827	70550609	70573024	70573067	117943726	101125088
70587391	70660367	70660376	70660380	70804218	70805566	70809141	70809202
70880362	118052861	70901131	70957397	70966917	70998354	71022166	71076277
71079990	71096750	71097116	71097205	118071063	71097207	71108305	71229805
71230334	71230336	71300683	71314295	71315359	71315723	118100424	71517539
71518049	71518050	71585323	71585325	71585528	71585635	71588001	71623259
71624445	118119024	71624446	71624447	71624552	71624553	71624554	71624555
71624556	71624661	71624662	71624663	118200818	71624664	71624780	71624894
71624895	71625025	71625026	71625027	71625028	71625029	71625148	118245120
71625149	71625277	71625278	71717732	71749208	71819401	76966059	81689704
85821085	86245744	118248663	87133322	88043979	88882626	89221731	89517043
89517090	89517112	89629139	89743541	89931686	118248841	89938472	89938493
89947154	89951002	89969769	89978603	89978715	89978785	89981558	89983705
118480001	90051037	90060829	90060837	90060840	90295854	90302273	90302451
90302521	90302523	90306838	118566514	101144993	90306855	90306887	90323554
90411173	90455320	90644874	90656664	90656666	90656667	90656668	118708105
90678812	90679038	90679039	90679041	90679043	90679045	90679046	90679047
90679048	90679049	118988707	90679050	90679051	90679052	91016062	91019274
91092725	91119418	91169607	91237622	9797380	118989481	9798735	9819364
9928583	9979004	122177867	122177869	122177870	122177871	122177877	122179050
122179055	101144994	122179059	122581380	122683773	122705357	123229194	123331467
123363504	123504188	123570156	123753536	10115837	123840616	123860047	126491388
126499132	126527132	126682458	126714502	129012978	129012984	129013018	101445074
129061338	129074295	129145633	129191805	129236961	129634426	129634451	129635088
129635171	129640447	101550853					

Table A.6. The 204 morphinan opioid GNPS spectrum IDs.

CCMSLIB00000568275	CCMSLIB00003134518	CCMSLIB00003136470	CCMSLIB00003136908
CCMSLIB00003138452	CCMSLIB00004679363	CCMSLIB00004679368	CCMSLIB00004721185
CCMSLIB00004721186	CCMSLIB00004721187	CCMSLIB00000085071	CCMSLIB00004721191
CCMSLIB00004721192	CCMSLIB00004721193	CCMSLIB00004721194	CCMSLIB00004721195
CCMSLIB00004721196	CCMSLIB00004721197	CCMSLIB00004721198	CCMSLIB00004721850
CCMSLIB00000085244	CCMSLIB00004722159	CCMSLIB00005730093	CCMSLIB00005730105
CCMSLIB00005731132	CCMSLIB00005731756	CCMSLIB00005732171	CCMSLIB00005733056
CCMSLIB00005734240	CCMSLIB00005734603	CCMSLIB00000085770	CCMSLIB00005735031
CCMSLIB00005735253	CCMSLIB00005736317	CCMSLIB00005771300	CCMSLIB00000086055
CCMSLIB00005771315	CCMSLIB00005771337	CCMSLIB00005771339	CCMSLIB00005771345
CCMSLIB00005771362	CCMSLIB00005771403	CCMSLIB00005771407	CCMSLIB00005771423
CCMSLIB00005771432	CCMSLIB00005771436	CCMSLIB00005771439	CCMSLIB00005771459
CCMSLIB00005771481	CCMSLIB00005771492	CCMSLIB00005771494	CCMSLIB00005771502
CCMSLIB00005771512	CCMSLIB00005771542	CCMSLIB00005771585	CCMSLIB00005772121
CCMSLIB00005772197	CCMSLIB00005772272	CCMSLIB00005772284	CCMSLIB00005772569
CCMSLIB00005772667	CCMSLIB00005772814	CCMSLIB00000206064	CCMSLIB00005773114
CCMSLIB00005773239	CCMSLIB00005773248	CCMSLIB00005773269	CCMSLIB00005773509
CCMSLIB00005773522	CCMSLIB00005773562	CCMSLIB00005773572	CCMSLIB00005773613
CCMSLIB00000206065	CCMSLIB00005773630	CCMSLIB00005773797	CCMSLIB00005773884
CCMSLIB00005773942	CCMSLIB00005773987	CCMSLIB00005774216	CCMSLIB00005774387
CCMSLIB00005774565	CCMSLIB00000077056	CCMSLIB00005774727	CCMSLIB00005774854
CCMSLIB00005774948	CCMSLIB00005774985	CCMSLIB00005774996	CCMSLIB00005775043
CCMSLIB00000206092	CCMSLIB00005775049	CCMSLIB00005775107	CCMSLIB00005775285
CCMSLIB00005775857	CCMSLIB00005775925	CCMSLIB00005776045	CCMSLIB00000206093
CCMSLIB00005776192	CCMSLIB00005776231	CCMSLIB00005776397	CCMSLIB00005776447
CCMSLIB00005776735	CCMSLIB00000206094	CCMSLIB00005776790	CCMSLIB00005776799
CCMSLIB00005776920	CCMSLIB00005777077	CCMSLIB00005777114	CCMSLIB00005777198
CCMSLIB00008851235	CCMSLIB00008851323	CCMSLIB00008851422	CCMSLIB00000206095
CCMSLIB00009919317	CCMSLIB00009919478	CCMSLIB00009919490	CCMSLIB00009919494
CCMSLIB00009919506	CCMSLIB00009919507	CCMSLIB00009919518	CCMSLIB00009919519
CCMSLIB00009919520	CCMSLIB00009919525	CCMSLIB00009919526	CCMSLIB00010007706
CCMSLIB00010011926	CCMSLIB00010129290	CCMSLIB00011429533	CCMSLIB00011429539
CCMSLIB00011429541	CCMSLIB00011429550	CCMSLIB00011429551	CCMSLIB00011429562
CCMSLIB00011429563	CCMSLIB00011429564	CCMSLIB00011429569	CCMSLIB00011429570
CCMSLIB00000206199	CCMSLIB00000206200	CCMSLIB00000206201	CCMSLIB00000078434
CCMSLIB00000206254	CCMSLIB00000206255	CCMSLIB00000206256	CCMSLIB00000206258
CCMSLIB00000206259	CCMSLIB00000206260	CCMSLIB00000206261	CCMSLIB00000078568
CCMSLIB00000206298	CCMSLIB00000206299	CCMSLIB00000206300	CCMSLIB00000206301
CCMSLIB00000210680	CCMSLIB00000210684	CCMSLIB00000210686	CCMSLIB00000210688
CCMSLIB00000210690	CCMSLIB00000078773	CCMSLIB00000210692	CCMSLIB00000210696

CCMSLIB00000210698	CCMSLIB00000210700	CCMSLIB00000210702	CCMSLIB00000210705
CCMSLIB00000210707	CCMSLIB00000210969	CCMSLIB00000078795	CCMSLIB00000210973
CCMSLIB00000210975	CCMSLIB00000210977	CCMSLIB00000210979	CCMSLIB00000210982
CCMSLIB00000210986	CCMSLIB00000210987	CCMSLIB00000210989	CCMSLIB00000210991
CCMSLIB00000079179	CCMSLIB00000210993	CCMSLIB00000210996	CCMSLIB00000211586
CCMSLIB00000211592	CCMSLIB00000211593	CCMSLIB00000211596	CCMSLIB00000211598
CCMSLIB00000084759	CCMSLIB00000211602	CCMSLIB00000211604	CCMSLIB00000211606
CCMSLIB00000211608	CCMSLIB00000211610	CCMSLIB00000211612	CCMSLIB00000211754
CCMSLIB00000211760	CCMSLIB00000084852	CCMSLIB00000211762	CCMSLIB00000211764
CCMSLIB00000211766	CCMSLIB00000211770	CCMSLIB00000211772	CCMSLIB00000211774
CCMSLIB00000211776	CCMSLIB00000211778	CCMSLIB00000211780	CCMSLIB00000084974

Table A.7. The 207 fentanyl opioid GNPS spectrum IDs.

CCMSLIB00009919393	CCMSLIB00009919394	CCMSLIB00009919395	CCMSLIB00009919397
CCMSLIB00009919399	CCMSLIB00009919400	CCMSLIB00009919402	CCMSLIB00009919404
CCMSLIB00009919406	CCMSLIB00009919408	CCMSLIB00000206356	CCMSLIB00009919409
CCMSLIB00009919410	CCMSLIB00009919411	CCMSLIB00009919412	CCMSLIB00009919414
CCMSLIB00009919415	CCMSLIB00009919416	CCMSLIB00009919417	CCMSLIB00009919418
CCMSLIB00009919420	CCMSLIB00000206392	CCMSLIB00009919422	CCMSLIB00009919424
CCMSLIB00009919426	CCMSLIB00009919427	CCMSLIB00009919428	CCMSLIB00009919429
CCMSLIB00009919430	CCMSLIB00009919431	CCMSLIB00009919432	CCMSLIB00009919433
CCMSLIB00000206393	CCMSLIB00009919434	CCMSLIB00009919435	CCMSLIB00009919437
CCMSLIB00009919439	CCMSLIB00009919440	CCMSLIB00009919443	CCMSLIB00009919444
CCMSLIB00009919446	CCMSLIB00009919447	CCMSLIB00009919451	CCMSLIB00000206394
CCMSLIB00010129297	CCMSLIB00010129299	CCMSLIB00010129300	CCMSLIB00010129301
CCMSLIB00010129305	CCMSLIB00010129306	CCMSLIB00010129308	CCMSLIB00010129309
CCMSLIB00010129310	CCMSLIB00010129311	CCMSLIB00000206395	CCMSLIB00010129312
CCMSLIB00010129313	CCMSLIB00010129314	CCMSLIB00010129316	CCMSLIB00010129317
CCMSLIB00010129318	CCMSLIB00010129319	CCMSLIB00010129320	CCMSLIB00010129321
CCMSLIB00010129322	CCMSLIB00000206396	CCMSLIB00010129323	CCMSLIB00010129324
CCMSLIB00010129325	CCMSLIB00010129326	CCMSLIB00010129327	CCMSLIB00010129328
CCMSLIB00010129329	CCMSLIB00010129330	CCMSLIB00010129331	CCMSLIB00010129332
CCMSLIB00000565838	CCMSLIB00010129333	CCMSLIB00010129334	CCMSLIB00010129335
CCMSLIB00010129336	CCMSLIB00010129337	CCMSLIB00010129338	CCMSLIB00010129339
CCMSLIB00010129341	CCMSLIB00010129342	CCMSLIB00010129343	CCMSLIB00003135084
CCMSLIB00010129344	CCMSLIB00010129345	CCMSLIB00010129346	CCMSLIB00010129347
CCMSLIB00010129348	CCMSLIB00010129350	CCMSLIB00010129351	CCMSLIB00010129353
CCMSLIB00010129355	CCMSLIB00011429480	CCMSLIB00003135626	CCMSLIB00011429481
CCMSLIB00011429482	CCMSLIB00011429483	CCMSLIB00011429485	CCMSLIB00011429487
CCMSLIB00011429488	CCMSLIB00011429489	CCMSLIB00011429490	CCMSLIB00011429494
CCMSLIB00011429499	CCMSLIB00003136417	CCMSLIB00011429500	CCMSLIB00011429502
CCMSLIB00011429503	CCMSLIB00011429504	CCMSLIB00011429508	CCMSLIB00011429509
CCMSLIB00011429510	CCMSLIB00011429511	CCMSLIB00011429513	CCMSLIB00011429514
CCMSLIB00003137026	CCMSLIB00011429517	CCMSLIB00011429518	CCMSLIB00003139043
CCMSLIB00003140014	CCMSLIB00005730630	CCMSLIB00005730640	CCMSLIB00005730959
CCMSLIB00005731011	CCMSLIB00005731434	CCMSLIB00005731800	CCMSLIB00005732135
CCMSLIB00000206007	CCMSLIB00005732920	CCMSLIB00005732921	CCMSLIB00005733673
CCMSLIB00005734514	CCMSLIB00005734519	CCMSLIB00005735192	CCMSLIB00005735321
CCMSLIB00005735795	CCMSLIB00005736118	CCMSLIB00005736282	CCMSLIB00000206008
CCMSLIB00005736347	CCMSLIB00005771157	CCMSLIB00005771175	CCMSLIB00005771183
CCMSLIB00005771198	CCMSLIB00005771252	CCMSLIB00005771333	CCMSLIB00005771341
CCMSLIB00005771355	CCMSLIB00005771363	CCMSLIB00000206009	CCMSLIB00005771449
CCMSLIB00005771536	CCMSLIB00005771537	CCMSLIB00005771576	CCMSLIB00005771587

CCMSLIB00005772956	CCMSLIB00005773583	CCMSLIB00000206010	CCMSLIB00005774238
CCMSLIB00005774296	CCMSLIB00005775363	CCMSLIB00005776242	CCMSLIB00005776967
CCMSLIB00005777047	CCMSLIB00005787991	CCMSLIB00009919009	CCMSLIB00009919334
CCMSLIB00009919339	CCMSLIB00000206352	CCMSLIB00009919341	CCMSLIB00009919343
CCMSLIB00009919348	CCMSLIB00009919350	CCMSLIB00009919353	CCMSLIB00009919355
CCMSLIB00009919357	CCMSLIB00009919360	CCMSLIB00009919361	CCMSLIB00009919363
CCMSLIB00000206353	CCMSLIB00009919366	CCMSLIB00009919370	CCMSLIB00009919372
CCMSLIB00009919373	CCMSLIB00009919374	CCMSLIB00009919375	CCMSLIB00009919376
CCMSLIB00009919377	CCMSLIB00009919378	CCMSLIB00009919380	CCMSLIB00000206354
CCMSLIB00009919381	CCMSLIB00009919383	CCMSLIB00009919384	CCMSLIB00009919385
CCMSLIB00009919386	CCMSLIB00009919387	CCMSLIB00009919388	CCMSLIB00009919389
CCMSLIB00009919390	CCMSLIB00009919391	CCMSLIB00000206355	

Table A.8. The 6 nitazene opioid GNPS spectrum IDs.

CCMSLIB00009919331	CCMSLIB00009919413	CCMSLIB00009919499	CCMSLIB00010129295
CCMSLIB00010129380	CCMSLIB00011429501		

Table A.9. The 190 randomly selected GNPS spectrum IDs used as negative experimental examples.

CCMSLIB00005719490	CCMSLIB00005721549	CCMSLIB00005721704	CCMSLIB00005722687
CCMSLIB00005724125	CCMSLIB00005725018	CCMSLIB00005725195	CCMSLIB00005725699
CCMSLIB00005725845	CCMSLIB00005727551	CCMSLIB00005731142	CCMSLIB00005732505
CCMSLIB00005735186	CCMSLIB00005736125	CCMSLIB00005737106	CCMSLIB00005740941
CCMSLIB00005745570	CCMSLIB00005745640	CCMSLIB00005747671	CCMSLIB00005755059
CCMSLIB00005760221	CCMSLIB00005761095	CCMSLIB00000221205	CCMSLIB00005763578
CCMSLIB00005767793	CCMSLIB00005769011	CCMSLIB00005771021	CCMSLIB00005773975
CCMSLIB00005775728	CCMSLIB00005776492	CCMSLIB00005778332	CCMSLIB00005883220
CCMSLIB00005883958	CCMSLIB00005884159	CCMSLIB00006114793	CCMSLIB00006116314
CCMSLIB00006118880	CCMSLIB00006119965	CCMSLIB00006120386	CCMSLIB00006122225
CCMSLIB00006122503	CCMSLIB00006124528	CCMSLIB00006356280	CCMSLIB00006357312
CCMSLIB00006357733	CCMSLIB00006358212	CCMSLIB00006360170	CCMSLIB00006360746
CCMSLIB00006361282	CCMSLIB00006363246	CCMSLIB00006363773	CCMSLIB00006365396
CCMSLIB00006367164	CCMSLIB00000075068	CCMSLIB00006375490	CCMSLIB00000564935
CCMSLIB00006386439	CCMSLIB00006386722	CCMSLIB00006386966	CCMSLIB00006387753
CCMSLIB00006389921	CCMSLIB00006392943	CCMSLIB00000568076	CCMSLIB00006403021
CCMSLIB00000568189	CCMSLIB00006404984	CCMSLIB00006405144	CCMSLIB00006407287
CCMSLIB00006408045	CCMSLIB00000568455	CCMSLIB00006414886	CCMSLIB00006418290
CCMSLIB00006435202	CCMSLIB00006440568	CCMSLIB00000845288	CCMSLIB00006444046
CCMSLIB00006444458	CCMSLIB00006448080	CCMSLIB00000080535	CCMSLIB00006449751
CCMSLIB00006450696	CCMSLIB00006451377	CCMSLIB00006451999	CCMSLIB00000846393
CCMSLIB00000847895	CCMSLIB00000851185	CCMSLIB00003134914	CCMSLIB00003137768
CCMSLIB00003138225	CCMSLIB00004681484	CCMSLIB00006510826	CCMSLIB00006516754
CCMSLIB00006517974	CCMSLIB00006531762	CCMSLIB00006533995	CCMSLIB00006536419
CCMSLIB00000085906	CCMSLIB00006549954	CCMSLIB00006551851	CCMSLIB00006553567
CCMSLIB00006553720	CCMSLIB00006556949	CCMSLIB00006559257	CCMSLIB00006559555
CCMSLIB00006559928	CCMSLIB00006561009	CCMSLIB00006567911	CCMSLIB00006568433
CCMSLIB00006573536	CCMSLIB00006573564	CCMSLIB00006573795	CCMSLIB00006575143
CCMSLIB00006583665	CCMSLIB00006583757	CCMSLIB00006673147	CCMSLIB00006675134
CCMSLIB00006677164	CCMSLIB00006678644	CCMSLIB00006680371	CCMSLIB00006684852
CCMSLIB00008851290	CCMSLIB00010007654	CCMSLIB00010104643	CCMSLIB00010105811
CCMSLIB00010106353	CCMSLIB00010106808	CCMSLIB00010106869	CCMSLIB00010107434
CCMSLIB00010107855	CCMSLIB00010108666	CCMSLIB00010108740	CCMSLIB00010109918
CCMSLIB00010110002	CCMSLIB00010110477	CCMSLIB00010113295	CCMSLIB00004690181
CCMSLIB00010115582	CCMSLIB00010115764	CCMSLIB00010117824	CCMSLIB00010118655
CCMSLIB00010121875	CCMSLIB00000204841	CCMSLIB00004691299	CCMSLIB00004691589
CCMSLIB00004693283	CCMSLIB00004693627	CCMSLIB00004695482	CCMSLIB00004695494
CCMSLIB00004695705	CCMSLIB00004695811	CCMSLIB00004696635	CCMSLIB00004696675
CCMSLIB00004696724	CCMSLIB00004698035	CCMSLIB00004698424	CCMSLIB00004699132

CCMSLIB00004701224	CCMSLIB00004701654	CCMSLIB00004701874	CCMSLIB00004701962
CCMSLIB00004702649	CCMSLIB00004703256	CCMSLIB00004703905	CCMSLIB00004705040
CCMSLIB00004705829	CCMSLIB00004705945	CCMSLIB00004708258	CCMSLIB00004708418
CCMSLIB00004708963	CCMSLIB00004710142	CCMSLIB00000207537	CCMSLIB00004712312
CCMSLIB00004712536	CCMSLIB00004713473	CCMSLIB00004713523	CCMSLIB00004713863
CCMSLIB00004716859	CCMSLIB00004716894	CCMSLIB00004717580	CCMSLIB00004717843
CCMSLIB00000208243	CCMSLIB00004720565	CCMSLIB00004721491	CCMSLIB00004721592
CCMSLIB00005463990	CCMSLIB00005464017	CCMSLIB00005464566	CCMSLIB00005489750
CCMSLIB00005716812	CCMSLIB00000209811		

Table A.10. The morphinan opioid spectrum metadata from the GNPS Drugs and Metabolites Library.

gnps_libid	name_compound	name_parent_compound	chemical_source
CCMSLIB00009919478	buprenorphine	buprenorphine	Medical
CCMSLIB00000206061	buprenorphine	buprenorphine	Medical
CCMSLIB00000206062	buprenorphine	buprenorphine	Medical
CCMSLIB00000206063	buprenorphine	buprenorphine	Medical
CCMSLIB00000206064	buprenorphine	buprenorphine	Medical
CCMSLIB00000206065	buprenorphine	buprenorphine	Medical
CCMSLIB00004722159	buprenorphine	buprenorphine	Medical
CCMSLIB00005771219	buprenorphine	buprenorphine	Medical
CCMSLIB00005771300	buprenorphine	buprenorphine	Medical
CCMSLIB00005771459	buprenorphine	buprenorphine	Medical
CCMSLIB00005771464	buprenorphine	buprenorphine	Medical
CCMSLIB00005771493	buprenorphine	buprenorphine	Medical
CCMSLIB00005774685	buprenorphine metabolite	buprenorphine	Drug metabolite
CCMSLIB00004721850	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772121	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772284	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772510	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772532	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772569	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005773047	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005773509	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005776799	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005777198	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00009919520	norbuprenorphine	buprenorphine	Drug metabolite
CCMSLIB00005772607	diamorphine	morphine	Drug metabolite
CCMSLIB00005773248	diamorphine	morphine	Drug metabolite
CCMSLIB00005774985	diamorphine	morphine	Drug metabolite
CCMSLIB00005774996	diamorphine	morphine	Drug metabolite
CCMSLIB00005775285	diamorphine	morphine	Drug metabolite
CCMSLIB00005776192	diamorphine	morphine	Drug metabolite
CCMSLIB00005776399	diamorphine	morphine	Drug metabolite
CCMSLIB00005776447	diamorphine	morphine	Drug metabolite
CCMSLIB00005776735	diamorphine	morphine	Drug metabolite
CCMSLIB00005777077	diamorphine	morphine	Drug metabolite
CCMSLIB00000208789	codeine	codeine	Medical
CCMSLIB00000210639	codeine	codeine	Medical
CCMSLIB00000210969	codeine	codeine	Medical
CCMSLIB00000210971	codeine	codeine	Medical
CCMSLIB00000210973	codeine	codeine	Medical
CCMSLIB00000210975	codeine	codeine	Medical
CCMSLIB00000210977	codeine	codeine	Medical
CCMSLIB00000210979	codeine	codeine	Medical
CCMSLIB00000210982	codeine	codeine	Medical
CCMSLIB00000210984	codeine	codeine	Medical
CCMSLIB00000210986	codeine	codeine	Medical
CCMSLIB00000210987	codeine	codeine	Medical
CCMSLIB00000210989	codeine	codeine	Medical
CCMSLIB00000210991	codeine	codeine	Medical
CCMSLIB00000210993	codeine	codeine	Medical

CCMSLIB00000210996	codeine	codeine	Medical
CCMSLIB00000222130	codeine	codeine	Medical
CCMSLIB00000222132	codeine	codeine	Medical
CCMSLIB00000222134	codeine	codeine	Medical
CCMSLIB00004679363	codeine	codeine	Medical
CCMSLIB00005725413	codeine	codeine	Medical
CCMSLIB00005725414	codeine	codeine	Medical
CCMSLIB00005725415	codeine	codeine	Medical
CCMSLIB00005725416	codeine	codeine	Medical
CCMSLIB00005725417	codeine	codeine	Medical
CCMSLIB00005725418	codeine	codeine	Medical
CCMSLIB00005772818	codeine	codeine	Medical
CCMSLIB00005772853	codeine	codeine	Medical
CCMSLIB00005772877	codeine	codeine	Medical
CCMSLIB00005773572	codeine	codeine	Medical
CCMSLIB00005773797	codeine	codeine	Medical
CCMSLIB00005774387	codeine	codeine	Medical
CCMSLIB00005775043	codeine	codeine	Medical
CCMSLIB00005775049	codeine	codeine	Medical
CCMSLIB00005775107	codeine	codeine	Medical
CCMSLIB00005776397	codeine	codeine	Medical
CCMSLIB00005776430	codeine	codeine	Medical
CCMSLIB00005776725	codeine	codeine	Medical
CCMSLIB00005776920	codeine	codeine	Medical
CCMSLIB00008851235	codeine	codeine	Medical
CCMSLIB00009919490	codeine	codeine	Medical
CCMSLIB00010007706	codeine	codeine	Medical
CCMSLIB00010011926	codeine	codeine	Medical
CCMSLIB00000568275	hydrocodone	hydrocodone	Medical
CCMSLIB00000569838	hydrocodone	hydrocodone	Medical
CCMSLIB00000570108	hydrocodone	hydrocodone	Medical
CCMSLIB00000570256	hydrocodone	hydrocodone	Medical
CCMSLIB00005730093	hydrocodone	hydrocodone	Medical
CCMSLIB00005731132	hydrocodone	hydrocodone	Medical
CCMSLIB00005732171	hydrocodone	hydrocodone	Medical
CCMSLIB00005733056	hydrocodone	hydrocodone	Medical
CCMSLIB00005735253	hydrocodone	hydrocodone	Medical
CCMSLIB00005736317	hydrocodone	hydrocodone	Medical
CCMSLIB00009919506	hydrocodone	hydrocodone	Medical
CCMSLIB00000208787	morphine	morphine	Medical
CCMSLIB00000210680	morphine	morphine	Medical
CCMSLIB00000210682	morphine	morphine	Medical
CCMSLIB00000210684	morphine	morphine	Medical
CCMSLIB00000210686	morphine	morphine	Medical
CCMSLIB00000210688	morphine	morphine	Medical
CCMSLIB00000210690	morphine	morphine	Medical
CCMSLIB00000210692	morphine	morphine	Medical
CCMSLIB00000210694	morphine	morphine	Medical
CCMSLIB00000210696	morphine	morphine	Medical
CCMSLIB00000210698	morphine	morphine	Medical
CCMSLIB00000210700	morphine	morphine	Medical
CCMSLIB00000210702	morphine	morphine	Medical
CCMSLIB00000210705	morphine	morphine	Medical
CCMSLIB00000210707	morphine	morphine	Medical

CCMSLIB00005772535	morphine	morphine	Medical
CCMSLIB00005772667	morphine	morphine	Medical
CCMSLIB00005772814	morphine	morphine	Medical
CCMSLIB00005773114	morphine	morphine	Medical
CCMSLIB00005773630	morphine	morphine	Medical
CCMSLIB00005773677	morphine	morphine	Medical
CCMSLIB00005774617	morphine	morphine	Medical
CCMSLIB00005774948	morphine	morphine	Medical
CCMSLIB00005775818	morphine	morphine	Medical
CCMSLIB00005775857	morphine	morphine	Medical
CCMSLIB00005775925	morphine	morphine	Medical
CCMSLIB00005775989	morphine	morphine	Medical
CCMSLIB00005776231	morphine	morphine	Medical
CCMSLIB00008851323	morphine	morphine	Medical
CCMSLIB00009919518	morphine	morphine	Medical
CCMSLIB00000206297	oxycodone	oxycodone	Medical
CCMSLIB00000206298	oxycodone	oxycodone	Medical
CCMSLIB00000206299	oxycodone	oxycodone	Medical
CCMSLIB00000206300	oxycodone	oxycodone	Medical
CCMSLIB00000206301	oxycodone	oxycodone	Medical
CCMSLIB00000222142	oxycodone	oxycodone	Medical
CCMSLIB00000222145	oxycodone	oxycodone	Medical
CCMSLIB00000222147	oxycodone	oxycodone	Medical
CCMSLIB00005771254	oxycodone	oxycodone	Medical
CCMSLIB00005771315	oxycodone	oxycodone	Medical
CCMSLIB00005771337	oxycodone	oxycodone	Medical
CCMSLIB00005771345	oxycodone	oxycodone	Medical
CCMSLIB00005771436	oxycodone	oxycodone	Medical
CCMSLIB00005772220	oxycodone	oxycodone	Medical
CCMSLIB00005772890	oxycodone	oxycodone	Medical
CCMSLIB00005773239	oxycodone	oxycodone	Medical
CCMSLIB00005773613	oxycodone	oxycodone	Medical
CCMSLIB00005773942	oxycodone	oxycodone	Medical
CCMSLIB00005774216	oxycodone	oxycodone	Medical
CCMSLIB00005774565	oxycodone	oxycodone	Medical
CCMSLIB00005774727	oxycodone	oxycodone	Medical
CCMSLIB00005774854	oxycodone	oxycodone	Medical
CCMSLIB00005774994	oxycodone	oxycodone	Medical
CCMSLIB00005776067	oxycodone	oxycodone	Medical
CCMSLIB00005776607	oxycodone	oxycodone	Medical
CCMSLIB00005776790	oxycodone	oxycodone	Medical
CCMSLIB00005777033	oxycodone	oxycodone	Medical
CCMSLIB00009919525	oxycodone	oxycodone	Medical
CCMSLIB00009919526	oxymorphone	oxymorphone	Medical
CCMSLIB00005731857	codeine metabolite	codeine	Drug metabolite
CCMSLIB00005772906	codeine metabolite	codeine	Drug metabolite
CCMSLIB00005774692	codeine metabolite	codeine	Drug metabolite
CCMSLIB00005774821	codeine metabolite	codeine	Drug metabolite
CCMSLIB00005776120	codeine metabolite	codeine	Drug metabolite
CCMSLIB00000210643	codeine-n-oxide	codeine	Drug metabolite
CCMSLIB00005730105	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005731756	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005734240	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005734603	dihydrocodeine	codeine	Drug metabolite

CCMSLIB00005735031	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005772197	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005772272	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005773269	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005773522	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005773562	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005773884	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005773987	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005776407	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00005777114	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00000568032	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00000568266	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00000568273	dihydrocodeine	codeine	Drug metabolite
CCMSLIB00009919507	hydromorphone	hydromorphone	Drug metabolite
CCMSLIB00000206252	morphine metabolite	morphine	Drug metabolite
CCMSLIB00000206253	morphine metabolite	morphine	Drug metabolite
CCMSLIB00005771210	morphine metabolite	morphine	Drug metabolite
CCMSLIB00005771269	morphine metabolite	morphine	Drug metabolite
CCMSLIB00000206254	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00000206255	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00000206256	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00005771339	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00005771492	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00005771494	morphine_3_glucuronide	morphine	Drug metabolite
CCMSLIB00000210641	norcodeine	codeine	Drug metabolite
CCMSLIB00010142146	nalfurafine	nalfurafine	Medical
CCMSLIB00010142147	nalfurafine	nalfurafine	Medical
CCMSLIB00010142149	nalfurafine	nalfurafine	Medical
CCMSLIB00010142145	nalfurafine	nalfurafine	Medical
CCMSLIB00010142143	nalfurafine	nalfurafine	Medical
CCMSLIB00010142148	nalfurafine	nalfurafine	Medical
CCMSLIB00010142144	nalfurafine	nalfurafine	Medical
CCMSLIB00010142150	nalfurafine	nalfurafine	Medical
CCMSLIB00000078795	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00000079179	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00000084974	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00003138432	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00004679248	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00004679249	levorphanol	dextromethorphan	Medical|Drug metabolite
CCMSLIB00000079174	nalbuphine	nalbuphine	Medical
CCMSLIB00000084982	nalbuphine	nalbuphine	Medical
CCMSLIB00000085926	nalbuphine	nalbuphine	Medical
CCMSLIB00003138388	nalbuphine	nalbuphine	Medical
CCMSLIB00003138410	nalbuphine	nalbuphine	Medical
CCMSLIB00000206197	butorphanol metabolite	butorphanol	Drug metabolite
CCMSLIB00000206198	butorphanol metabolite	butorphanol	Drug metabolite
CCMSLIB00005771190	butorphanol metabolite	butorphanol	Drug metabolite
CCMSLIB00005771607	butorphanol metabolite	butorphanol	Drug metabolite
CCMSLIB00000206199	hydroxybutorphanol	butorphanol	Drug metabolite
CCMSLIB00000206200	hydroxybutorphanol	butorphanol	Drug metabolite
CCMSLIB00000206201	hydroxybutorphanol	butorphanol	Drug metabolite
CCMSLIB00005771432	hydroxybutorphanol	butorphanol	Drug metabolite
CCMSLIB00005771512	hydroxybutorphanol	butorphanol	Drug metabolite
CCMSLIB00005771585	hydroxybutorphanol	butorphanol	Drug metabolite

CCMSLIB00012465953	nalbuphine	nalbuphine	Medical
CCMSLIB00012465954	nalbuphine	nalbuphine	Medical
CCMSLIB00012465955	nalbuphine	nalbuphine	Medical
CCMSLIB00012465956	nalbuphine	nalbuphine	Medical
CCMSLIB00012465957	nalbuphine_metabolite_004	nalbuphine	Drug metabolite
CCMSLIB00012465958	nalbuphine_metabolite_006	nalbuphine	Drug metabolite
CCMSLIB00012465959	nalbuphine_metabolite_017	nalbuphine	Drug metabolite
CCMSLIB00010141341	naloxegol	naloxegol	Medical
CCMSLIB00010141338	naloxegol	naloxegol	Medical
CCMSLIB00010141340	naloxegol	naloxegol	Medical
CCMSLIB00010141339	naloxegol	naloxegol	Medical
CCMSLIB00010141342	naloxegol	naloxegol	Medical
CCMSLIB00010141350	naloxegol	naloxegol	Medical
CCMSLIB00010141349	naloxegol	naloxegol	Medical
CCMSLIB00010141332	naloxegol	naloxegol	Medical
CCMSLIB00010141344	naloxegol	naloxegol	Medical
CCMSLIB00010141337	naloxegol	naloxegol	Medical
CCMSLIB00010141335	naloxegol	naloxegol	Medical
CCMSLIB00010141333	naloxegol	naloxegol	Medical
CCMSLIB00010141334	naloxegol	naloxegol	Medical
CCMSLIB00010141336	naloxegol	naloxegol	Medical
CCMSLIB00010141345	naloxegol	naloxegol	Medical
CCMSLIB00010141346	naloxegol	naloxegol	Medical
CCMSLIB00010141348	naloxegol	naloxegol	Medical
CCMSLIB00010141347	naloxegol	naloxegol	Medical
CCMSLIB00010141343	naloxegol	naloxegol	Medical
CCMSLIB00009919519	naloxone	naloxone	Medical
CCMSLIB00000077056	naloxone	naloxone	Medical
CCMSLIB00000078568	naloxone	naloxone	Medical
CCMSLIB00000084852	naloxone	naloxone	Medical
CCMSLIB00000085244	naloxone	naloxone	Medical
CCMSLIB00000085770	naloxone	naloxone	Medical
CCMSLIB00004721185	naloxone	naloxone	Medical
CCMSLIB00004721186	naloxone	naloxone	Medical
CCMSLIB00004721187	naloxone	naloxone	Medical
CCMSLIB00004721188	naloxone	naloxone	Medical
CCMSLIB00004721189	naloxone	naloxone	Medical
CCMSLIB00004721190	naloxone	naloxone	Medical
CCMSLIB00004721191	naloxone	naloxone	Medical
CCMSLIB00004721192	naloxone	naloxone	Medical
CCMSLIB00004721193	naloxone	naloxone	Medical
CCMSLIB00004721194	naloxone	naloxone	Medical
CCMSLIB00004721195	naloxone	naloxone	Medical
CCMSLIB00004721196	naloxone	naloxone	Medical
CCMSLIB00004721197	naloxone	naloxone	Medical
CCMSLIB00004721198	naloxone	naloxone	Medical
CCMSLIB00004721199	naloxone	naloxone	Medical
CCMSLIB00000078434	naltrexone	naltrexone	Medical
CCMSLIB00000085211	naltrexone	naltrexone	Medical
CCMSLIB00000206257	naltrexone	naltrexone	Medical
CCMSLIB00000206258	naltrexone	naltrexone	Medical
CCMSLIB00000206259	naltrexone	naltrexone	Medical
CCMSLIB00000206260	naltrexone	naltrexone	Medical
CCMSLIB00000206261	naltrexone	naltrexone	Medical

CCMSLIB00000211754	naltrexone	naltrexone	Medical
CCMSLIB00000211756	naltrexone	naltrexone	Medical
CCMSLIB00000211758	naltrexone	naltrexone	Medical
CCMSLIB00000211760	naltrexone	naltrexone	Medical
CCMSLIB00000211762	naltrexone	naltrexone	Medical
CCMSLIB00000211764	naltrexone	naltrexone	Medical
CCMSLIB00000211766	naltrexone	naltrexone	Medical
CCMSLIB00000211768	naltrexone	naltrexone	Medical
CCMSLIB00000211770	naltrexone	naltrexone	Medical
CCMSLIB00000211772	naltrexone	naltrexone	Medical
CCMSLIB00000211774	naltrexone	naltrexone	Medical
CCMSLIB00000211776	naltrexone	naltrexone	Medical
CCMSLIB00000211778	naltrexone	naltrexone	Medical
CCMSLIB00000211780	naltrexone	naltrexone	Medical
CCMSLIB00003134518	naltrexone	naltrexone	Medical
CCMSLIB00003138452	naltrexone	naltrexone	Medical
CCMSLIB00005771294	naltrexone	naltrexone	Medical
CCMSLIB00005771362	naltrexone	naltrexone	Medical
CCMSLIB00005771439	naltrexone	naltrexone	Medical
CCMSLIB00005771481	naltrexone	naltrexone	Medical
CCMSLIB00005771542	naltrexone	naltrexone	Medical
CCMSLIB00009919317	naltrexone	naltrexone	Medical
CCMSLIB00012465291	naloxone	naloxone	Medical
CCMSLIB00012465292	naloxone	naloxone	Medical
CCMSLIB00012465293	naloxone	naloxone	Medical
CCMSLIB00012465294	naloxone_metabolite_005	naloxone	Drug metabolite
CCMSLIB00012465295	naloxone_metabolite_020	naloxone	Drug metabolite
CCMSLIB00012465296	naloxone_metabolite_022	naloxone	Drug metabolite
CCMSLIB00012467249	naltrexone	naltrexone	Medical
CCMSLIB00012467250	naltrexone	naltrexone	Medical
CCMSLIB00012467251	naltrexone	naltrexone	Medical
CCMSLIB00012467252	naltrexone	naltrexone	Medical
CCMSLIB00012467253	naltrexone_metabolite_001	naltrexone	Drug metabolite
CCMSLIB00012467254	naltrexone_metabolite_009	naltrexone	Drug metabolite
CCMSLIB00012467255	naltrexone_metabolite_011	naltrexone	Drug metabolite
CCMSLIB00012467256	naltrexone_metabolite_017	naltrexone	Drug metabolite
CCMSLIB00012467257	naltrexone_metabolite_019	naltrexone	Drug metabolite
CCMSLIB00012467258	naltrexone_metabolite_019	naltrexone	Drug metabolite
CCMSLIB00012467259	naltrexone_metabolite_020	naltrexone	Drug metabolite

Table A.11. The fentanyl opioid spectrum metadata from the GNPS Drugs and Metabolites Library

gnps_libid	name_compound	name_parent_compound	chemical_ source
CCMSLIB000 09919373	2,3-benzodioxole fentanyl	2,3-benzodioxole fentanyl	Medical
CCMSLIB000 09919375	2',3'-dimethoxy fentanyl (hydrochloride)	2',3'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919377	2',4'-dimethoxy fentanyl (hydrochloride)	2',4'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919339	2',5'-dimethoxy fentanyl (hydrochloride)	2',5'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919378	2',6'-dimethoxy fentanyl (hydrochloride)	2',6'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919384	3',4'-dimethoxy fentanyl (hydrochloride)	3',4'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919386	3',5'-dimethoxy fentanyl (hydrochloride)	3',5'-dimethoxy fentanyl (hydrochloride)	Medical
CCMSLIB000 09919370	n-(2,5-dma) fentanyl (hydrochloride)	n-(2,5-dma) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919406	n-(3,4,5-tma) fentanyl (hydrochloride)	n-(3,4,5-tma) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919353	(?)-cis-3-methyl norfentanyl	(?)-cis-3-methyl norfentanyl	Medical
CCMSLIB000 09919355	(?)-cis-isofentanyl (hydrochloride)	(?)-cis-isofentanyl (hydrochloride)	Medical
CCMSLIB000 09919380	2'-fluorofentanyl (hydrochloride)	2'-fluorofentanyl (hydrochloride)	Medical
CCMSLIB000 09919341	2'-methyl acetyl fentanyl (hydrochloride)	2'-methyl acetyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919343	2'-methyl fentanyl (hydrochloride)	2'-methyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919390	3-fluorofentanyl (hydrochloride)	3-fluorofentanyl (hydrochloride)	Medical
CCMSLIB000 09919388	3'-fluoro?ortho-fluorofentanyl (hydrochloride)	3'-fluoro?ortho-fluorofentanyl (hydrochloride)	Medical
CCMSLIB000 09919348	3'-methyl acetyl fentanyl (hydrochloride)	3'-methyl acetyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919399	4'-methyl fentanyl (hydrochloride)	4'-methyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919409	cyclopropaneacetyl fentanyl (hydrochloride)	cyclopropaneacetyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919415	meta-fluoro acrylfentanyl	meta-fluoro acrylfentanyl	Medical
CCMSLIB000 09919363	meta-fluoro furanyl fentanyl (hydrochloride)	meta-fluoro furanyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919417	meta-fluoro valeryl fentanyl (hydrochloride)	meta-fluoro valeryl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919420	meta-methoxy furanyl fentanyl	meta-methoxy furanyl fentanyl	Medical
CCMSLIB000 09919366	meta-methyl acetyl fentanyl (hydrochloride)	meta-methyl acetyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919372	n-(2-apb) fentanyl	n-(2-apb) fentanyl	Medical

CCMSLIB000 09919376	n-(2c-b-fly) fentanyl (hydrochloride)	n-(2c-b-fly) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919374	n-(2c-b) fentanyl (hydrochloride)	n-(2c-b) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919381	n-(2c-c) fentanyl (hydrochloride)	n-(2c-c) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919383	n-(2c-d) fentanyl (hydrochloride)	n-(2c-d) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919385	n-(2c-e) fentanyl (hydrochloride)	n-(2c-e) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919387	n-(2c-g) fentanyl (hydrochloride)	n-(2c-g) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919389	n-(2c-i) fentanyl (hydrochloride)	n-(2c-i) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919391	n-(2c-ip) fentanyl (hydrochloride)	n-(2c-ip) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919393	n-(2c-n) fentanyl (hydrochloride)	n-(2c-n) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919394	n-(2c-p) fentanyl (hydrochloride)	n-(2c-p) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919397	n-(2c-t-2) fentanyl (hydrochloride)	n-(2c-t-2) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919400	n-(2c-t-4) fentanyl (hydrochloride)	n-(2c-t-4) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919402	n-(2c-t-7) fentanyl (hydrochloride)	n-(2c-t-7) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919395	n-(2c-t) fentanyl	n-(2c-t) fentanyl	Medical
CCMSLIB000 09919404	n-(2c-tfm) fentanyl (hydrochloride)	n-(2c-tfm) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919408	n-(3c-b-fly) fentanyl (hydrochloride)	n-(3c-b-fly) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919410	n-(6-apb) fentanyl	n-(6-apb) fentanyl	Medical
CCMSLIB000 09919412	n-(6-apdb) fentanyl	n-(6-apdb) fentanyl	Medical
CCMSLIB000 09919422	n-(dob) fentanyl (hydrochloride)	n-(dob) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919424	n-(dobu) fentanyl (hydrochloride)	n-(dobu) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919427	n-(doc) fentanyl (hydrochloride)	n-(doc) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919429	n-(doet) fentanyl (hydrochloride)	n-(doet) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919431	n-(doi) fentanyl (hydrochloride)	n-(doi) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919433	n-(dom) fentanyl (hydrochloride)	n-(dom) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919434	n-(mda) fentanyl (hydrochloride)	n-(mda) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919439	n-(phentermine) fentanyl (hydrochloride)	n-(phentermine) fentanyl (hydrochloride)	Medical
CCMSLIB000 09919418	n-benzyl phenyl norfentanyl (hydrochloride)	n-benzyl phenyl norfentanyl (hydrochloride)	Medical

CCMSLIB000 09919414	n-benzyl?para-fluoro cyclopropyl norfentanyl (hydrochloride)	n-benzyl?para-fluoro cyclopropyl norfentanyl (hydrochloride)	Medical
CCMSLIB000 09919416	n-benzyl?para-fluoro norfentanyl (hydrochloride)	n-benzyl?para-fluoro norfentanyl (hydrochloride)	Medical
CCMSLIB000 09919435	n-methyl norfentanyl (hydrochloride)	n-methyl norfentanyl (hydrochloride)	Medical
CCMSLIB000 09919426	ortho-fluoro valeryl fentanyl (hydrochloride)	ortho-fluoro valeryl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919443	para-bromofentanyl	para-bromofentanyl	Medical
CCMSLIB000 09919334	para-chloro acrylfentanyl	para-chloro acrylfentanyl	Medical
CCMSLIB000 09919428	para-chloro furanyl fentanyl 3-furancarboxamide	para-chloro furanyl fentanyl 3-furancarboxamide	Medical
CCMSLIB000 09919430	para-chloroacetyl fentanyl (hydrochloride)	para-chloroacetyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919432	para-hydroxy butyryl fentanyl	para-hydroxy butyryl fentanyl	Medical
CCMSLIB000 09919447	para-toluoyl fentanyl (hydrochloride)	para-toluoyl fentanyl (hydrochloride)	Medical
CCMSLIB000 09919437	thiophene fentanyl 3-thiophenecarboxamide (hydrochloride)	thiophene fentanyl 3-thiophenecarboxamide (hydrochloride)	Medical
CCMSLIB000 09919440	tigloyl fentanyl	tigloyl fentanyl	Medical
CCMSLIB000 09919350	3-methyl fentanyl	3-methyl fentanyl	Medical
CCMSLIB000 05787991	acrylfentanyl	acrylfentanyl	Medical
CCMSLIB000 00206006	alfentanil	alfentanil	Medical
CCMSLIB000 00206007	alfentanil	alfentanil	Medical
CCMSLIB000 00206008	alfentanil	alfentanil	Medical
CCMSLIB000 00206009	alfentanil	alfentanil	Medical
CCMSLIB000 00206010	alfentanil	alfentanil	Medical
CCMSLIB000 05771449	alfentanil	alfentanil	Medical
CCMSLIB000 05771536	alfentanil	alfentanil	Medical
CCMSLIB000 05771537	alfentanil	alfentanil	Medical
CCMSLIB000 05771587	alfentanil	alfentanil	Medical
CCMSLIB000 05771599	alfentanil	alfentanil	Medical
CCMSLIB000 00565838	fentanyl	fentanyl	Medical
CCMSLIB000 00568443	fentanyl	fentanyl	Medical
CCMSLIB000 00568476	fentanyl	fentanyl	Medical

CCMSLIB000 00568546	fentanyl	fentanyl	Medical
CCMSLIB000 00568596	fentanyl	fentanyl	Medical
CCMSLIB000 00568626	fentanyl	fentanyl	Medical
CCMSLIB000 00568690	fentanyl	fentanyl	Medical
CCMSLIB000 00568713	fentanyl	fentanyl	Medical
CCMSLIB000 00569036	fentanyl	fentanyl	Medical
CCMSLIB000 03135084	fentanyl	fentanyl	Medical
CCMSLIB000 03135626	fentanyl	fentanyl	Medical
CCMSLIB000 03136417	fentanyl	fentanyl	Medical
CCMSLIB000 03137026	fentanyl	fentanyl	Medical
CCMSLIB000 03139043	fentanyl	fentanyl	Medical
CCMSLIB000 03140014	fentanyl	fentanyl	Medical
CCMSLIB000 05730959	fentanyl	fentanyl	Medical
CCMSLIB000 05731434	fentanyl	fentanyl	Medical
CCMSLIB000 05731800	fentanyl	fentanyl	Medical
CCMSLIB000 05732920	fentanyl	fentanyl	Medical
CCMSLIB000 05733673	fentanyl	fentanyl	Medical
CCMSLIB000 05734519	fentanyl	fentanyl	Medical
CCMSLIB000 05735321	fentanyl	fentanyl	Medical
CCMSLIB000 05736118	fentanyl	fentanyl	Medical
CCMSLIB000 05736282	fentanyl	fentanyl	Medical
CCMSLIB000 05772528	fentanyl	fentanyl	Medical
CCMSLIB000 05774214	fentanyl	fentanyl	Medical
CCMSLIB000 05774296	fentanyl	fentanyl	Medical
CCMSLIB000 05775363	fentanyl	fentanyl	Medical
CCMSLIB000 05776242	fentanyl	fentanyl	Medical
CCMSLIB000 09919009	fentanyl	fentanyl	Medical

CCMSLIB000 00206352	remifentanil	remifentanil	Medical
CCMSLIB000 00206353	remifentanil	remifentanil	Medical
CCMSLIB000 00206354	remifentanil	remifentanil	Medical
CCMSLIB000 00206355	remifentanil	remifentanil	Medical
CCMSLIB000 00206356	remifentanil	remifentanil	Medical
CCMSLIB000 05771175	remifentanil	remifentanil	Medical
CCMSLIB000 05771198	remifentanil	remifentanil	Medical
CCMSLIB000 05771333	remifentanil	remifentanil	Medical
CCMSLIB000 05771341	remifentanil	remifentanil	Medical
CCMSLIB000 05771355	remifentanil	remifentanil	Medical
CCMSLIB000 09919444	remifentanil	remifentanil	Medical
CCMSLIB000 00206392	sufentanil	sufentanil	Medical
CCMSLIB000 00206393	sufentanil	sufentanil	Medical
CCMSLIB000 00206394	sufentanil	sufentanil	Medical
CCMSLIB000 00206395	sufentanil	sufentanil	Medical
CCMSLIB000 00206396	sufentanil	sufentanil	Medical
CCMSLIB000 05771157	sufentanil	sufentanil	Medical
CCMSLIB000 05771183	sufentanil	sufentanil	Medical
CCMSLIB000 05771252	sufentanil	sufentanil	Medical
CCMSLIB000 05771363	sufentanil	sufentanil	Medical
CCMSLIB000 05771576	sufentanil	sufentanil	Medical
CCMSLIB000 05730630	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05730640	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05731011	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05732135	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05732921	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05734514	norfentanyl	fentanyl	Drug metabolite

CCMSLIB000 05735192	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05735795	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05736347	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05772956	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05773583	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05774238	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05776967	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 05777047	norfentanyl	fentanyl	Drug metabolite
CCMSLIB000 09919446	remifentanil acid	remifentanil	Drug metabolite
CCMSLIB000 09919451	remifentanil acid (trifluoroacetate salt)	remifentanil	Drug metabolite

Table A.12. The nitazene opioid spectrum metadata from the GNPS Drugs and Metabolites Library

gnps_libid	name_compound	name_parent_compound	chemical_source
CCMSLIB00009919413	etonitazene	etonitazene	Medical
CCMSLIB00009919331	isotonitazene	isotonitazene	Medical
CCMSLIB00009919499	metonitazene	metonitazene	Medical

Table A.13. Opioid Identifications per sample from Barkholtz lab spectral library matching.

SID	Identified Compounds	Morphinan	Fentanyl	Nitazene	Morphinan Number	Fentanyl Number	Nitazene Number
22UW-0001		0	0	0			
22UW-0002		0	0	0			
22UW-0003	Fentanyl	0	1	0		1	
22UW-0004		0	0	0			
22UW-0005		0	0	0			
22UW-0006	Codeine, Oxycodone	1	0	0	2		
22UW-0007		0	0	0			
22UW-0008	Fentanyl	0	1	0		1	
22UW-0009	Norbuprenorphine	1	0	0	1		
22UW-0010		0	0	0			
22UW-0011	Fentanyl	0	1	0		1	
22UW-0012		0	0	0			
22UW-0013	Hydrocodone	1	0	0	1		
22UW-0014		0	0	0			
22UW-0015		0	0	0			
22UW-0016		0	0	0			
22UW-0017	Dextromethorphan	1	0	0	1		
22UW-0018		0	0	0			
22UW-0019		0	0	0			
22UW-0020		0	0	0			
22UW-0021	Fentanyl	0	1	0		1	
22UW-0022	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0023	Fentanyl	0	1	0		1	
22UW-0024		0	0	0			
22UW-0025		0	0	0			
22UW-0026		0	0	0			
22UW-0027	Oxycodone	1	0	0	1		
22UW-0028	Dextromethorphan, Fentanyl	1	1	0	1	1	
22UW-0029		0	0	0			
22UW-0030		0	0	0			
22UW-0031	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0032	Fentanyl	0	1	0	1		
22UW-0033		0	0	0			
22UW-0034	Dextromethorphan	1	0	0	1		
22UW-0035		0	0	0			
22UW-0036		0	0	0			
22UW-0037		0	0	0			

ID	Drug						
22UW-0038		0	0	0			
22UW-0039		0	0	0			
22UW-0040		0	0	0			
22UW-0041	Fentanyl	0	1	0		1	
22UW-0042		0	0	0			
22UW-0043		0	0	0			
22UW-0044		0	0	0			
22UW-0045		0	0	0			
22UW-0046		0	0	0			
22UW-0047		0	0	0			
22UW-0048		0	0	0			
22UW-0049		0	0	0			
22UW-0050	Naltrexone	1	0	0	1		
22UW-0051		0	0	0			
22UW-0052		0	0	0			
22UW-0053		0	0	0			
22UW-0054		0	0	0			
22UW-0055		0	0	0			
22UW-0056	Fentanyl, Naloxone	1	1	0	1	1	
22UW-0057		0	0	0			
22UW-0058		0	0	0			
22UW-0059		0	0	0			
22UW-0061	Isotonitazene, Metonitazene	0	0	1			2
22UW-0060		0	0	0			
22UW-0062		0	0	0			
22UW-0063		0	0	0			
22UW-0064	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		6	
22UW-0065		0	0	0			
22UW-0066		0	0	0			
22UW-0067		0	0	0			
22UW-0068	Hydrocodone	1	0	0	1		
22UW-0069	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0070		0	0	0			
22UW-0071		0	0	0			
22UW-0072		0	0	0			
22UW-0073		0	0	0			
22UW-0074		0	0	0			
22UW-0075	Fentanyl	0	1	0		1	
22UW-0076	Fentanyl	0	1	0		1	

22UW-0077		0	0	0			
22UW-0078		0	0	0			
22UW-0079		0	0	0			
22UW-0080		0	0	0			
22UW-0081		0	0	0			
22UW-0082	Fentanyl	0	1	0		1	
22UW-0083	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, Fentanyl	0	1	0		4	
22UW-0084		0	0	0			
22UW-0085		0	0	0			
22UW-0086	Naloxone	1	0	0	1		
22UW-0087		0	0	0			
22UW-0088		0	0	0			
22UW-0089	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl	0	1	0		3	
22UW-0090		0	0	0			
22UW-0091		0	0	0			
22UW-0092		0	0	0			
22UW-0093		0	0	0			
22UW-0094		0	0	0			
22UW-0095	beta-hydroxyfentanyl	0	1	0		1	
22UW-0096	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, Norfentanyl, Morphine	1	1	0	1	4	
22UW-0097	Norfentanyl	0	1	0		1	
22UW-0098		0	0	0			
22UW-0099	Acetylfentanyl, beta-hydroxyfentanyl, buprenorphine, Fentanyl, Norfentanyl	1	1	0	1	4	
22UW-0100		0	0	0			
22UW-0101		0	0	0			
22UW-0102		0	0	0			
22UW-0103		0	0	0			
22UW-0104		0	0	0			
22UW-0105		0	0	0			
22UW-0106		0	0	0			
22UW-0107		0	0	0			
22UW-0108		0	0	0			
22UW-0109		0	0	0			
22UW-0110		0	0	0			
22UW-0111	Norbuprenorphine	1	0	0	1		
22UW-0112	Fentanyl	0	1	0		1	

22UW-0113		0	0	0			
22UW-0114		0	0	0			
22UW-0115	Fentanyl, Morphine	1	1	0	1	1	
22UW-0116		0	0	0			
22UW-0117	Naloxone	1	0	0	1		
22UW-0118	Fentanyl, Norfentanyl, Metonitazene	0	1	1		2	1
22UW-0119		0	0	0			
22UW-0120	Desproprionylfentanyl	0	0	0			
22UW-0121	Fentanyl	0	1	0		1	
22UW-0122		0	0	0			
22UW-0123		0	0	0			
22UW-0124		0	0	0			
22UW-0125		0	0	0			
22UW-0126	Fentanyl, Norfentanyl, Metonitazene	0	1	1		2	1
22UW-0127	Hydrocodone, Morphine	1	0	0	2		
22UW-0128		0	0	0			
22UW-0129	beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		3	
22UW-0130		0	0	0			
22UW-0131	3-Fluorofentanyl, 4-Fluorofentanyl, Fentanyl	0	1	0		3	
22UW-0132		0	0	0			
22UW-0133		0	0	0			
22UW-0134		0	0	0			
22UW-0135		0	0	0			
22UW-0136	Fentanyl, Metonitazene	0	1	1		1	1
22UW-0137	Fentanyl	0	1	0		1	
22UW-0138		0	0	0			
22UW-0139	Fentanyl	0	0	0		1	
22UW-0140		0	0	0			
22UW-0141	Dextromethorphan	1	0	0	1		
22UW-0142		0	0	0			
22UW-0143	Fentanyl	0	1	0		1	
22UW-0144		0	0	0			
22UW-0145		0	0	0			
22UW-0146		0	0	0			
22UW-0147		0	0	0			
22UW-0148		0	0	0			
22UW-0149		0	0	0			
22UW-0150	Acetylfentanyl, Fentanyl, Morphine, Naloxone	1	1	0	2	2	
22UW-0151		0	0	0			

22UW-0152		0	0	0			
22UW-0153		0	0	0			
22UW-0154		0	0	0			
22UW-0155		0	0	0			
22UW-0156		0	0	0			
22UW-0157	Naloxone	1	0	0	1		
22UW-0158		0	0	0			
22UW-0159	Naloxone	1	0	0	1		
22UW-0160		0	0	0			
22UW-0161		0	0	0			
22UW-0162	Dextromethorphan	1	0	0	1		
22UW-0163	Dextromethorphan	1	0	0	1		
22UW-0164		0	0	0			
22UW-0165		0	0	0			
22UW-0166		0	0	0			
22UW-0167		0	0	0			
22UW-0168		0	0	0			
22UW-0169	Fentanyl, Metonitazene	0	1	1		1	1
22UW-0170		0	0	0			
22UW-0171		0	0	0			
22UW-0172	beta-hyroxyfentanyl	0	1	0		1	
22UW-0173		0	0	0			
22UW-0174		0	0	0			
22UW-0175	Metonitazene	0	0	1			1
22UW-0176		0	0	0			
22UW-0177	Metonitazene	0	0	1			1
22UW-0178	4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		4	
22UW-0179		0	0	0			
22UW-0180	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl	0	1	0		3	
22UW-0181	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentany,	0	1	0		3	
22UW-0182		0	0	0			
22UW-0183	beta-hydroxyfentanyl, Fentanyl	0	1	0		2	
22UW-0184		0	0	0			
22UW-0185	beta-hydroxyfentanyl, Fentanyl	0	1	0		2	
22UW-0186		0	0	0			
22UW-0187		0	0	0			
22UW-0188		0	0	0			
22UW-0189		0	0	0			
22UW-0190		0	0	0			

ID	Substance						
22UW-0191		0	0	0			
22UW-0192		0	0	0			
22UW-0193		0	0	0			
22UW-0194		0	0	0			
22UW-0195		0	0	0			
22UW-0196		0	0	0			
22UW-0197		0	0	0			
22UW-0198		0	0	0			
22UW-0199		0	0	0			
22UW-0200		0	0	0			
22UW-0201		0	0	0			
22UW-0202		0	0	0			
22UW-0203	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0204		0	0	0			
22UW-0205		0	0	0			
22UW-0206	beta-hydroxyfentanyl, Fentanyl	0	1	0		2	
22UW-0207		0	0	0			
22UW-0208		0	0	0			
22UW-0209		0	0	0			
22UW-0210		0	0	0			
22UW-0211		0	0	0			
22UW-0212		0	0	0			
22UW-0213		0	0	0			
22UW-0214		0	0	0			
22UW-0215		0	0	0			
22UW-0216		0	0	0			
22UW-0217		0	0	0			
22UW-0218	Norfentanyl, Methoxyacetylfentanyl	0	1	0		2	
22UW-0219		0	0	0			
22UW-0220		0	0	0			
22UW-0221		0	0	0			
22UW-0222		0	0	0			
22UW-0223		0	0	0			
22UW-0224		0	0	0			
22UW-0225	Fentanyl, Naloxone	1	1	0	1	1	
22UW-0226		0	0	0			
22UW-0227		0	0	0			
22UW-0228	Fentanyl, Metonitazene, Naloxone	1	1	1	1	1	1
22UW-0229	Fentanyl, Metonitazene	0	1	1		1	1
22UW-0230	Naloxone	1	0	0	1		

ID	Substance						
22UW-0231		0	0	0			
22UW-0232		0	0	0			
22UW-0233	Fentanyl	0	1	0		1	
22UW-0234		0	0	0			
22UW-0235		0	0	0			
22UW-0236		0	0	0			
22UW-0237	Hydromorphone	1	0	0	1		
22UW-0238		0	0	0			
22UW-0239		0	0	0			
22UW-0240		0	0	0			
22UW-0241		0	0	0			
22UW-0242		0	0	0			
22UW-0243		0	0	0			
22UW-0244		0	0	0			
22UW-0245		0	0	0			
22UW-0246		0	0	0			
22UW-0247		0	0	0			
22UW-0248		0	0	0			
22UW-0249		0	0	0			
22UW-0250		0	0	0			
22UW-0251		0	0	0			
22UW-0252	Dextromethorphan	1	0	0	1		
22UW-0253		0	0	0			
22UW-0254	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0255		0	0	0			
22UW-0256		0	0	0			
22UW-0257		0	0	0			
22UW-0258		0	0	0			
22UW-0259		0	0	0			
22UW-0260	Hydrocodone, Naloxone, Oxycodone	1	0	0	3		
22UW-0261		0	0	0			
22UW-0262	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl	0	1	0		5	
22UW-0263	Fentanyl, Norfentanyl	0	1	0		2	
22UW-0264		0	0	0			
22UW-0265		0	0	0			
22UW-0266	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		6	

ID	Substances						
22UW-0267	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		6	
22UW-0268	beta-hydroxyfentanyl, Fentanyl, Norfentanyl, Oxycodone	1	1	0	1	3	
22UW-0269		0	0	0			
22UW-0270		0	0	0			
22UW-0271		0	0	0			
22UW-0272		0	0	0			
22UW-0273		0	0	0			
22UW-0274		0	0	0			
22UW-0275		0	0	0			
22UW-0276		0	0	0			
22UW-0277		0	0	0			
22UW-0278		0	0	0			
22UW-0279		0	0	0			
22UW-0280	Hydrocodone	1	0	0	1		
22UW-0281		0	0	0			
22UW-0282		0	0	0			
22UW-0283		0	0	0			
22UW-0284		0	0	0			
22UW-0285		0	0	0			
22UW-0286		0	0	0			
22UW-0287		0	0	0			
22UW-0288	2-Fluorofentanyl, 3-Fluorofentanyl, 4-Fluorofentanyl, beta-hydroxyfentanyl, Fentanyl, Norfentanyl, Isonitazene, Metonitazene	0	1	1		6	2
22UW-0289		0	0	0			
22UW-0290		0	0	0			
22UW-0291		0	0	0			
22UW-0292	Dextromethorphan	1	0	0	1		
22UW-0293		0	0	0			
22UW-0294		0	0	0			
22UW-0295		0	0	0			
22UW-0296		0	0	0			
22UW-0297		0	0	0			
22UW-0298		0	0	0			
22UW-0299		0	0	0			
22UW-0300		0	0	0			
22UW-0301		0	0	0			

22UW-0302		0	0	0			
22UW-0303		0	0	0			
22UW-0304	Fentanyl	0	1	0	1		
22UW-0305		0	0	0			
22UW-0306	beta-hydroxyfentanyl, Fentanyl, Norfentanyl	0	1	0		3	
22UW-0307		0	0	0			
22UW-0308		0	0	0			
22UW-0309		0	0	0			
22UW-0310		0	0	0			
22UW-0311		0	0	0			
22UW-0312		0	0	0			
22UW-0313		0	0	0			
22UW-0314		0	0	0			
22UW-0315		0	0	0			
22UW-0316		0	0	0			
22UW-0317		0	0	0			
22UW-0318	Fentanyl	0	1	0		1	
22UW-0319	Dextromethorphan	1	0	0	1		
22UW-0320		0	0	0			
22UW-0321		0	0	0			
22UW-0322		0	0	0			
22UW-0323		0	0	0			
22UW-0324		0	0	0			
22UW-0325		0	0	0			
22UW-0326		0	0	0			
22UW-0327		0	0	0			
22UW-0328		0	0	0			
22UW-0329		0	0	0			
22UW-0330		0	0	0			
22UW-0331	Fentanyl	0	1	0		1	
22UW-0332		0	0	0			
	Total	33	58	10	40	120	12

Table A.14. Fentanyl classification model feature identifications

Compound Relation/ID	Possible Identity	Modified Cosine Score	Number of Matching Signals	Feature ID	Number of Identifications	Adduct
Norfentanyl	Norfentanyl	0.91	4	3160	51	M+H
Acrylfentanyl + 18.0097	Hydroxy-fentanyl	0.61	3	4563	23	M+H
Fentanyl + 15.9958	Hydroxy-fentanyl	0.84	8	4903	40	M+Na
Fentanyl + 46.0059	methoxy-hydroxy-fentanyl	0.98	3	5055	10	M+H
Fentanyl + 15.9961	Hydroxy-fentanyl	0.96	6	5505	49	
N/A	N/A	N/A	N/A	3226	N/A	N/A
Norfentanyl + 72.0201	Unknown	0.6	3	6225	17	M+H
Fentanyl	Fentanyl	0.97	10	6322	72	M+H
Dimethoxy-fentanyl - 44.0263	Hydroxy-fentanyl	0.7	3	7485	7	M+H
N/A	N/A	N/A	N/A	7712	N/A	N/A
Remifentanil Acid + 104.0380	Unknown	0.54	4	8978	2	M+H
Remifentanil Acid + 118.0536	Unknown	0.54	4	9897	2	M+H
N/A	N/A	N/A	N/A	11495	N/A	N/A
N/A	N/A	N/A	N/A	12510	N/A	N/A

Table A.15. Morphinan classification model feature identifications.

Compound Relation/ID	Possible Identity	Modified Cosine Score	Number of Matching Signals	Feature ID	Number of Identifications	Adduct
Morphine/Norcodeine	Morphine/ Norcodeine	0.68	7	1093	65	M+H
Dextrorphan + 176.03104	Dextrorphan Glucuronide	0.72	3	1239	12	M+H
Naloxone + 2.0158	N-propylnoroxymorphone	0.91	5	1347	13	M+H
Naloxone	Naloxone	0.98	14	1412	41	M+H
Naloxone - 18.01056	In-source fragment - Loss of water	0.65	24	1420	11	M+H
Oxycodone - 2.0158	6-Oxycodol	0.9	3	1506	9	M+H
3-hydroxy-N-methylmorphinan + 65.94898	Unknown	0.91	6	1766	6	M+K
Hydrocodone/Codeine	Hydrocodone/ Codeine	0.97	6	2155	14	M+H
Dextrorphan + 79.9559	Dextrorphan Sulphate	0.83	7	2527	11	M+H
3-Hydroxymorphinan	3-Hydroxymorphinan	0.9	18	3247	11	M+H
Dextrorphan/Levorphanol	Dextrorphan/ Levorphanol	0.98	6	3329	13	M+H
Norbuprenorphine + 238.1406	False Positive / Contaminant	0.73	4	4212	287	M+H
Norbuprenorphine + 252.1203	False Positive / Contaminant	0.93	4	4262	254	M+H
Codeine + 3.9733	Possible Fluoromorphine, MS1 match of 6ppm	0.72	7	4358	5	M+Na
Codeine + 26.0524	Unknown	0.78	3	5884	3	M+H
Dextromethorphan	Dextromethorphan	1	12	6244	17	M+H
Nalbuphine + 160.0847	Unknown	0.84	8	6698	7	M+H
Nalbuphine - 102.0480	Unknown	0.9	3	6871	5	M+Na
Nalbuphine + 112.06619	Possible buprenorphine impurity or lipid	0.72	14	7731	1	M+H

	analogue of nalbuphine					
Nalbuphine + 130.07283	Unknown	0.74	10	7732	3	M+H
Nalbuphine + 2.0224	Unknown	0.78	7	8478	2	M+Na
Naloxone + 154.1352	Unknown	0.66	18	8711	30	M+H
Norcodeine + 192.1181	Unknown	0.88	3	8862	3	M+H
Butorphanol + 0.0023	False Positive / Contaminant	0.93	3	12401	331	M+H
Oxycodone - 2.0540	False Posiutive / Contaminant	0.99	3	12887	245	M+H
Naltrexone + 58.04415	False Positive / Contaminant	0.86	3	13046	324	M+H
Dihydrocodeine + 44.0636	False Positive / Contaminant	0.67	13	13195	258	M+H
Nalbuphine - 29.9728	False Positive / Contaminant	0.9	6	13492	249	M+H
Norbuprenorphine - 104.0468	False Positive / Contaminant	0.98	3	13502	297	M+H

Appendix B

Supplementary Figures for Chapter 4

Figure B.1. Stacked bar graph demonstrating strict fentanyl compound identifications by the fentanyl ML classifier.

Figure B.2. Stacked bar graph demonstrating relaxed fentanyl compound identifications by the fentanyl ML classifier.

Figure B.3. Stacked bar graph demonstrating strict morphinan compound identifications by the morphinan ML classifier.
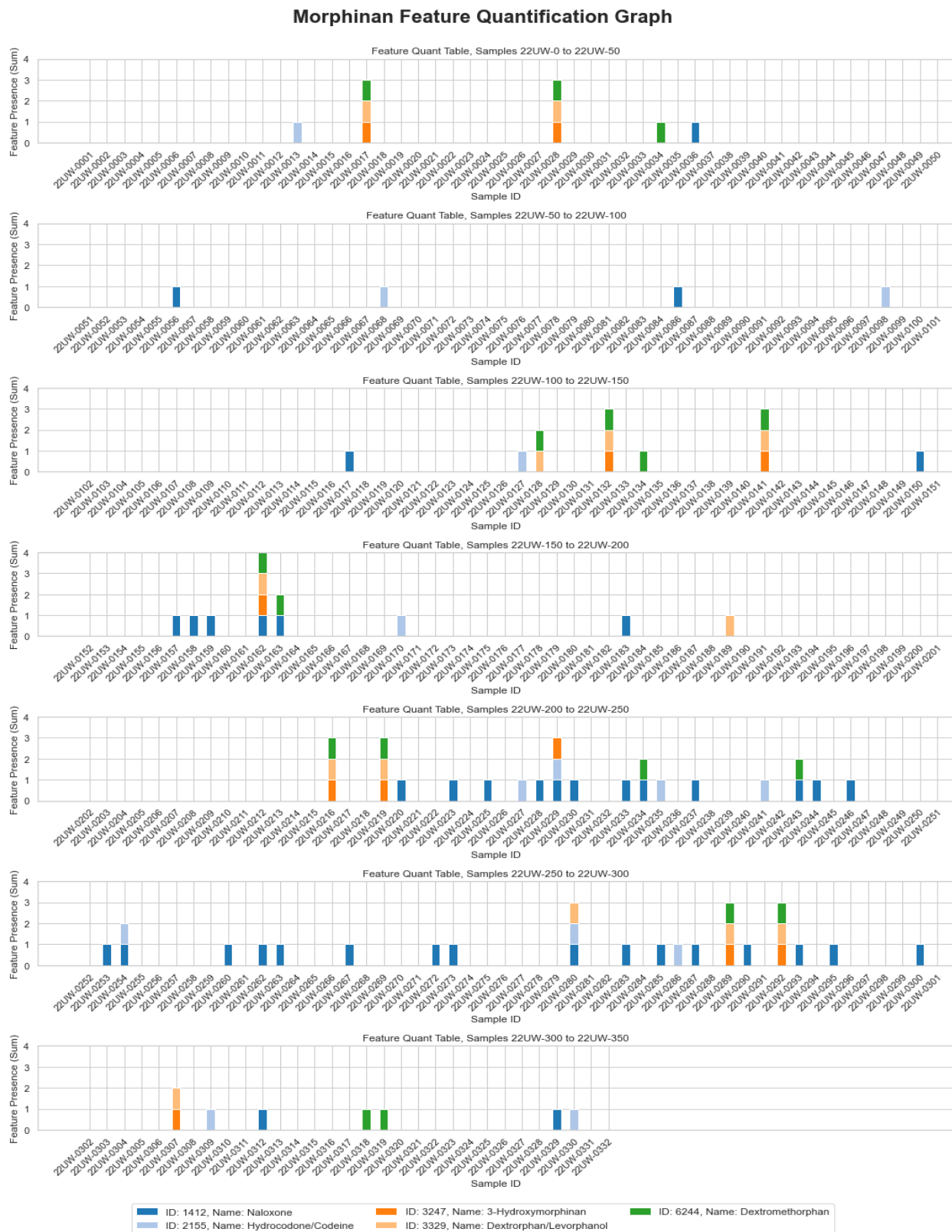
Figure B.3. Stacked bar graph demonstrating relaxed morphinan compound identifications by the morphinan ML classifier.
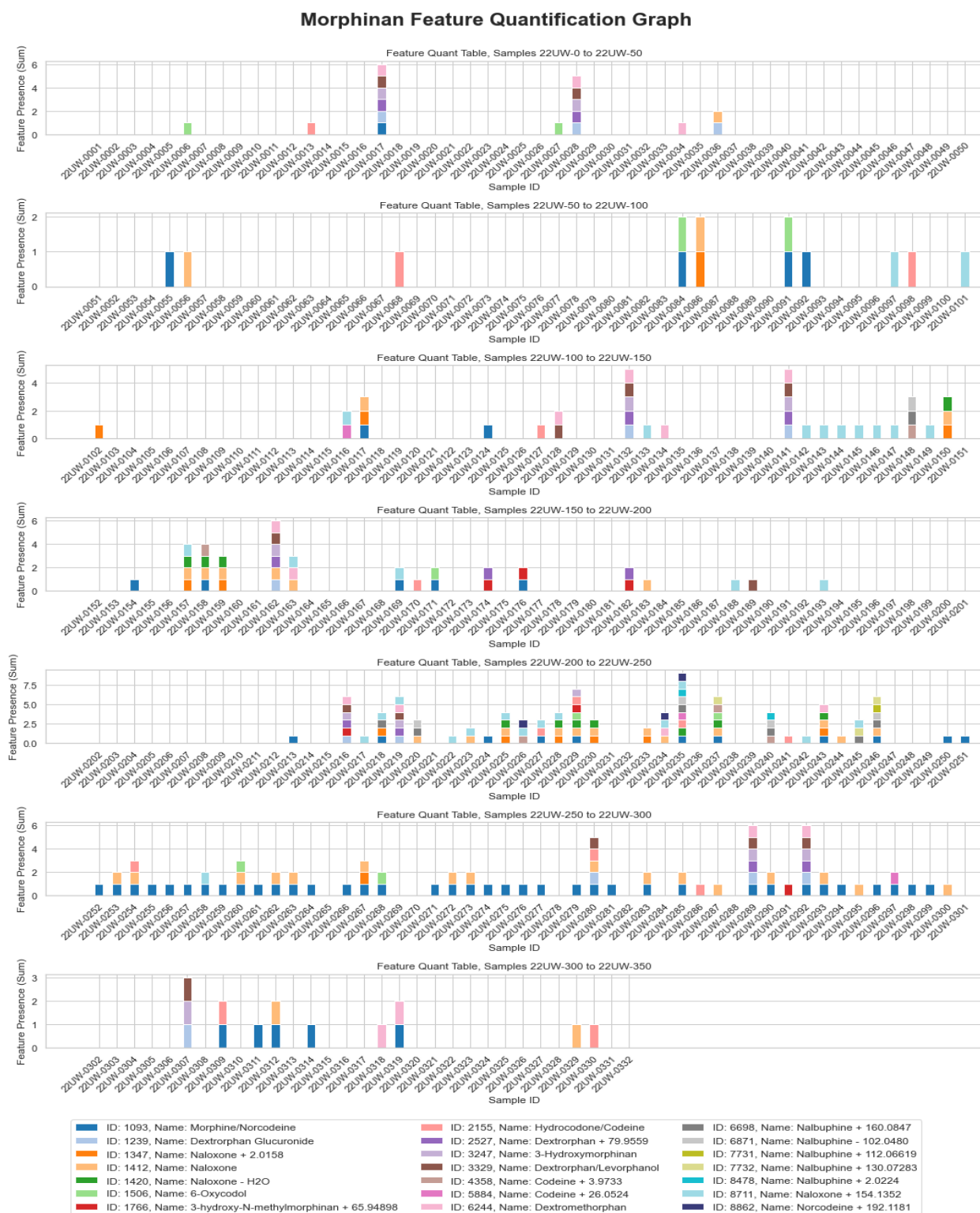

Morphinan Feature Quantification Graph

Figure B.4. Stacked bar graph demonstrating compound identification of fentanyl compounds by spectral library searching.

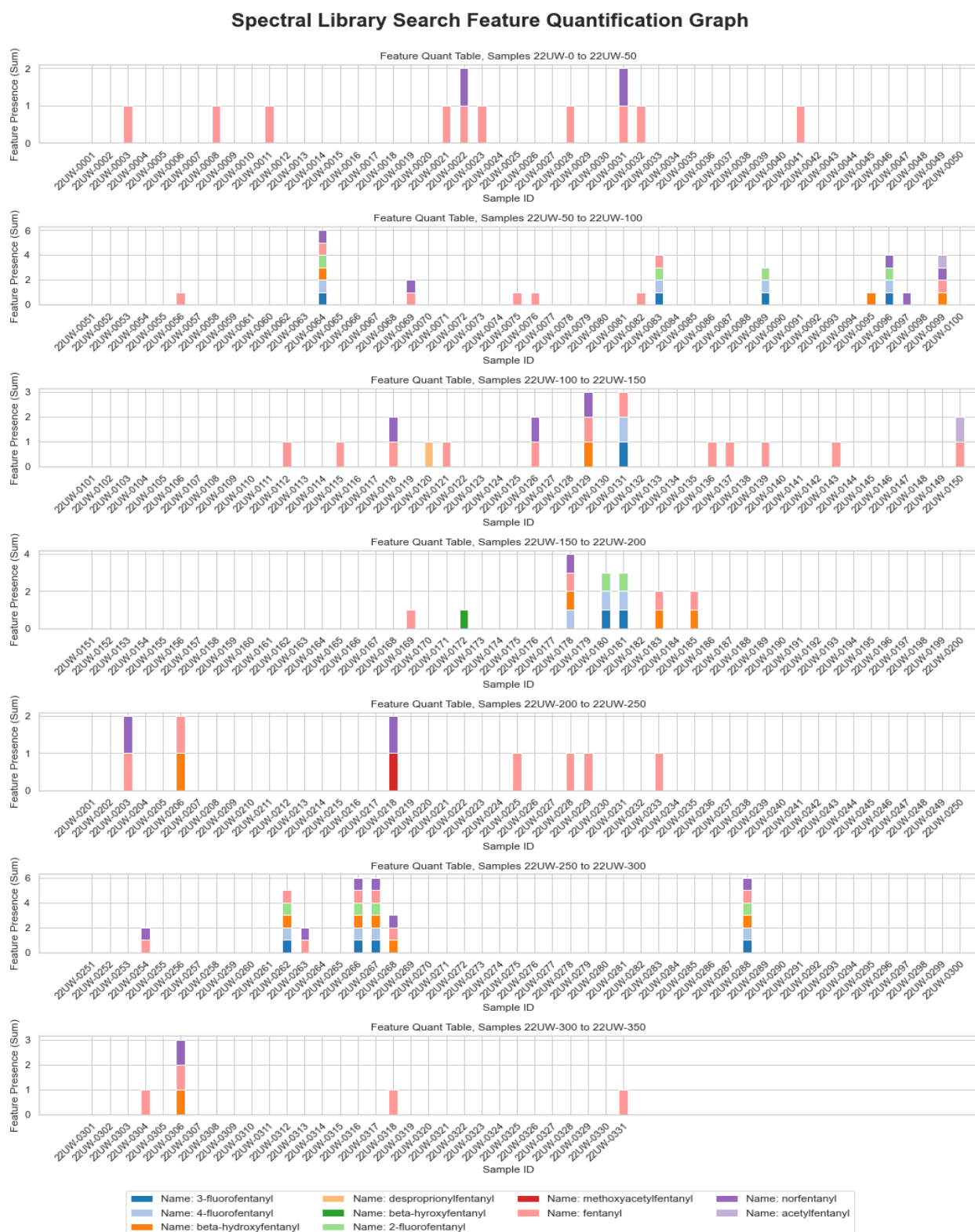Spectral Library Search Feature Quantification Graph

Figure B.5. Stacked bar graph demonstrating compound identification of morphinan compounds by spectral library searching.
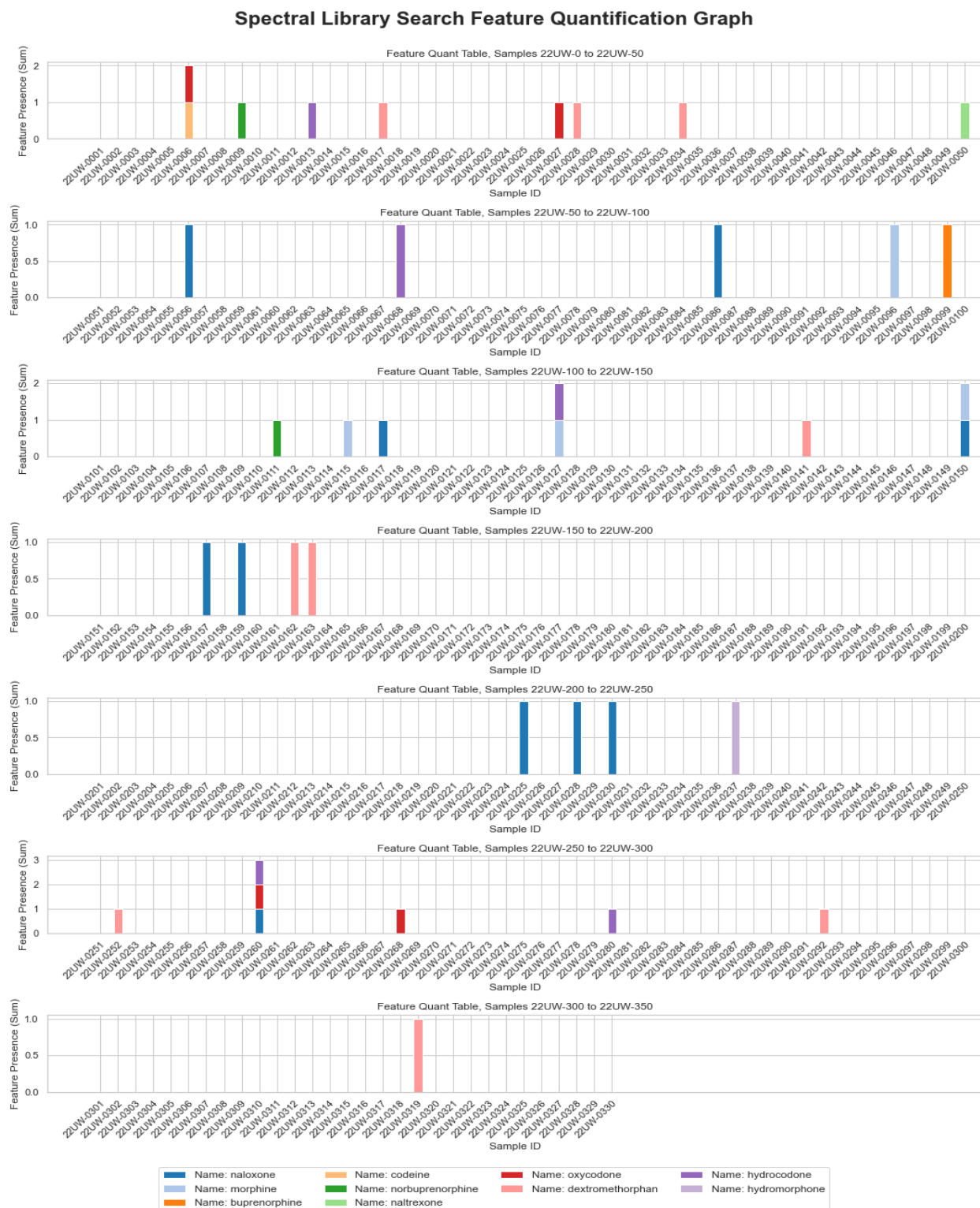
Figure B.6. Stacked bar graph demonstrating compound identification of nitazene compounds by spectral library searching.
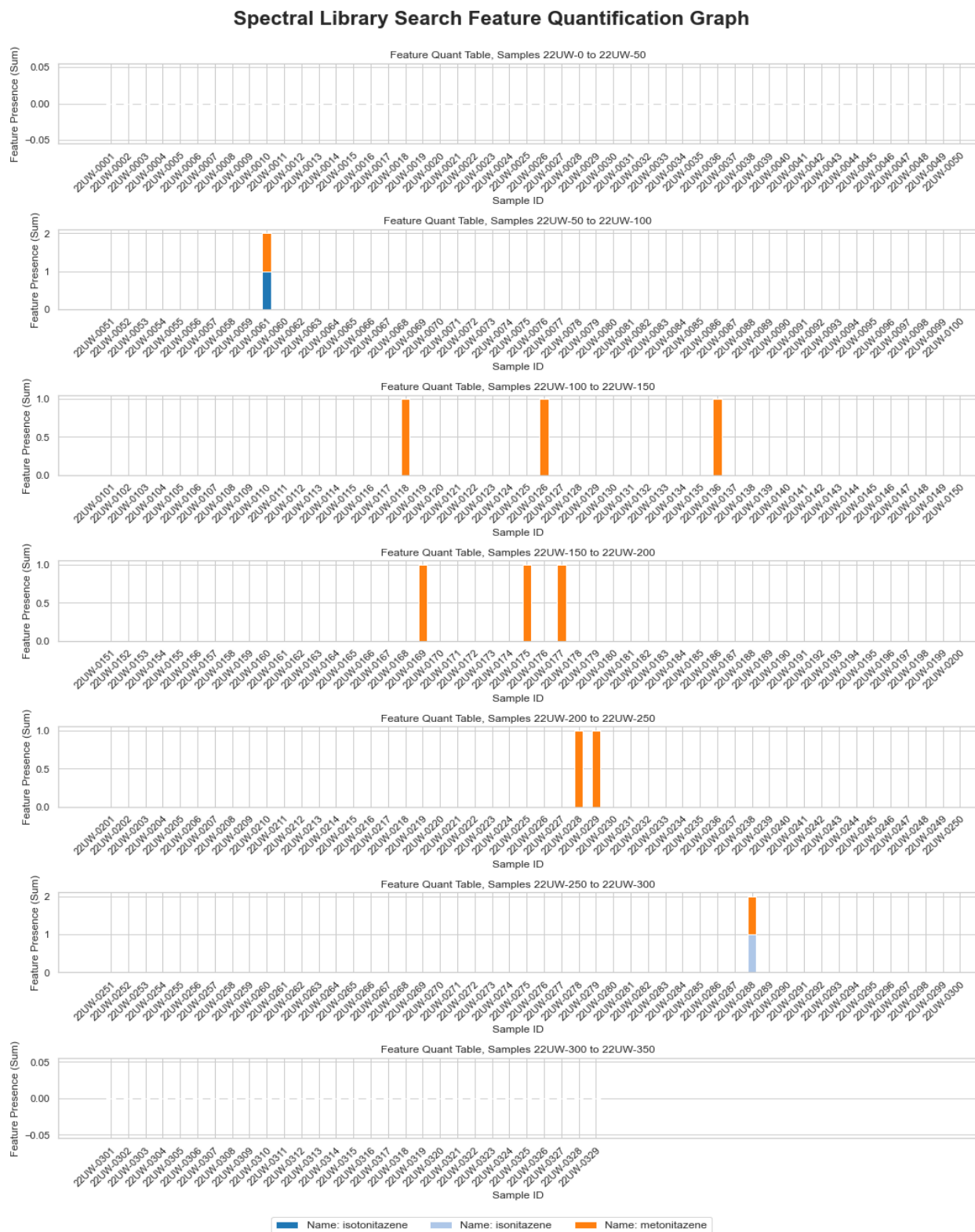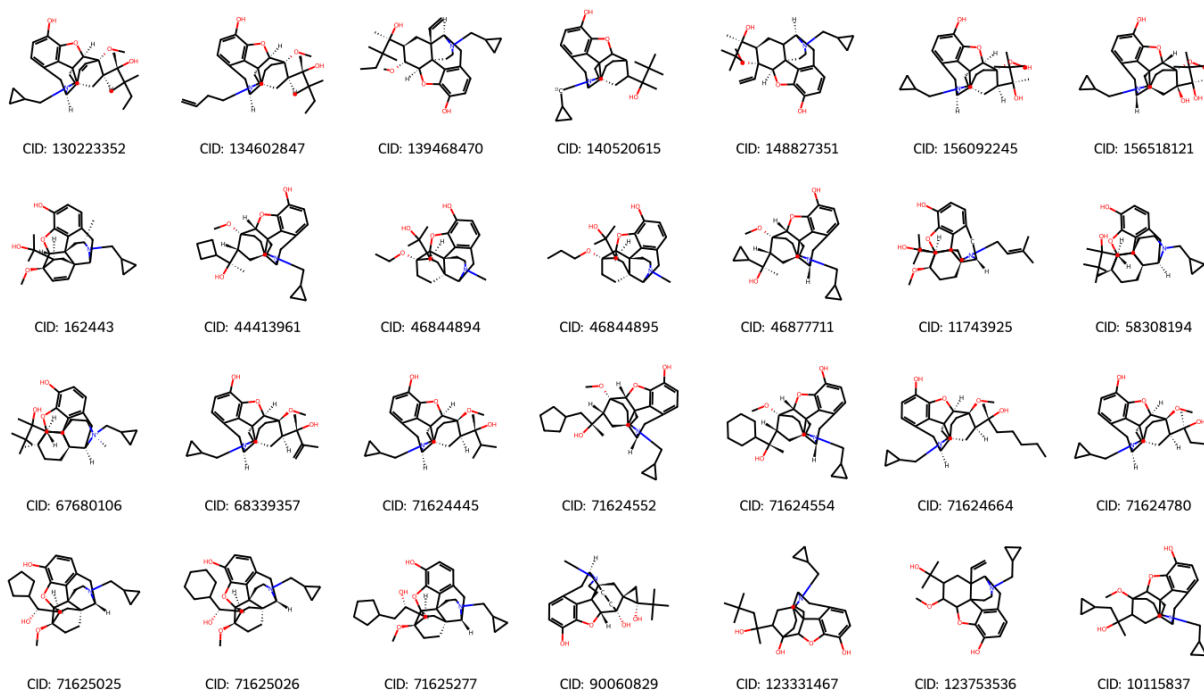
Figure B.7 Morphinan opioid group depicting prototypical morphinan opioid features and buprenorphine like opioids in t-SNE clusters.

Buprenorphine-Like Compounds:

| CID: 130223352 | CID: 134602847 | CID: 139468470 | CID: 140520615 | CID: 148827351 | CID: 156092245 | CID: 156518121 |

| CID: 162443 | CID: 44413961 | CID: 46844894 | CID: 46844895 | CID: 46877711 | CID: 11743925 | CID: 58308194 |

| CID: 67680106 | CID: 68339357 | CID: 71624445 | CID: 71624552 | CID: 71624554 | CID: 71624664 | CID: 71624780 |

| CID: 71625025 | CID: 71625026 | CID: 71625277 | CID: 90060829 | CID: 123331467 | CID: 123753536 | CID: 10115837 |

Morphine-Like Compounds:

| CID: 129684446 | CID: 13065599 | CID: 134186042 | CID: 138720655 | CID: 140976824 | CID: 10215960 | CID: 141341942 |

| CID: 151056914 | CID: 154290160 | CID: 161460093 | CID: 161636666 | CID: 21279955 | CID: 22870688 | CID: 10928736 |

| CID: 44282413 | CID: 44293036 | CID: 44293051 | CID: 44293481 | CID: 44293935 | CID: 46878175 | CID: 46878185 |