Computational and machine learning approaches for functional characterization of chemical compounds using chemical-genetic interactions

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Hamid Safizadeh

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Chad L. Myers, Ph.D.

May 2023

© Hamid Safizadeh 2023
ALL RIGHTS RESERVED

Acknowledgements

The research work in this document is the product of a close collaboration among the researchers at the University of Minnesota, University of Toronto, and RIKEN Center for Sustainable Resource Science in Japan. These researchers are listed by their institution as follows:

- University of Minnesota: Chad Myers, Scott W. Simpkins, and Justin Nelson
- University of Toronto: Charles Boone, Sheena C. Li, Luis A. Vega Isuhuaylas, and Zhijian Li
- RIKEN Institute: Minoru Yoshida, Jeff S. Piotrowski, Hiroyuki Osada, Yoko Yashiroda, Hiroyuki Hirano, and Mami Yoshimura

The details of their contributions are listed on the first pages of chapters 2–4.

During my doctoral program, I experienced difficulties including financial issues. I would like to acknowledge the office of International Student and Scholar Services (ISSS) that guided me through available financial resources for international students. Without the ISSS financial support, the completion of my doctoral program was surely impossible. In specific, I would like to thank Duane Rohovit and Drew Smith, the former ISSS advisors who always helped me navigate a solution to financial and academic problems.

Moreover, I would like to acknowledge and thank those who stayed with me during the dark time of my health crisis. First and foremost, I would like to thank Azadeh Etemadi, Christine Johns, Payman Najafi, and Shirin Asadi, who greatly supported me during this difficult time. I would also like to thank Alireza Nabavizadeh, Borzoo Baradaran, Paul and Adrienne Volk, Mohammad Amin Tadayon, Kiarash Yoosefi, Apollo Arman, Shahrbanou

Amirkeivan, Naty Lopez, Michael Jacobson, Spencer and Christine Chow, Cathy Moore, Wen Wang, Mahfuz Rahman, Farhad Shahriari Nogorani, Timothy McDonald, Scott and Kassandra Shimotsu, Mohsen Mahmoodi, Navid Noori, Julianne Pinke Vanderaa, Robert and Susan Osburn, Farhad Kosari, Alireza Foroozan, Morteza Salehi, and many others who provided much encouragement and comfort during the time of my health crisis. Without their support, the completion of this doctoral dissertation would have been very difficult, if not impossible. In addition, I would like to thank my siblings who always supported me during my doctoral program.

Finally, I would like to thank Professor Chad Myers, my advisor, who took the risk to accept me with limited background in biology and genetics as a doctoral student in his research group. Chad provided me with a unique opportunity to work in a supportive and collaborative research environment, where I received rigorous training in computational and systems biology. During my doctoral program, I encountered unexpected difficulties. I would like to sincerely thank Chad for his patience and support as I navigated solutions to those challenges. I would also like to thank Professor Mostafa (Mos) Kaveh, Professor Daniel Boley, and Professor Marc Riedel, who agreed to serve on my doctoral examination committee.

Dedication

To the loving memory of my parents, especially my amazing mother.

To the memory of Manouchehr Harandi, who greatly assisted my academic pursuits.

To Azadeh Etemadi for her endless support throughout the most difficult times of my life in the United States.

Abstract

Chemical-genetic interactions provide an invaluable source of information about the cellular functions perturbed by compounds. The chemical-genetic interaction profile of a compound against genome-wide mutant collections captures the effects of that compound on the fitness of each mutant in the collection. Since such profiles cover the entire genome, they can be used as a systematic, unbiased functional proxy for the effects of compounds on cells. Therefore, exploring the chemical-genetic interaction profiles of compounds can elucidate the modes of actions of compounds and their genetic targets.

Using large libraries of chemical-genetic interaction profiles in *Saccharomyces cerevisiae*, we improved the functional prediction of compounds from chemical structures. We benchmarked a set of molecular fingerprints and similarity coefficients to find the pair with superior prediction power for connecting the compound structural space to functional space. We also developed a machine learning approach for predicting compound functional similarities from molecular fingerprints, which demonstrated substantial improvement over existing similarity measurements based on chemical structure.

Moreover, using large libraries of multimodal chemical-genetic interaction profiles in *S. cerevisiae*, we developed two scalable computational pipelines for target identification of compounds. In one pipeline, we integrated chemical-genetic interactions from several modes to assign a unified target score to each compound–gene pair and prioritize the most promising candidates for further experimental studies. In another pipeline, we modeled the gene target of a compound as the rumor source spreading inhibitory signals across a genetic network. Our validation results confirmed novel targets for several studied compounds.

Table of Contents

Acknowledgements ..... **i**

Dedication ..... **iii**

Abstract ..... **iv**

List of Tables ..... **viii**

List of Figures ..... **x**

1 Introduction **1**

1.1 High-throughput screening in drug discovery 5

1.1.1 Generating genetic interactions as phenotypic profiles 5

1.1.2 Generating and interpreting chemical-genetic interaction profiles 7

1.2 Computational approaches in drug discovery 9

1.2.1 Prioritizing compounds using virtual screening 9

1.2.2 Accelerating discoveries using artificial intelligence 10

1.2.3 Identifying gene targets of chemical compounds 12

1.3 Outline 13

2 Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions **15**

2.1 Abstract 16

2.2 Introduction 17

2.3 Results and discussion 20

2.3.1 Establishing a systematic benchmark for structural similarity measures 20

2.3.2 Evaluating the performance of structural similarity measures 25

2.3.3 Optimizing the depth of molecular fingerprints 33

2.3.4 Improving the prediction performance using SVM models 34

2.3.5 Exploring the basis of the predictive power of SVM models 38
2.3.6 Example use case: retrieving novel functionally analogous compounds for virtual screening applications 41
2.3.7 Evaluating the reciprocal predictive power of structural and functional similarities 44
2.4 Conclusion 45
2.5 Materials and methods 47
2.5.1 Data collections 47
2.5.2 Structural description of compounds 48
2.5.3 Establishing the gold standard for biological activity 49
2.5.4 Designing the support vector machine learning pipeline 50
2.5.5 Estimating the diversity of compound collections 51
3 Target Identification of Bioactive Compounds Using a High-Resolution, Integrated Chemical-Genetic Approach 53
3.1 Abstract 55
3.2 Introduction 56
3.3 Target identification pipeline 59
3.3.1 Assembly of whole genome, temperature sensitive, heterozygous diploid, and MoBY overexpression profiles for compounds from diverse libraries 59
3.3.2 Deprioritizing noisy data screens 61
3.3.3 Outlier analysis for prioritizing gene targets 64
3.3.4 Detecting genes exhibiting frequent drop-outs across many compounds 65
3.3.5 Specificity analysis in gene target predictions 66
3.3.6 Target scores for identifying specific, strong gene targets 68
3.3.7 Driver scores of bioprocess target predictions for target identification 75
3.3.8 Bioprocess target prediction agreements across different mutant pools 76
3.3.9 Structural similarity measurements 78

3.4 Results and Discussion 78

3.4.1 Novel mutant reagent sets for high resolution target identification in *Saccharomyces cerevisiae* 78

3.4.2 Examples of compound–gene target validations 80

3.5 Conclusion 88

4 Target Identification of Bioactive Compounds Using a Network Diffusion Approach 90

4.1 Abstract 91

4.2 Introduction 92

4.3 Results and Discussion 95

4.3.1 Modeling the rumor graph for a compound 95

4.3.2 Evaluating the PINT framework against random rumor graphs 97

4.3.3 Exploring the correlation of rumor centrality and CG-TARGET driver scores 100

4.3.4 Evaluating gene targets using heterozygous diploid mutants 101

4.3.5 Example compound case: Benomyl 105

4.4 Conclusion 107

5 Conclusion and Future Work 108

References 112

Appendix A: Supporting Information for Chapter 2 140

Appendix B: Supporting Information for Chapter 3 153

Appendix C: Supporting Information for Chapter 4 155

List of Tables

Table 2.1. Molecular fingerprints 22

Table 2.2. Similarity coefficients 23

Table 3.1. Data generation using several diverse compound libraries 61

Table A.1. Precision at several recall thresholds for all the combinations of a molecular fingerprint and a similarity coefficient using the RIKEN high-confidence set 152

Table A.2. Precision at several recall thresholds for all the combinations of a molecular fingerprint and a similarity coefficient using the NCI/NIH/GSK high-confidence set 152

Table A.3. Functional clustering using the K-means algorithm and functional enrichment analysis using the hypergeometric test for the RIKEN high-confidence set 152

Table A.4. Functional clustering using the K-means algorithm and functional enrichment analysis using the hypergeometric test for the NCI/NIH/GSK high-confidence set 152

Table A.5. Structural clustering of our RIKEN and NCI/NIH/GSK high-confidence sets using the K-medoids algorithm 152

Table A.6. Ranked lists of compound pairs selected from the top 1000 machine learning-derived true positive pairs for our RIKEN and NCI/NIH/GSK high-confidence sets 152

Table A.7. Comprehensive ranked lists of structural and chemical-genetic similarities for our RIKEN and NCI/NIH/GSK high-confidence sets 152

Table B.1. Ranked lists of frequent flier genes and frequent flier compounds using the HET data for our RIKEN high-confidence set 154

Table B.2. Target score examples 154

List of Figures

Figure 2.1. Ligand-based virtual screening of a target compound 18

Figure 2.2. Performance of selected prediction models using our RIKEN high-confidence set 29

Figure 2.3. Performance of selected prediction models using our NCI/NIH/GSK high-confidence set 31

Figure 2.4. Impact of the describing depth of molecular fingerprints on the RIKEN high-confidence set 34

Figure 2.5. Prediction performance of machine learning models 36

Figure 2.6. Functional and structural clusters of top true positive pairs for our RIKEN high-confidence set 39

Figure 2.7. Functional and structural clusters of top true positive pairs for our NCI/NIH/GSK high-confidence set 40

Figure 2.8. Reciprocal evaluation of prediction performance of structural vs. functional similarity and machine-learning-based virtual screening of a target 43

Figure 3.1. Removing noisy screens and demonstrating outlier analysis concepts 62

Figure 3.2. Frequent flier analysis 67

Figure 3.3. Integration of information from three distinct chemical genomic approaches for target identification of a bioactive compound 70

Figure 3.4. Comparison of chemical-genetic interaction profiles with genetic interaction profiles enabled compound–target identification 71

Figure 3.5. Pipeline for bioprocess target prediction agreements across data pools 77

Figure 3.6. Priority gene targets for NPE71 from the RIKEN bioactive compounds 82
Figure 3.7. Priority gene targets for RP-3-161 from the University of British Columbia natural products library 84
Figure 3.8. Priority gene targets for NPD6433 from the RIKEN bioactive set and NPD5728 from the RIKEN high-confidence set 87
Figure 4.1. Modeling gene targets of compounds as rumor sources using a global genetic similarity network 94
Figure 4.2. Evaluating the PINT framework against random rumor graphs 98
Figure 4.3. Correlation analysis of rumor centrality scores with CG-TARGET driver scores 101
Figure 4.4. Evaluation of PINT gene targets using HET data 103
Figure 4.5. Target identification for Benomyl using PINT 106
A.1. SRD analysis for molecular fingerprints 142
A.2. SRD analysis for similarity coefficients 144
A.3. Impact of the describing depth of molecular fingerprints on the NCI/NIH/GSK high-confidence set 146
A.4. Pairwise structural vectors and bootstrapping used by our machine learning pipeline 147
A.5. Correlation analysis of the predicted structural similarities with the chemical-genetic similarities 148
A.6. Spearman's rank correlation distribution between the machine-learning-derived and the ASP/Braun-Blanquet-derived structural similarity predictions 149

A.7. Distribution of the predicted biological functions across 17 broad, previously defined functional neighborhoods 150
A.8. Prediction performance of the machine learning models for the 5% cutoff (as a more stringent cutoff than 10%) on the functional similarity gold standard 151
C.1. Concept of the PINT framework 156
C.2. Mapping approach for connecting rumor centralities with HET data 157

1 Introduction

Biopharmaceutical drug discovery activities are time-consuming and expensive procedures. In 2003, the total development cost per drug was estimated at $802 million (1). In 2006, the average capitalized development cost per new drug was estimated at $1 billion using the publicly accessible database Pharmaprojects (2). This cost has continued to climb to the total amount of nearly $2.8 billion for a biopharmaceutical (3,4), implying that the cost of drug discovery and developments in biopharmaceutical industry has been rising dramatically. The increasing cost of drug discovery procedures for new biopharmaceuticals was originated from high costs for development phases (i.e., pre-clinical phases and clinical trial phases I to III), low success rates, long development times, and high capital costs (5). A recent study conducted at Massachusetts Institute of Technology found that only 13.8% of all drug development programs lead to FDA approval (6). This study emphasized that the development of the drugs that treated rare diseases, also known as orphan drugs, had significantly lower success rates, with only 6.2% of drug development projects reaching the market. The FDA approval rate was even much lower for oncology drugs, where the probability of success of drug development projects was as low as 3.4%. Even when a drug finally receives FDA approval, the drug can sometimes be removed from the market due to severe potential side effects that only appear after the drug is used by many patients (7–9).

An FDA-approved biopharmaceutical may initially be effective against its target(s) but eventually become ineffective due to various drug resistance mechanisms of action that a pathogen or tumor cell develops to defeat the drug. Such resistance is more prevalent for

antibiotic (10–15), antifungal (16–20), antiviral (21–25), and anticancer agents (26–28). The ongoing emergence of antimicrobial-resistant pathogens that do not respond to existing antibiotics is considered a serious global health risk and is prioritized as one of the top ten global public health threats by World Health Organization (29). Especially alarming is the rapid global spread of multi- and pan-resistant bacteria, also known as "superbugs", that cause infections not treatable with existing antimicrobial medicines such as antibiotics. In addition, many fungal infections have existing treatability issues such as toxicity especially for patients with other underlying infections (e.g., Human Immunodeficiency Virus or HIV infection). Antiviral drug resistance is also an increasing concern in immunocompromised patient populations, where ongoing viral replication and prolonged drug exposure lead to the selection of resistant strains (29). Resistance has already developed to most antivirals including antiretroviral drugs. All antiretroviral drugs, including newer classes, are at risk of becoming partially or fully inactive because of the emergence of drug-resistant HIV (i.e., HIVDR). People receiving antiretroviral therapy can acquire HIVDR, and people can also be infected with HIV that is already drug resistant. For cancer therapeutics, drug resistance usually develops via different mechanisms of action than the ones for antibiotic, antifungal, and antiviral agents. However, promising targeted cancer therapies are also at the high risk of resistance, and drug-resistant cancerous tumors are commonly even more aggressive and more difficult to be treated (30–33).

The difficulties pertaining to the resistance of current biopharmaceuticals and the cost of developing new biopharmaceuticals necessitate novel methods for the discovery of new drug agents. To develop safe and effective drugs that will provide the maximum health

benefits to humans and minimize potential risks and side effects, efficient approaches are required to evaluate the interactions between chemical compounds and biological systems.

The biopharmaceutical industry generally relies on two different strategies for the discovery of new drugs – target-based and phenotypic strategies, which are both reviewed here. Target-based drug development strategies refer to the hypothesis-driven methods with the promise of targeting specific gene products and identifying successful medicines tailored to individual patients. These methods usually need a priori knowledge of the target of a chemical compound. Drug discovery efforts in the past three decades focused on the promise of these strategies and successfully discovered biopharmaceutical therapeutics for specific mutations or cellular defects. Success stories of target-based drug discovery efforts include development of tyrosine kinase inhibitors for cancer treatments (34). For example, Gefitinib, sold under the brand name Iressa, targets EGFR (35), Imatinib, sold under the brand name Glivec/Gleevec, targets BCR-ABL (36), Sorafenib, sold under the brand name Nexavar, targets Raf (37), and Sunitinib, sold under the brand name Sutent, targets VEGFR and PDGFR phosphorylation (38). Target-based drug discovery approaches were also used for the development of many other classes of biopharmaceuticals including antivirals. For example, zanamivir, sold under the brand name Relenza, targets influenza neuraminidase (39,40), maraviroc, sold under the brand name Selzentry, is a CCR5 co-receptor antagonist for the combination antiretroviral treatment of adults infected with only CCR5-tropic HIV (41,42), and raltegravir, sold under the brand name Isentress, is a HIV integrase inhibitor (43–45).

Prior to the introduction of target-based drug development schemes, drug discovery was driven primarily by phenotypic assays, often with limited knowledge of the molecular mechanisms of disease. Phenotypic drug development strategies concentrate on evaluating different compounds against phenotypes in biological systems such as animals, cells, yeast, or bacteria and measuring the phenotypic response of the biological systems in the presence of a compound (46). These strategies do not need a priori knowledge of the potential target of a compound as they perform phenotypic screens of a compound in a whole cell context to observe all the interactions of the compound with the entire biological system. Thorough examination of the function of a compound against a biological system provides detailed insights into the target(s) and side effects of the compound. Although phenotypic screening of compounds is much more expensive than target-based drug discovery approaches, the results of a study showed that the contribution of phenotypic screening to the discovery of small-molecule drugs exceeded that of target-based approaches in the period between 1999 and 2008, an era in which the major focus of drug discovery efforts was on target-based approaches (47). Examples of drug discoveries using phenotypic approaches are linezolid (48), which is an oxazolidinone antibiotic identified in bacterial assays and is sold under the brand name Zyvox, and vorinostat (49), which is the first histone deacetylase inhibitor for cancer treatment and is sold under the brand name Zolinza. The results presented in this document are primarily based on a phenotypic screening strategy. For functional profiling of our chemical compounds using an unbiased phenotypic screening approach, chemical-genetic interaction profiles were generated and interpreted. These profiles were used as a part of the computational frameworks for drug discovery efforts in this dissertation.

1.1 High-throughput screening in drug discovery

1.1.1 Generating genetic interactions as phenotypic profiles

Recent improvements in high-throughput screening and whole genome sequencing technologies have accelerated the generation of quantitative, high-dimensional functional profiles that reflect the effects of cellular perturbations in response to specific stimuli (e.g., gene deletions, drug treatments, environmental stress conditions, etc.). Such genome-wide functional profiles can be treated as phenotypic responses in regard to the utilized stimuli. Early genome-wide phenotypic functional profiles were generated using gene expression or transcription measurements of cellular perturbations (50–54). Because these functional profiles are quantitative and cover the whole genome, a wide array of mathematical and statistical methods can be applied to these profiles. For example, functional profiles of gene expressions may be compared to assess the functional similarity of genes, assuming that different genes are in similar functional states if the gene expression profiles associated to the genes are similar. The functional similarity comparisons of the phenotypic profiles of the genes opens up an avenue for characterizing yet unknown genes using large databases of genes with known functions. For example, in one research study, a reference database or compendium of gene expression profiles corresponding to 300 diverse mutations and chemical treatments in *Saccharomyces cerevisiae* was used to identify the functions of eight previously uncharacterized genes using a pattern matching approach that searched for the genes in the compendium with the highest profile similarity to those eight genes (55).

A genetic interaction occurs when the combination of two genetic mutations leads to an unexpected phenotype (56). Genetic interactions can generate novel phenotypes when genetic mutations are found in combination with one another (57). Large-scale analysis of genome-wide genetic interactions is a powerful approach for interrogating an organism and discovering new genetic dependencies (58). The yeast *S. cerevisiae* is a well-characterized, eukaryotic model organism for which a large number of tools and methods to interrogate cellular functions exist. Construction of genome-wide mutant collections for *S. cerevisiae*, with the aid of high-throughput screening technologies and the emergence of systematic approaches to biology, offers a new direction for functional characterization of genes (59). To quantify genetic interactions, the fitness of a mutant under a second genetic perturbation is evaluated. The fitness of a resulting double-gene knockout mutant can be measured in comparison to the null expectation of fitness in order to score the genetic interactions and determine functional genetic dependencies (60,61). If the observed fitness of a double-gene knockout mutant is greater than the expected fitness, the interaction is scored as a positive genetic interaction, including genetic suppression, where the fitness defect of a mutant is alleviated by a mutation in a second gene (62). On the other hand, if the observed fitness of a double-gene knockout mutant is more severe than the expected fitness, the genetic interaction is scored as a negative interaction, which manifests high genetic dependencies of the two genes and their associated pathways for cell viability and survival. An extreme example of a negative genetic interaction is synthetic lethality, where the combined effect of the mutant and the genetic perturbation causes cell death (63).

In an impressive study, the genome of *S. cerevisiae* was interrogated for all pairwise viable mutants (64). The budding yeast genome includes approximately 6,000 genes, about 1,000 of which are essential. The viable mutants can be constructed by knocking out any of the approximately 5,000 non-essential genes, by reducing the expression of the essential genes, or by partially compromising the functionality of the gene products. In this study, a double-gene knockout mutant was constructed by introducing a second gene mutation to about 4,000 viable mutants, resulting in about 5.4 million double-gene knockout mutants. This study included a large fraction of all possible double-gene knockout mutants in yeast, covering most non-essential genes and about 300 essential genes. Analysis of these mutants for fitness growth or defect produced genetic interaction profiles that aided with functional annotation of poorly characterized or even uncharacterized genes. The efforts in this study resulted in a genetic similarity network by computing pairwise profile similarity of genes. These genetic interactions were further extended to all combinations of essential and non-essential genes that formed a global genetic similarity network for *S. cerevisiae* (65).

1.1.2 Generating and interpreting chemical-genetic interaction profiles

If an array of genetic mutants is exposed to a compound, the resulting fitness-based interaction scores will represent the chemical-genetic interaction profile for the compound (66,67). The generation of these profiles for hundreds or even thousands of compounds has become practically feasible by genetically barcoding mutants and utilizing next generation sequencing technologies (68–70).

Barcode sequencing technologies enable the fitness-based interaction screening of mutants in a single, pooled culture that offers a substantial gain in efficiency over screening arrays of individual mutants (68,69). Barcode sequencing, which is performed by inserting unique DNA barcodes or tags together with common primer sites for Polymerase Chain Reaction (PCR)-based amplifications into individual mutants, allows the entire collection of mutants to be pooled in a single plate and grown in competition. For such experiments, the abundance of genetic barcodes can be used to measure the fitness of genetic mutants. This approach is particularly beneficial for generating chemical-genetic interaction profiles for rare or expensive compounds as the pooled format substantially reduces the amount of the compounds required for the experiments. In addition, highly multiplexed sequencing technologies enabled a substantial improvement in the throughput of barcoded chemical-genetic interaction screens (70). Current next-generation sequencing technologies provide the ability to simultaneously quantify a mutant collection perturbed by several compounds through massive parallelism.

Generally, a chemical-genetic interaction profile of a compound contains the degree of sensitivity or resistance of each genetic mutant to the compound. This quantitative, high-dimensional representation of a compound's biological function can be interpreted using a global compendium of genetic interaction profiles (64,65) systematically mapped through pairwise, double-gene knockout screens in *S. cerevisiae*. Specifically, the chemical-genetic interaction profile of a compound mimics the genetic interaction profile obtained through genetic knockout of the biomolecular target of the compound (71). These chemical-genetic interaction profiles can be treated as unbiased functional profiles for compounds.

Although barcode and next generation sequencing technologies enhance the speed of generation of chemical-genetic interaction profiles for compounds, screening thousands of genetic mutants (e.g., in whole-genome sequencing) is still a bottleneck. One strategy to facilitate the generation of chemical-genetic interaction profiles is to take advantage of the functional redundancy in genome-wide functional profiles and establish a subset of the whole-genome mutants as a miniaturized, diagnostic mutant collection that captures most of the functional information contained within genome-wide genetic interaction profiles (72). This strategy provides a powerful mechanism to scale up the generation of chemical-genetic interaction profiles at lower experimental time and costs.

We recently generated and published a compendium of more than 13,000 chemical-genetic interaction profiles obtained in *S. cerevisiae* (73). These compounds originate from the RIKEN Natural Product Depository and several diverse compound collections. These chemical-genetic interaction profiles were processed using the computational tools (74,75), validated in experimental settings (73), and deposited to the MOSAIC public database (76). I have primarily used a sub-collection of about 1,500 compounds that we annotated to their bioprocess predictions with high-confidence (73) for the purpose of this dissertation.

1.2 Computational approaches in drug discovery

1.2.1 Prioritizing compounds using virtual screening

The most dominant technique for the identification of new lead compounds in drug discovery is the physical screening of large libraries of small molecules against a biological target using high-throughput screening technologies. Virtual screening of compounds, as a

complementary approach to high-throughput screening, is to computationally screen large libraries of small molecules for the compounds that will likely bind to a target (e.g., protein receptors, enzymes, etc.) (77–79). With the aid of virtual screening approaches, only the candidates predicted to possess the highest likelihood of binding to a specific target are experimentally tested (instead of the whole library). The main idea behind virtual screening is to narrow down the vast, growing libraries of the chemical space to biologically relevant candidates for further investigations in experimental settings.

Virtual screening methods fall into two general classes: structure-based and ligand-based virtual screening methods (80). Structure-based virtual screening mainly utilizes the three-dimensional structure of the biological target, obtained from X-ray, nuclear magnetic resonance spectroscopy, or even computational modeling, to dock a library of compounds into the binding site and select a subset of these compounds based on the predicted binding scores for further experimental studies (81–83). On the other hand, ligand-based virtual screening approaches use the information of known active ligands rather than the structure of a target protein for lead identification and optimization (84). These approaches usually utilize the two-dimensional structural representation of molecules and chemical similarity searching methods (85) to mine large libraries of compounds such as ZINC (86) for targets.

1.2.2 Accelerating discoveries using artificial intelligence

Developing a new biopharmaceutical currently costs multiple billion dollars for the industry. Most of this money gets wasted as this includes the money spent on the nine out of ten candidates that fail somewhere between phase I trials and regulatory approvals (87).

As a result, few people in the field doubt the need for more efficient drug discovery and development strategies.

Artificial intelligence in general, and machine learning approaches in specific, have gained popularity in predicting a variety of needs related to the field of drug discovery and development (88,89). These applications include predictions of physiochemical properties, bioactivity, toxicity, and protein targets for chemical compounds. For example, AlphaFold is an artificial intelligence-based system developed by DeepMind that predicts the three-dimensional structure of a protein from its amino acid sequence (90), and Chemprop is a message passing neural network that predicts molecular properties directly from the graph structure of the molecule (91). Moreover, the DeepTox framework is a machine learning-based tool that predicts the toxicity of potential drugs (92).

Recently, artificial intelligence approaches have demonstrated their power in the discovery of new antimicrobials or agents against the COVID-19 disease. Halicin, a novel antibiotic discovered using deep neural network models, is a great example. Using machine learning models to predict the bioactivity and toxicity of small molecules, researchers were able to identify a kinase inhibitor from the Drug Repurposing Hub as a potent inhibitor of *E. coli* growth (93). In a different study, researchers found two new drug combinations of remdesivir/reserpine and remdesivir/IQ-1S as effective therapeutic solutions for treatment of SARS-CoV-2, the cause of the COVID-19 disease, using deep neural network models that jointly learned drug-target interactions and drug-drug synergy to mine new synergistic drug combinations (94).

1.2.3 Identifying gene targets of chemical compounds

The early functional profiling experiments with the aim of identifying drug-target interactions were conducted in *S. cerevisiae* using a collection of heterozygous diploid mutants, where the copy number of an essential gene was reduced from two to one for a genetic mutant (95). The general intuition behind using heterozygous diploid mutants for target identification of compounds is that, if a gene is essential for growth, the heterozygous diploid mutant of the gene whose product is targeted by a drug is sensitive to the drug. This sensitivity is observable through a negative genetic interaction and interpreted as the gold standard for the strength of the drug–target interaction. Although the heterozygous diploid mutants are referred as the gold standard for target identification of compounds, additional experiments such as gene overexpression data (96) and other experiments related to both essential and non-essential genes have also proven their value in this space.

In addition, computational approaches are powerful tools for target identification of compounds. For example, we previously developed a computational pipeline called CG-TARGET that annotated large libraries of compounds to their biological processes using chemical-genetic interaction profiles (74). The CG-TARGET tool integrated these large-scale chemical-genetic interaction data with a global network of genetic interactions in *S. cerevisiae* (64,65) to predict the biological processes perturbed by compounds. We applied CG-TARGET to several large compound libraires including a collection of nearly 14,000 compounds screened in *S. cerevisiae*, where we prioritized ~1,500 compounds with high-confidence biological process predictions for further study (73).

1.3 Outline

In this dissertation, I present my efforts to advance computational approaches for drug discovery and development using large libraries of chemical-genetic interactions. As part of this research work, I discuss novel analytical, computational, and machine learning frameworks for functional characterization and target identification of compounds.

Chapter 2 presents a benchmarking system for identifying the structural similarity measures that accurately connect the structural molecular space to the functional space for chemical compounds. The results of this chapter introduce the new measure of All-Shortest Path fingerprints combined with the Braun-Blanquet similarity coefficient as the superior choice over all other studied combinations. Notably, this new structural similarity measure outperformed the measure resulted from the use of the Morgan fingerprints and Tanimoto similarity coefficient, which are commonly used in chemical structural similarity efforts. Moreover, I developed a machine learning pipeline that improved functional prediction of compounds from their chemical structures. This machine learning framework was able to assign high predictive scores to a compound's functional analogs that also exhibited high structural diversity from the compound. This research work is published by the American Chemical Society at the Journal of Chemical Information and Modeling, Volume 61, Issue 9, 2021.

Chapter 3 presents a computational and analytical approach for target identification of compounds by integrating chemical-genetic interaction data from several experimental modalities. This computational approach integrated inhibitory signals from several pools, among which heterozygous diploid mutants were used as the gold standard for gene target

identifications. This computational framework provided an accurate ranking system that assigned to each compound-gene combination a unified score reflecting the likelihood of the compound to target the gene. The potential compound-target pairs that had high ranking scores were tested in experimental settings, which resulted in the identification of targets for several chemical compounds.

Chapter 4 presents a computational framework for target identification of chemical compounds using a network of genetic similarities. In this approach, I modeled a potential gene target of a compound as a rumor source that could spread its inhibitory signals across a genetic similarity network (or generally, any functional network, such as a protein-protein interaction network). This framework received a set of bioprocess target predictions as its input and produced rumor centrality scores for the associated genes, where a higher score reflected a higher likelihood of the gene to be the target of the compound. I evaluated these rumor centrality scores in several computational settings.

Chapter 5 summarizes the conclusions from across all the projects I have completed as part of my dissertation and my overall contributions to the field of drug discovery. This chapter also provides future directions related to my findings of this dissertation.

2 Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions*

This research work was published by the American Chemical Society at the Journal of Chemical Information and Modeling, Volume 61, Issue 9, 2021. Other contributors to this research project, ordered by their contribution, were Scott W. Simpkins, Justin Nelson, Sheena C. Li, Jeff S. Piotrowski, Mami Yoshimura, Yoko Yashiroda, Hiroyuki Hirano, Hiroyuki Osada, Minoru Yoshida, Charles Boone, and Chad L. Myers.

H.S. (dissertation's author) and C.L.M. conceived the project and established the methodologies. H.S. developed the benchmarking systems and machine learning models, generated the molecular fingerprints for the models, and performed all the analyses. H.S. constructed and validated the VS-SVM machine learning-based prediction framework for ligand-based virtual screening. S.W.S. contributed to the conception of the benchmarking systems, generation of the molecular fingerprints using the jCompoundMapper tool, and provision of detailed feedback on the results and visualizations. S.W.S., J.N., S.C.L., J.S.P., Mami Yoshimura, Y.Y., H.H., and H.O. generated and validated yeast chemical-genetic interaction data. Minoru Yoshida, C.B., and C.L.M. supervised the project and acquired funding. H.S. wrote the original draft, and all the authors reviewed and edited this draft.

---

* This chapter is reprinted with permission from Safizadeh, H. *et al.* Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical–Genetic Interactions. *Journal of Chemical Information and Modeling* **61**, 4156–4172 (2021). Copyright 2021 American Chemical Society. URL: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00993.

2.1 Abstract

A common strategy for identifying molecules likely to possess a desired biological activity is to search large databases of compounds for high structural similarity to a query molecule that demonstrates this activity, under the assumption that structural similarity is predictive of similar biological activity. However, efforts to systematically benchmark the diverse array of available molecular fingerprints and similarity coefficients have been limited by a lack of large-scale datasets that reflect biological similarities of compounds. To elucidate the relative performance of these alternatives, we systematically benchmarked 11 different molecular fingerprint encodings, each combined with 13 different similarity coefficients, using a large set of chemical-genetic interaction data from *Saccharomyces cerevisiae* as a systematic proxy for biological activity. We found that the performance of different molecular fingerprint encodings and similarity coefficients varied substantially, and that the All-Shortest Path (ASP) fingerprint encoding paired with the Braun-Blanquet similarity coefficient provided superior performance that was robust across several compound collections. We further proposed a machine learning pipeline based on support vector machines that offered a fivefold improvement relative to the best unsupervised approach. Our results generally suggest that using high-dimensional chemical-genetic data as a basis for refining molecular fingerprints can be a powerful approach for improving prediction of biological function from chemical structure.

2.2 Introduction

The development of new drugs that demonstrate desired biochemical behavior against a biomolecular target is challenging. Despite the scientific and technological advances in drug discovery during the past 60 years, the ratio of drugs approved to money spent on R&D has halved roughly every 9 years since 1950 (Eroom's Law in contrast to Moore's Law) (97), and has now dipped below one drug per billion USD (3,97,98). Following the similar property principle (SPP) (99), ligand-based virtual screening has been commonly used to filter candidates prior to high-throughput screening efforts by ranking compounds from a large database in descending order of their structural similarity to a reference or target molecule with known biological activity (**Figure 2.1**) (100,101). According to SPP, structurally similar molecules will more likely possess similar biological activities and physicochemical properties. Despite limitations to SPP (102,103), such as activity cliffs that manifest when a small structural modification drastically alters the biological properties of a compound (104), this structure-activity relationship is broadly consistent throughout the larger flat regions of activity landscapes (105,106). Improving the structural similarity-based retrieval of biologically similar compounds will therefore benefit a multitude of drug discovery efforts.

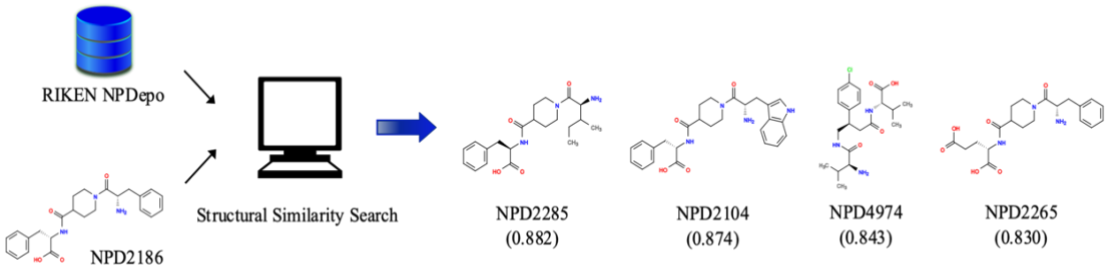


Figure 2.1. Ligand-based virtual screening of a target compound (e.g., NPD2186 from RIKEN Natural Product Depository). We ranked all compounds (top four shown) of the MOSAIC database (http://mosaic.cs.umn.edu) (76) in descending order of structural similarity to the target molecule based on SPP. We described our compounds using All-Shortest Path (ASP) fingerprints (depth 8) and measured structural similarity using the Braun-Blanquet similarity coefficient. In this example ranked list, three compounds except NPD4974 have a very similar chemical-genetic profile to that of NPD2186.

The chemical informatics community has suggested a wide range of molecular fingerprint encodings and similarity coefficients to quantify structural similarity between two molecules (85,107,108), which is then used as a proxy for biological similarity. The most widely used class of molecular encodings is chemical fingerprints (109–113), where a molecular graph is represented by a bit-vector that encodes the presence or absence of molecular features such as paths between atoms, substructural fragments, and pharmacophores. The degree of similarity of two structural vectors describing two different compounds is usually measured by similarity coefficients. The Tanimoto (aka Jaccard) coefficient, formulated as the fraction of features in common between two molecules relative to the total number of features presented in either molecule, remains the coefficient

of choice to capture the highest level of intermolecular similarity and thus biological activity (114–117), although this coefficient suffers from an intrinsic bias toward selecting smaller compounds (85,118–120). However, a major challenge for the research community has been the lack of a systematic benchmarking framework based on biological activity that covers a broad range of protein targets. Such a framework will enable a definitive assessment of the performance of molecular fingerprints and similarity coefficients to predict biological similarity between two molecules.

Chemical genomic approaches, which involve the systematic mapping of chemical-genetic interactions, offer a valuable new source of data to connect chemical structure to biological function. The yeast *Saccharomyces cerevisiae* is a well-characterized eukaryote for which ~5000 viable deletion mutants have been identified (121). Interrogating these deletion mutants with a bioactive compound generates a chemical-genetic interaction profile containing the degree of sensitivity or resistance of each genetic mutant to the compound (66,67). This quantitative, high-dimensional representation of biological function can be interpreted using a global compendium of genetic interaction profiles (64,65) systematically mapped through pairwise, double-knockout screens in *S. cerevisiae*. Specifically, the chemical-genetic interaction profile of a compound will resemble the genetic interaction profile obtained through the genetic knockout of its biomolecular target (71). These chemical-genetic interaction profiles can be used as a gold standard to evaluate structural similarity measures (each defined as one molecular fingerprint encoding paired with one similarity coefficient), even for compounds where the actual biological targets are not yet known, since these profiles accurately reflect compound function.

We recently generated and published a compendium of over 13,000 chemical-genetic interaction profiles obtained in *S. cerevisiae* (73), providing a unique basis to independently assess the performance of structural similarity measures. These compounds originate from the RIKEN Natural Product Depository (NPDepo) and several NCI, NIH, and GlaxoSmithKline (GSK) compound collections. We systematically benchmarked 11 different molecular fingerprint encodings, each combined with 13 different similarity coefficients, to identify the pair that best predicted biological similarity as measured by chemical-genetic interaction profile similarity (in brief, chemical-genetic similarity). Given recent advances in machine learning approaches for drug discovery (93,122–125), we, furthermore, developed several supervised machine learning models to improve our predictions of the biological activity of our compounds from chemical structures, achieving higher predictive power than using similarity measures alone. We found that support vector machines (SVM) models (126,127) can significantly enhance the power of our chemical fingerprints for predicting the biological activity of compounds.

2.3 Results and discussion

2.3.1 Establishing a systematic benchmark for structural similarity measures

We compiled a list of commonly used molecular fingerprint encodings (**Table 2.1**) and similarity coefficients (**Table 2.2**) to capture the structural similarity of compounds in different modes (e.g., exploring molecular paths or radial atom environments). To quantify the ability of different structural similarity measures for predicting functionally analogous compounds, we used chemical-genetic interaction profiles from *S. cerevisiae* as a

systematic genome-wide standard for biological activity. We previously published chemical-genetic interaction profiles for 13,431 chemical compounds from several diverse compound collections and identified a subset of these compounds that were predicted with high-confidence to perturb predefined biological processes based on integrating these profiles with genetic interaction profiles (73). We included in our benchmarking system two independent subsets of these high-confidence compounds: (1) 826 compounds from the RIKEN NPDepo collection, which we hereafter call the "RIKEN high-confidence set", and (2) 659 compounds from several NCI/NIH/GSK collections, which we hereafter call the "NCI/NIH/GSK high-confidence set" (Materials and methods).

Since we require a binary gold standard to calculate prediction performance using precision and recall, we selected the 10% most similar compound pairs based on the (cosine) similarity of their chemical-genetic interaction profiles and labeled them as our gold standard for true positive compound pairs. This assumption is imperfect because chemical-genetic similarity of compounds does not always imply structural similarity and two functionally analogous compounds may have distinct chemical structures. However, a large fraction of these true positives (functionally similar compound pairs) should still be identified from the structural similarity of compounds; therefore, this chemical-genetic-derived similarity serves as a reasonable basis for identifying compounds with similar biological effects.

Table 2.1. Molecular fingerprints. 11 different molecular fingerprint encodings generated by the jCompoundMapper tool (128) or RDKit toolkit (version 2020.09.4) (129) for describing the compounds in our datasets. The fourth column, Features, represents the total number of features that the jCompoundMapper tool or RDKit toolkit generated for describing our RIKEN high-confidence set (826 compounds). The molecular features we counted were only the ones present in the description of at least one compound of this collection, and the molecular fingerprints that required a depth of description were measured at depth 8.

ID	Name	Description	Features	Reference(s)
FP1	AP2D	Topological atom pairs	1211	(130)
FP2	ASP	All-shortest paths	26194	(131)
FP3	AT2D	Topological atom triplets	56963	(130)
FP4	DFS	All-paths (depth-first search)	48448	(132)
FP5	ECFP	Extended connectivity fingerprints	42672	(133)
FP6	LSTAR	Local path environments	85232	(134)
FP7	MACCS	MDL public keys (166 keys)	155	(135)
FP8	PHAP2POINT2D	Topological pharmacophore pairs	17	(136)
FP9	PHAP3POINT2D	Topological pharmacophore triplets	302	(136)
FP10	RAD2D	Topological molprint-like fingerprints	92191	(134)
FP11	RDKit	Topological daylight-like fingerprints	65183	(129,137)

While studying compounds using yeast chemical-genetic interaction screens benefits from many advantages (e.g., genome-wide detection of potential targets), some important aspects of compound activity such as pharmacokinetic properties will be missed. Since our modeling approaches are based on functional similarity as detected in yeast chemical-genetic interactions, our results and conclusions about the functional similarity predictions of compounds will not reflect these characteristics, including pharmacokinetic properties, of compounds.

Table 2.2. Similarity coefficients. 13 different similarity coefficients (several of these coefficients were collected by Raymond and Willett) (138) for measuring the degree of structural similarity of two compounds described by a given molecular fingerprint encoding. Here, x is the number of bits set in both fingerprints, y is the number of bits set in the first fingerprint, z is the number of bits set in the second fingerprint, and w is the total number of bits in the bit string. For the Tullos similarity coefficient, $X = \log\left(1 + \frac{min(y,z)}{\max(y,z)}\right)/\log(2)$, $Y = \left(\log\left(2 + \frac{min(y,z)-x}{x+1}\right)/\log(2)\right)^{-\frac{1}{2}}$, and $Z = \log\left(1 + \frac{x}{y}\right)\log\left(1 + \frac{x}{z}\right)/\log^2(2)$. For the asymmetric evaluation of the Tversky similarity coefficient, $\alpha = 0.9$. Here, we assume that the parameters of the Tversky coefficient in its original formulation, $Tversky = \frac{x}{x+p(y-x)+q(z-x)}$, will follow p + q = 1. The Dice and Tanimoto similarity coefficients are two symmetric instances of the Tversky coefficient, where p = q = 0.5 and p = q = 1 respectively.

Name	Measurement	Range	Reference(s)
Braun-Blanquet	$x/\max(y,z)$	0 to 1	(139,140)
Cosine	$\frac{x}{\sqrt{yz}}$	0 to 1	(141,142)
Dice	$\frac{2x}{y+z}$	0 to 1	(143,144)
Dot-product	x	0 to ∞	N/A
Euclidean	$\frac{1}{1+\sqrt{y+z-2x}}$	0 to 1	N/A
Kulczynski	$\frac{x(y+z)}{2yz}$	0 to 1	(145)
McConnaughey	$\frac{x(y+z)-yz}{yz}$	-1 to 1	(146)
Russel/Rao	x/w	0 to 1	(147)
Simpson	$x/\min(y,z)$	0 to 1	(148)
Sokal/Sneath	$\frac{x}{2y+2z-3x}$	0 to 1	(149)
Tanimoto	$\frac{x}{y+z-x}$	0 to 1	(150,151)
Tullos	XYZ	0 to 1	(152)
Tversky	$\frac{x}{x+\alpha(y-x)+(1-\alpha)(z-x)}$ $\alpha\epsilon[0,1]$	0 to 1	(153)

We evaluated both components (molecular fingerprints and similarity coefficients) of our systematic benchmark to find both the fingerprint and the coefficient that best predicted the biological activity of our compounds (**Algorithm A.1**). We described all our compounds in 11 different fingerprint-based structural encoding spaces (**Table 2.1**), where a compound was described by a bit-vector that indicated the presence or absence of certain molecular features. The number of features required for the description of the compounds in a structural space varied based on the space definition and collection properties (e.g., only 155 out of 166 predefined molecular substructures from MDL Public Keys were used in the MACCS description of our RIKEN high-confidence set, while RAD2D structural encodings generated 91082 features to describe the same collection). We used 13 widely used similarity coefficients (**Table 2.2**) to measure the degree of similarity of two chemical compounds described by a given molecular fingerprint.

2.3.2 Evaluating the performance of structural similarity measures

We computed pairwise structural similarities for the compounds in our RIKEN high-confidence set (340,725 different compound pairs) using every combination of a molecular fingerprint and a similarity coefficient (143 different prediction models). For every model, we sorted pairwise structural similarity scores in the descending order against our binarized gold-standard functional (chemical-genetic) similarities and calculated precision at predefined recall thresholds (**Table A.1** for all precision values and complete evaluation of all prediction models). To isolate the best-performing molecular fingerprint, we examined the precision of all prediction models at several predefined recall thresholds

for our RIKEN high-confidence set (**Table A.1**). Based on these evaluation metrics, the ASP, LSTAR, and RAD2D fingerprints emerged as the molecular fingerprint encodings with superior predictive power (**Figure 2.2** and **Table A.1**). The wide range of precision values achieved by different molecular fingerprints revealed that our chemical-genetic interaction profiles can highly discriminate molecular fingerprints in terms of their efficacy for predicting the biological activity of our compounds.

Notably, among the molecular fingerprints that we evaluated was the Morgan fingerprint (154), also known as the extended-connectivity fingerprint (ECFP) (133), which has recently been one of the best performing fingerprints in a variety of applications such as small molecule virtual screening (155–157). However, ECFP was generally outperformed by ASP, LSTAR, and RAD2D fingerprints, except for a small number of recall thresholds (**Figure 2.2B**). Because this finding might simply be a result of the RIKEN NPDepo compound collection characteristics, we used our NCI/NIH/GSK high-confidence set to validate the prediction performance of ASP, LSTAR, and RAD2D fingerprints, using the same evaluation method designed for our RIKEN high-confidence set. Our analysis on this second large set of compounds strongly confirmed the superiority of these fingerprints over ECFP at many recall thresholds (**Figure 2.3** and **Table A.2** for complete evaluation of all prediction models).

We determined that the ASP and LSTAR fingerprints best captured the relationship between chemical structure and biological activity of our compounds. ASP fingerprints encode graph traversals over all atoms in a molecular graph but store only the shortest paths between atoms, whereas LSTAR and RAD2D fingerprints describe the radial environment

of all atoms in the molecular graph.(128) As a result, ASP structural encodings needed fewer features than LSTAR and RAD2D encodings to describe our compound collections (**Table 2.1** – Features column) although the prediction performance of ASP fingerprints was higher or comparable with that of LSTAR and RAD2D fingerprints at most recall thresholds. Moreover, LSTAR fingerprints generally exhibited higher performance than RAD2D fingerprints at several recall thresholds (**Figure 2.2** and **Figure 2.3**), which could be driven by the additional bond information that LSTAR fingerprints would collect from the radial environment of the atoms.

While the choice of a molecular fingerprint is important (as demonstrated above) to maximize the biological relevance of chemical similarity calculations, such fingerprints must be paired with a similarity coefficient to compute chemical similarity. To characterize the utility of different similarity coefficients, we benchmarked a large set of 13 different similarity coefficients (**Table 2.2**) that were applied to the complete set of candidate molecular fingerprints. Again, using our RIKEN high-confidence set as the basis, we measured the prediction performance of every coefficient over all molecular fingerprints to find the best performing similarity coefficient (**Table A.1**). We found that several (nine out of 13) similarity coefficients were able to consistently exhibit high performance across all molecular fingerprints; however, four coefficients (Dot-product, Euclidean, Russel/Rao, and Simpson similarity coefficients) failed in several molecular fingerprint spaces because their precision substantially dropped at many recall thresholds (**Table A.1**). Because these four coefficients were unable to capture the biological similarity of our compounds given the provided molecular fingerprints, we focused only on the nine remaining coefficients.

A

	Recall = 0.002			Recall = 0.005			Recall = 0.02		
	BB	Cos	Tan	**BB**	Cos	Tan	**BB**	Cos	Tan
FP1	0.861	0.850	0.850	0.676	0.684	0.683	0.370	0.355	0.357
ASP	0.912	**0.919**	**0.919**	0.830	0.829	0.819	**0.586**	0.557	0.567
FP3	0.866	0.849	0.850	0.770	0.785	0.785	0.523	0.515	0.515
FP4	0.871	0.877	0.888	0.740	0.765	0.767	0.543	0.540	0.544
FP5	0.861	0.855	0.854	0.777	0.743	0.749	0.556	0.555	0.556
LSTAR	0.883	0.883	0.883	0.828	0.833	0.833	0.560	0.565	0.566
FP7	0.850	0.834	0.834	0.677	0.668	0.670	0.464	0.487	0.488
FP8	0.138	0.138	0.138	0.044	0.044	0.044	0.049	0.049	0.049
FP9	0.056	0.056	0.056	0.081	0.081	0.081	0.108	0.108	0.108
RAD2D	0.861	0.875	0.866	**0.834**	0.810	0.814	0.548	0.535	0.536
FP11	0.804	0.883	0.872	0.746	0.772	0.770	0.498	0.501	0.503
Mean	**0.724**	**0.729**	**0.728**	**0.637**	**0.638**	**0.638**	**0.437**	**0.433**	**0.435**

	Recall = 0.05			Recall = 0.2			Area Under the Curve (AUC)		
	BB	Cos	Tan	**BB**	Cos	Tan	**BB**	Cos	Tan
FP1	0.230	0.221	0.222	0.156	0.154	0.154	0.593	0.590	0.592
ASP	**0.294**	0.288	0.287	0.158	0.154	0.156	0.584	0.575	0.579
FP3	0.256	0.254	0.259	0.152	0.152	0.153	0.584	0.581	0.583
FP4	0.279	0.267	0.270	0.158	0.154	0.156	0.580	0.572	0.576
FP5	0.277	0.263	0.268	0.153	0.146	0.149	0.584	0.574	0.578
LSTAR	**0.294**	0.288	0.290	**0.171**	0.168	0.170	**0.596**	0.587	0.590
FP7	0.261	0.278	0.278	0.157	0.159	0.159	0.583	0.587	0.587
FP8	0.062	0.062	0.062	0.091	0.091	0.091	0.501	0.499	0.499
FP9	0.109	0.109	0.109	0.101	0.098	0.098	0.492	0.486	0.488
RAD2D	**0.294**	0.289	0.291	0.164	0.161	0.162	0.588	0.579	0.582
FP11	0.228	0.235	0.231	0.145	0.143	0.144	0.574	0.573	0.574
Mean	**0.235**	**0.232**	**0.233**	**0.146**	**0.144**	**0.145**	**0.569**	**0.564**	**0.566**

B

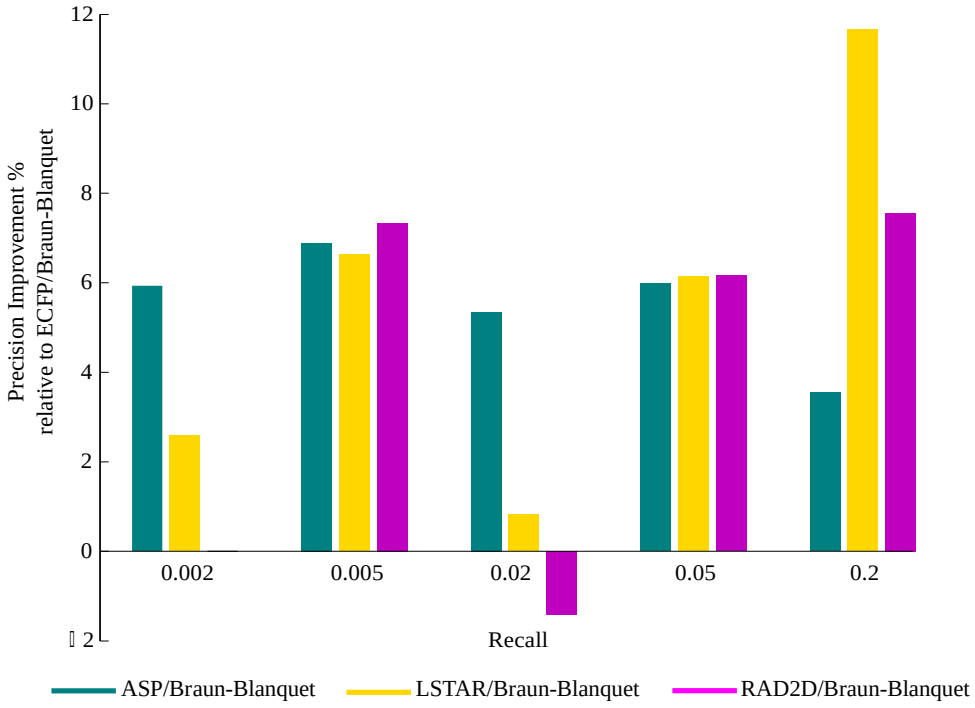


Figure 2.2. Performance of selected prediction models using our RIKEN high-confidence set (**Table A.1** for the complete evaluation of all prediction models). **(A)** Precision at several recall thresholds and the area under the ROC curve for each model, where a molecular fingerprint was paired with the Braun-Blanquet, Cosine, or Tanimoto similarity coefficient, evaluated based on chemical-genetic similarity as the gold standard for biological activity. The blue values represent the highest precision achieved at a given recall, and the green values represent the average precision over all molecular fingerprints for a similarity coefficient at a specific recall threshold. **(B)** Relative performance of ASP, LSTAR, and RAD2D fingerprints to ECFP. For all the molecular fingerprints that required a depth of description, precision was measured at depth 8. With the exception of ASP, LSTAR, and RAD2D, the remaining molecular fingerprints are coded FP1-FP11 (**Table 2.1**).

We found that the Braun-Blanquet similarity coefficient (139,140) resulted in the highest precision at many recall thresholds compared to all other coefficients (**Table A.1**), including the Tanimoto and cosine coefficients (**Figure 2.2A**), which have commonly been used by the chemical informatics community. The Braun-Blanquet coefficient is simply computed as the fraction of features shared between two molecules to the total number of features presented in the larger molecule. For the Braun-Blanquet coefficient, the average and maximum precisions and the average and maximum areas under the receiver operating characteristic (ROC) curves across all molecular fingerprints were slightly higher at many recall thresholds compared to those of the Tanimoto and cosine coefficients (**Figure 2.2A** and **Table A.1** – green and blue values), suggesting that this simple coefficient of structural similarity can confidently be used in ligand-based virtual screening as a substitute for the traditional Tanimoto coefficient.

A

	Recall = 0.002			Recall = 0.005			Recall = 0.02		
	BB	Cos	Tan	**BB**	Cos	Tan	**BB**	Cos	Tan
FP1	0.781	0.818	0.813	0.644	0.594	0.609	0.332	0.312	0.319
ASP	**0.932**	0.839	0.839	0.813	0.789	0.810	0.529	0.535	0.541
FP3	0.884	0.841	0.828	0.801	0.819	0.819	0.461	0.518	0.495
FP4	0.916	0.828	0.828	0.807	0.772	0.789	0.490	0.491	0.495
FP5	0.916	0.897	0.897	0.795	0.783	0.783	0.400	0.373	0.395
LSTAR	0.916	0.897	0.897	**0.871**	0.856	0.870	0.520	0.533	0.527
FP7	0.812	0.812	0.812	0.805	0.785	0.782	0.429	0.464	0.461
FP8	0.110	0.110	0.110	0.144	0.144	0.144	0.185	0.185	0.185
FP9	0.280	0.280	0.280	0.234	0.234	0.234	0.175	0.175	0.175
RAD2D	0.878	0.861	0.861	0.858	0.851	0.851	0.472	0.456	0.462
FP11	0.811	0.796	0.796	0.795	0.779	0.778	0.545	0.550	**0.554**
Mean	**0.749**	**0.725**	**0.724**	**0.688**	**0.673**	**0.679**	**0.413**	**0.417**	**0.419**

	Recall = 0.05			Recall = 0.2			Area Under the Curve (AUC)		
	BB	Cos	Tan	**BB**	Cos	Tan	**BB**	Cos	Tan
FP1	0.220	0.233	0.235	0.156	0.153	0.156	**0.577**	0.572	0.575
ASP	0.290	0.272	0.282	0.162	0.156	0.161	0.576	0.567	0.572
FP3	0.265	0.273	0.273	**0.165**	0.154	0.161	0.573	0.567	0.571
FP4	0.291	0.282	0.293	0.162	0.156	0.161	0.575	0.566	0.571
FP5	0.216	0.202	0.208	0.137	0.128	0.132	0.558	0.544	0.550
LSTAR	0.284	0.272	0.279	0.160	0.150	0.154	0.575	0.563	0.568
FP7	0.274	0.271	0.272	0.158	0.160	0.162	0.568	0.562	0.565
FP8	0.164	0.164	0.164	0.136	0.136	0.136	0.550	0.550	0.550
FP9	0.148	0.148	0.148	0.129	0.129	0.129	0.542	0.546	0.545
RAD2D	0.270	0.243	0.247	0.156	0.144	0.147	0.574	0.561	0.566
FP11	**0.299**	0.283	0.292	0.157	0.149	0.154	0.566	0.557	0.563
Mean	**0.247**	**0.240**	**0.245**	**0.152**	**0.147**	**0.150**	**0.567**	**0.560**	**0.563**

B

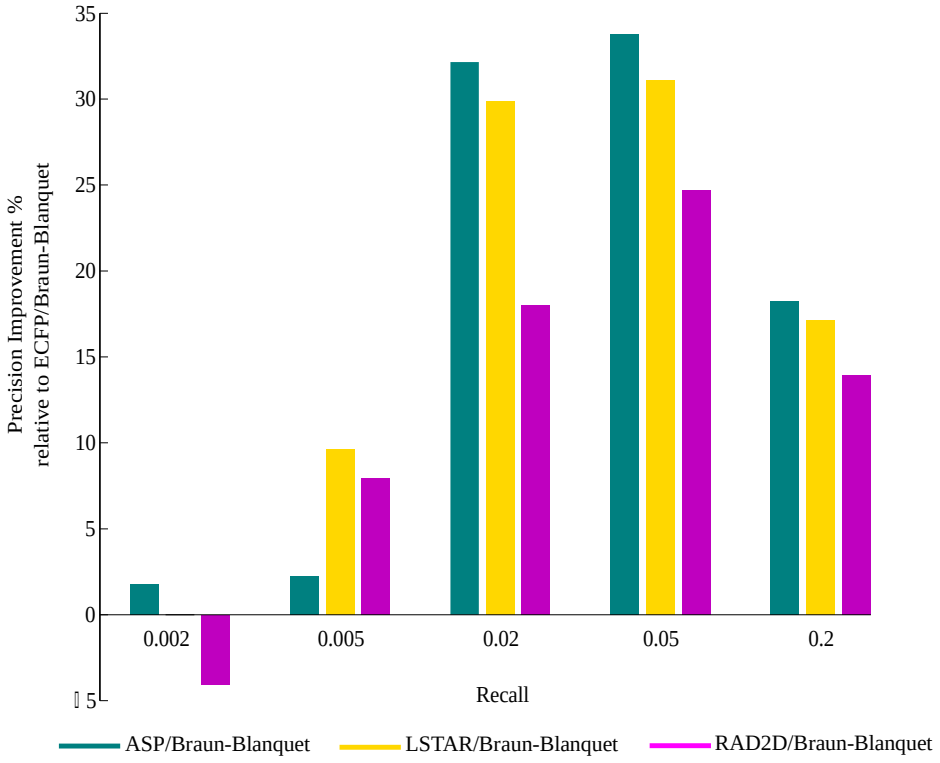


Figure 2.3. Performance of selected prediction models using our NCI/NIH/GSK high-confidence set (**Table A.2** for the complete evaluation of all prediction models). **(A)** Precision at several recall thresholds and the area under the ROC curve for each model, where a molecular fingerprint was paired with the Braun-Blanquet, Cosine, or Tanimoto similarity coefficient, evaluated based on chemical-genetic similarity as the gold standard for biological activity. The blue values represent the highest precision achieved at a given recall threshold, and the green values represent the average precision over all molecular fingerprints for a given similarity coefficient at a specific recall threshold. **(B)** Relative performance of ASP, LSTAR, and RAD2D fingerprints to ECFP. For all the molecular fingerprints that required a depth of description, precision was measured at depth 8. With the exception of ASP, LSTAR, and RAD2D, the remaining molecular fingerprints are coded FP1-FP11 (**Table 2.1**).

We further measured the performance of our prediction models using our NCI/NIH/GSK high-confidence set to validate the superiority of the Braun-Blanquet coefficient over other similarity coefficients (**Figure 2.3A** and **Table A.2**).

Given the prevalence of the Tanimoto similarity coefficient throughout chemical similarity calculations, we found our evaluation results especially interesting in that the Tanimoto similarity coefficient was outperformed in our systematic benchmark by the Braun-Blanquet similarity coefficient. Thus, we offer a potential explanation: For a pair of small molecules where the structural features of one molecule are a subset of the structural features of the other molecule, both similarity coefficients yield the same similarity values. However, these similarity values diverge if new features are added to the molecule with fewer structural features, and therefore, its feature set is no longer a subset of the feature

set of the larger molecule. Specifically, Tanimoto-based structural similarity decreases because the size of the total feature set across the two molecules increases, while Braun-Blanquet-based structural similarity remains unchanged. Thus, when retrieving similar molecules to a target molecule, the Tanimoto similarity coefficient will bias the results toward the sub- and supersets of the target molecule and against the molecules that share a common core with the target molecule but incorporate additional unique features. Braun-Blanquet-based similarity does not suffer from this bias, as the denominator of the Braun-Blanquet coefficient is merely the size of the largest molecular feature set, not the total number of features observed across both molecules. The higher performance of the Braun-Blanquet coefficient against our functional gold standard suggests that the bias inherent in the definition of the Tanimoto coefficient affects its performance in predicting the true functional similarity of compounds, and that a similarity coefficient such as the Braun-Blanquet coefficient can mitigate this issue.

Overall, we identified the combination of the Braun-Blanquet similarity coefficient and either ASP or LSTAR molecular fingerprints as the highest performing approach for measuring chemical structural similarity in ligand-based virtual screening. This conclusion was further confirmed through more formal "Sum of Ranking Distances" (SRD) analysis (**Figure A.1**, **Figure A.2**, and **Table A.1** and **Table A.2** – SRD Analysis spreadsheets) (158–160).

To our knowledge, no study has reported or used the combination of the ASP or LSTAR fingerprints paired with the Braun-Blanquet similarity coefficient as a measure of structural similarity.

2.3.3 Optimizing the depth of molecular fingerprints

One major parameter involved in the structural description of compound collections is the describing depth; a high depth generates features that describe the local and global environments of each atom, whereas a low depth focuses only on the local neighborhood of atoms in the molecular graph. We assessed the impact of the depth of five molecular fingerprints (ASP, DFS, ECFP, LSTAR, and RAD2D) in predicting the biological activity of our compounds using the Braun-Blanquet similarity coefficient (**Figure 2.4**).

At predefined recall thresholds, we observed that precision beyond a depth of 8 at best minimally improved the prediction of functional similarity from chemical structure and at worst decreased it markedly. Describing our RIKEN high-confidence set at a high depth generally resulted in strong predictions at lower recall thresholds. Structural description at a high depth was able to predict the compound pairs that were structurally and therefore functionally similar according to SPP. On the other hand, a low describing depth was able to capture the similarity of two molecular graphs in the local subgraphs that were essential for functional similarity, resulting in reasonable predictions at lower recall thresholds. We further evaluated these results using our NCI/NIH/GSK high-confidence set (**Figure A.3**), which confirmed a similar general trend across a range of recall thresholds using 10 different describing depths. Therefore, the structural description of a compound collection at high depths does not appear to be beneficial from our evaluation, and thus, the additional computational complexity and storage space for these structural vectors across hundreds or thousands of compounds may not be justified. For evaluations in this study, we selected depth 8 as the focus of our analysis.

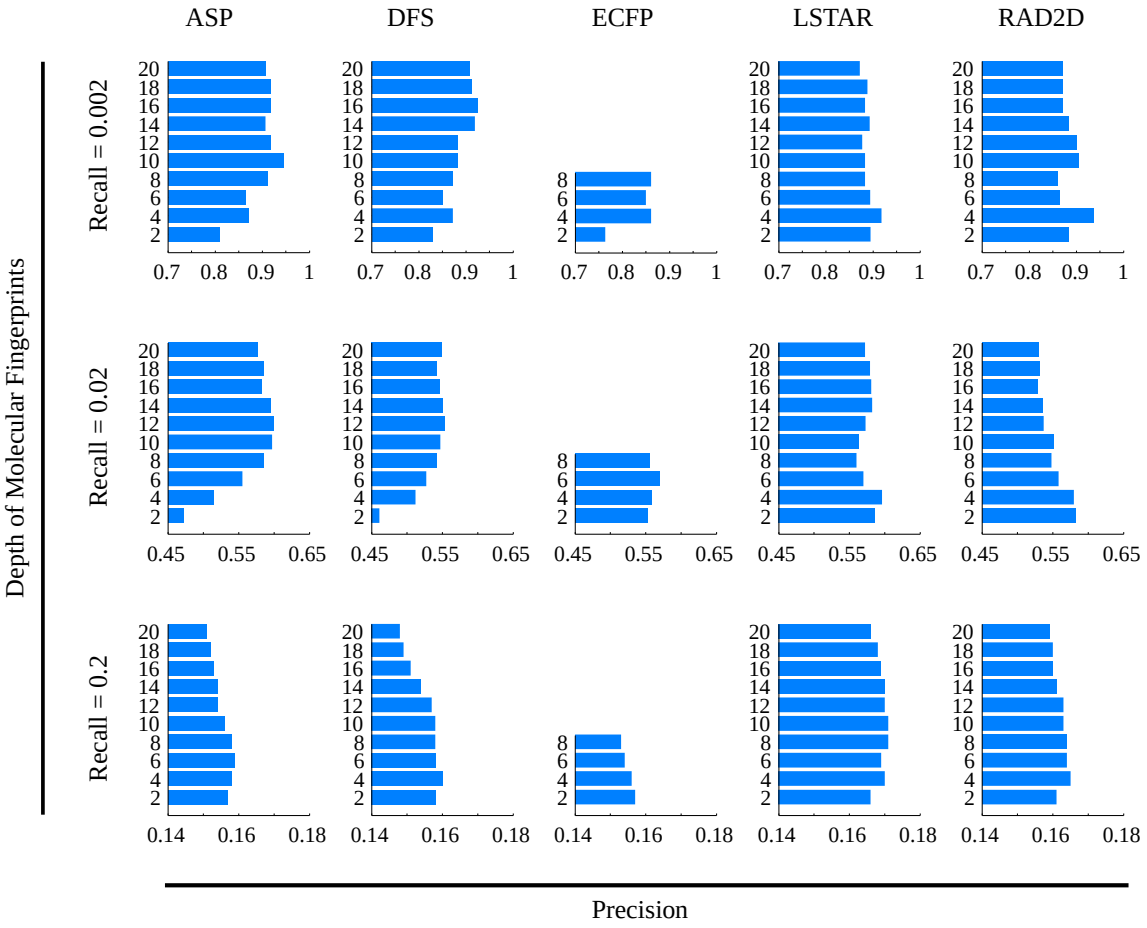


Figure 2.4. Impact of the describing depth of molecular fingerprints on the RIKEN high-confidence set. We measured the precision of our prediction models at 10 molecular depths, ranging from 2 to 20, for 5 different molecular fingerprints. Similarities were calculated with the Braun-Blanquet similarity coefficient, and the precision at three different recall thresholds for each molecular depth are shown.

2.3.4 Improving the prediction performance using SVM models

To increase our ability to predict the biological activity of compounds from their chemical structure, we designed several supervised machine learning models that leveraged chemical-genetic interaction profiles as training data. As a preprocessing step to reduce the

size of the feature space, we used supervised principal component analysis (161) to extract the most informative structural features given our functional similarity standard.

We designed a learning pipeline (**Figure 2.5A**) for predicting chemical-genetic similarities by creating bootstraps (162) from our compound collections and generating "pairwise structural vectors" (**Figure A.4**) that labeled the shared molecular features between two molecules (Materials and methods). This pipeline was implemented using support vector regression (SVR) models (126,127) with a Radial Basis Function (RBF) kernel. We employed precision-recall (PR) curves to evaluate the prediction performance of our learning models for different molecular fingerprints, where the compound pairs from our high-confidence sets with the highest 10% of chemical-genetic similarities were labeled as gold standard for true positives. Using this machine learning approach, we found that a subset of our molecular fingerprints (FP1-FP6, FP10, and FP11) enabled substantial improvements in performance for predicting compound functional similarity than our best-performing structural similarity measure (ASP/Braun-Blanquet measure). To evaluate the characteristics of our predicted supervised structural similarities, we computed the Pearson correlation between these pairwise similarities and the observed chemical-genetic similarity for both our compound collections (**Figure A.5**). This analysis demonstrated that predicted similarities were correlated with chemical-genetic similarities in both collections, and that the correlation increased using our machine learning approach.

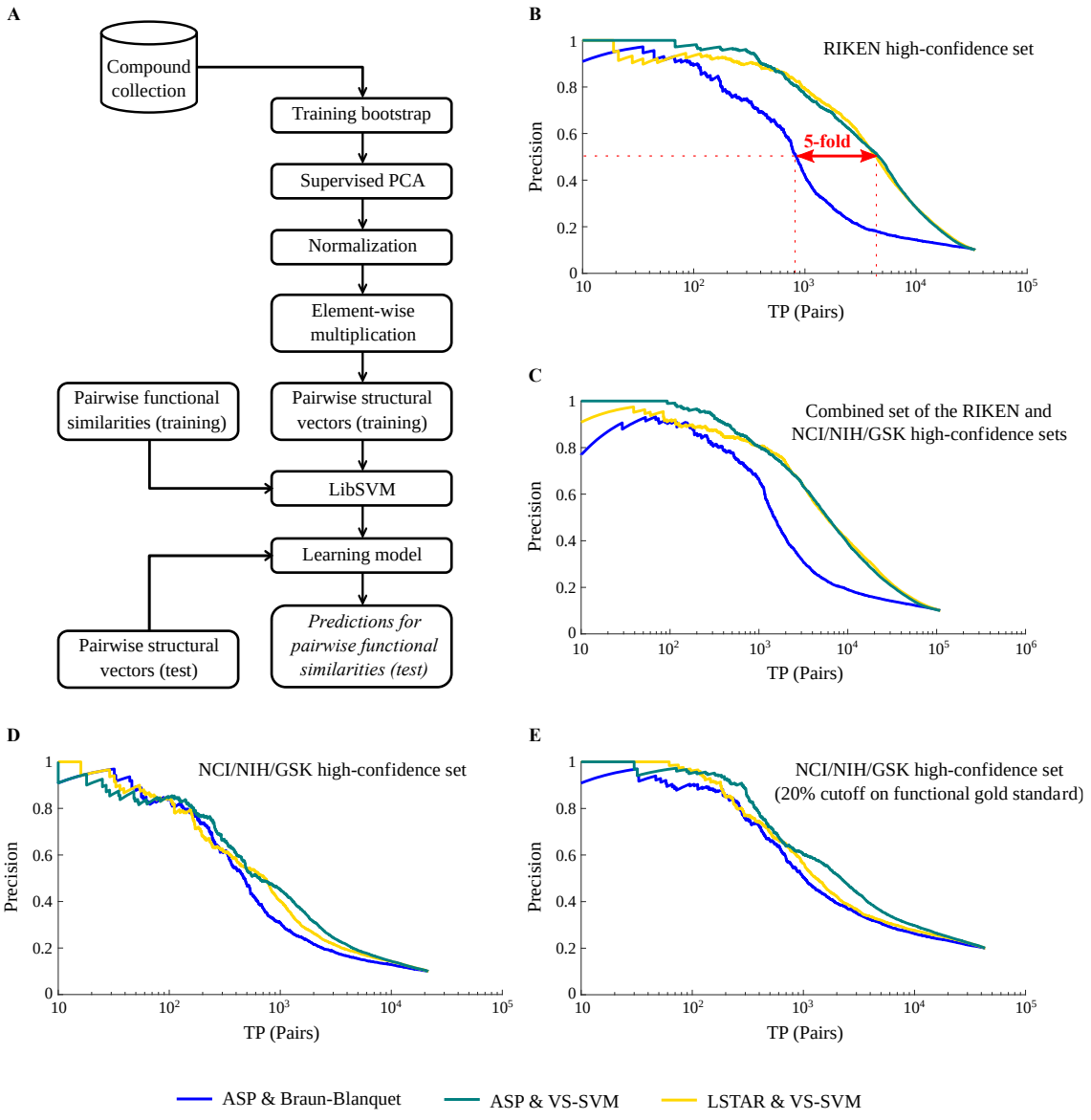


Figure 2.5. Prediction performance of machine learning models. **(A)** Learning pipeline for one bootstrap using pairwise structural vectors (Materials and methods). **(B)** Model performance for our RIKEN high-confidence set. The blue PR curve represents the prediction performance of our best structural similarity measure (ASP/Braun-Blanquet), whereas the teal and gold PR curves represent the performance of our machine learning models using ASP and LSTAR fingerprints, respectively. A prediction is considered a true positive if the compound pair is within the top 10%

of functionally similar compound pairs using chemical-genetic interaction profiles. We used pairwise true positives or TP (pairs) as a general form of recall in our PR curves. **(C)** Model performance for the combined RIKEN and NCI/NIH/GSK high-confidence sets. **(D)** Model performance for the NCI/NIH/GSK high-confidence set. **(E)** Model performance for the NCI/NIH/GSK high-confidence set (as in D), except using top 20% of pairwise chemical-genetic similarities to define true positives.

The PR curves of our learning models built on ASP and LSTAR fingerprints (**Figure 2.5B**) exhibited a fivefold improvement in the recall of biologically similar compounds at a precision of 50%. However, the degree of improvement was dependent on the functional diversity of our datasets, where performance was higher for the collections with lower functional diversity. For instance, predictions for the NCI/NIH/GSK high-confidence set improved by only about 2-fold based on the recall of biologically similar compounds at a precision of 50% (**Figure 2.5D–E**). We speculated that this relatively modest improvement (compared to that of the RIKEN high-confidence set) may be explained by the higher functional diversity of the compounds in our NCI/NIH/GSK high-confidence set. Specifically, we derived a functional diversity score based on a divisive clustering approach and observed scores of ~25.3 and ~14.6 for the NCI/NIH/GSK and RIKEN high-confidence sets, respectively, using the chemical-genetic interaction profiles of compounds (Materials and methods). This difference in functional diversity contrasted with the near-identical structural diversity observed between the two collections (score of ~62 using ASP molecular fingerprints). The relatively higher functional diversity of our NCI/NIH/GSK high-confidence set arises from its 6 functionally different sub-collections,

which consequently affected the ability of our learning pipeline to learn models that capture chemical-genetic similarities and generalize across many compounds in this collection. Although the performance of our supervised learning models for the NCI/NIH/GSK high-confidence set was affected by high functional diversity, even for this collection, the boost in performance offered by our machine learning models was still substantial (**Figure 2.5D–E**). Furthermore, we combined the two collections, which added not only more compounds but also more diversity to the resulting set. Making predictions for the combined dataset, we achieved about a 4.5–fold improvement in the recall of biologically similar compounds at a precision of 50% (**Figure 2.5C**).

We conclude that compound collections with low functional diversity substantially benefit from our machine learning models. These models can more appropriately adjust the influence of specific substructural features within large clusters of structurally similar compounds given an independent standard of shared biological function.

2.3.5 Exploring the basis of the predictive power of SVM models

To investigate the compounds driving our prediction models, we clustered our compound collections into 10 functional and 10 structural clusters using K-means and K-medoids clustering algorithms, respectively (**Table A.3**, **Table A.4**, and **Table A.5**). We then mapped the 1000 true positive pairs at the top of our PR curves for ASP fingerprints to their corresponding functional and structural clusters (**Figure 2.6** and **Figure 2.7**).

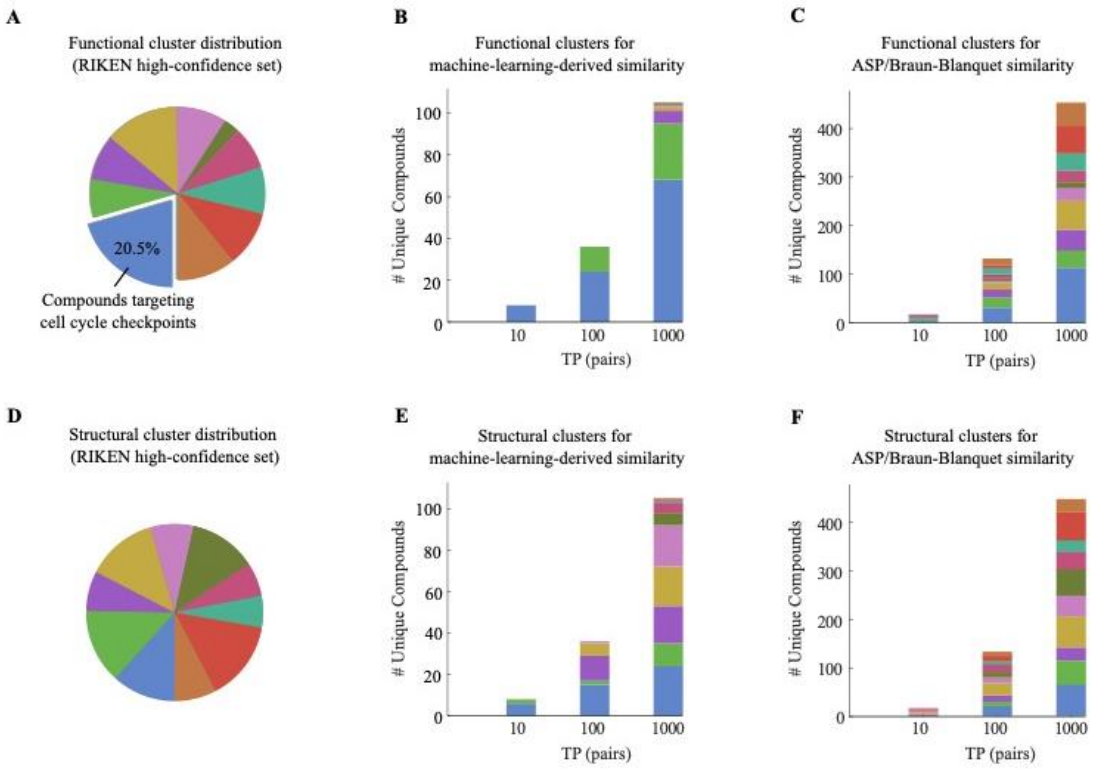


Figure 2.6. Functional and structural clusters of top true positive pairs for our RIKEN high-confidence set. **(A)** Distribution of 10 functional clusters generated by the K-means clustering algorithm using our chemical-genetic interaction profiles. The blue cluster represents the largest functional cluster. **(B–C)** Contribution of these functional clusters to the top true positive pairs retrieved by **(B)** our machine learning model and **(C)** our best structural similarity measure (ASP/Braun-Blanquet). **(D)** Distribution of 10 structural clusters generated by the K-medoids clustering algorithm using ASP fingerprints. **(E–F)** Contribution of these structural clusters to the top true positive pairs introduced by **(E)** our machine learning model and **(F)** our best structural similarity measure.

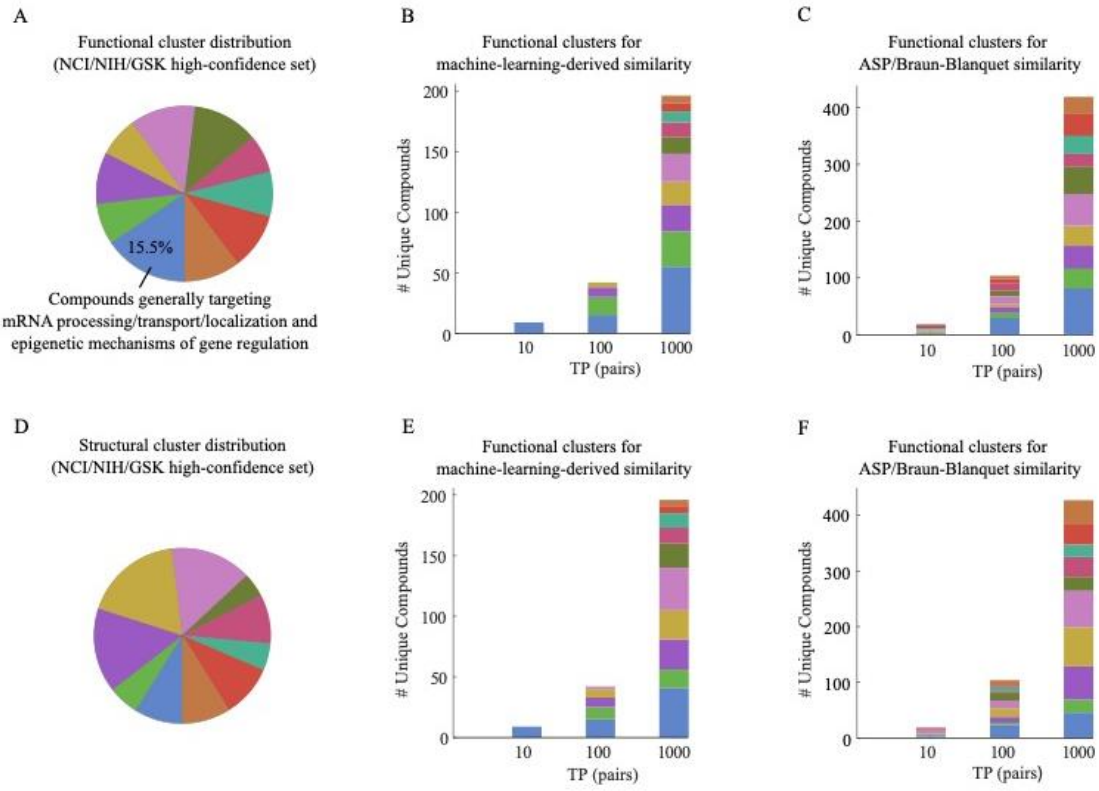


Figure 2.7. Functional and structural clusters of top true positive pairs for our NCI/NIH/GSK high-confidence set. **(A)** Distribution of 10 functional clusters generated by the K-means clustering algorithm using our chemical-genetic interaction profiles. **(B–C)** Contribution of these functional clusters to the top true positive pairs retrieved by **(B)** our machine learning model and **(C)** our best structural similarity measure (ASP/Braun-Blanquet). **(D)** Distribution of 10 structural clusters generated by the K-medoids clustering algorithm using ASP structural profiles. **(E–F)** Contribution of these structural clusters to the top true positive pairs introduced by **(E)** our machine learning model and **(F)** our best structural similarity measure.

A large group of the compounds generating high predictive scores at the top of our learning PR curves belonged to the same functional clusters (**Figure 2.6B** and **Figure 2.7B**), while the baseline PR curve for the ASP fingerprints and Braun-Blanquet similarity

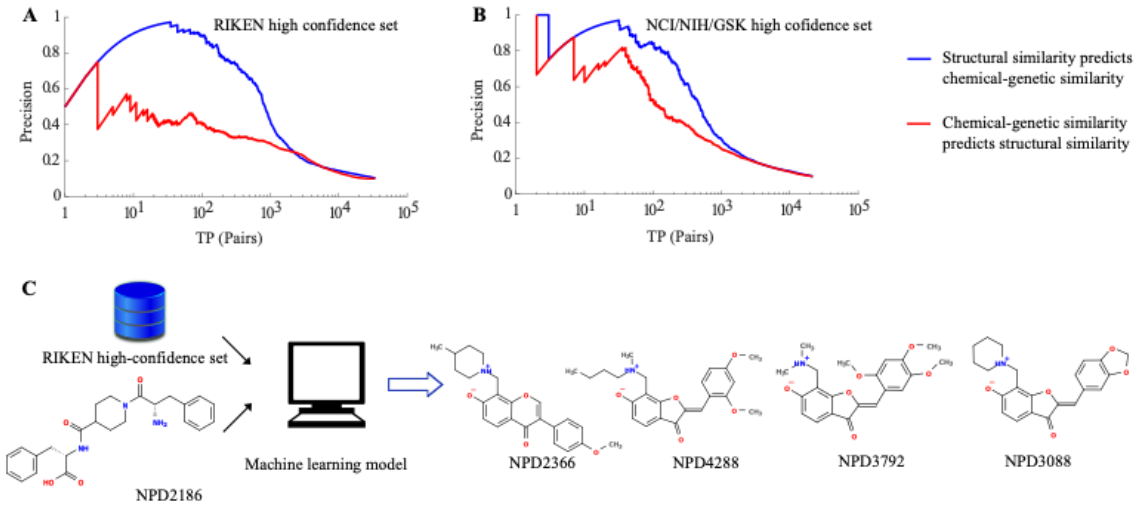
coefficient included several functional clusters even at lower recall thresholds (**Figure 2.6C** and **Figure 2.7C**). This indicates that our machine learning-based models improved performance in a non-uniform fashion, boosting prediction performance substantially for a few specific functional clusters. The first functional cluster that we observed among the highest predictions for the RIKEN high-confidence set (blue cluster in **Figure 2.6A** named Cluster 1 in **Table A.3**) was enriched for compounds targeting cell cycle checkpoint processes (**Table A.3** and **Table A.4** for the enrichment analysis of the functional clusters in the RIKEN and NCI/NIH/GSK high-confidence sets; predictions from the MOSAIC database (76)). Despite this enrichment for the predicted target processes, several distinct structural classes were clearly represented among the compound pairs with the highest similarity predicted by our machine learning models (**Figure 2.6E**), suggesting that compounds considered diverse based on molecular structures alone were predicted to have similar biological activity using our machine learning models (**Table A.6**).

2.3.6 Example use case: retrieving novel functionally analogous compounds for virtual screening applications

To demonstrate the impact of our machine-leaning-based approach for predicting biologically similar compounds, we illustrate with a specific example. We chose a query compound of interest from our RIKEN high-confidence set, NPD2186. We first ranked all other compounds in this collection based on their structural similarity to this compound as measured by our best-performing structural similarity measure (i.e., ASP/Braun-Blanquet) (**Figure 2.1**). To contrast, we also compiled a ranking based on our machine-leaning-

derived structural similarity scores computed from ASP fingerprints (**Figure 2.8C**). For example, among the top 10 most similar compounds identified by our machine-leaning-derived similarity, none were identified in the top 10 by the ASP/Braun-Blanquet similarity measure (or any of the other structural similarity measures). Among the top 100 and top 200 most similar compounds identified by our machine-leaning-derived similarity, only 29 and 88 were also identified in the top 100 and top 200 most similar compounds retrieved by the ASP/Braun-Blanquet similarity measure, respectively. Overall, the Spearman's rank correlation between the two recalled lists of chemical compounds for our query compound, NPD2186, was 0.31. Therefore, our machine-leaning-based structural similarity approach provided a substantially different set of similar compounds for this example.

The trends that we observed for NPD2186 generalize to the broader collection of compounds as well (**Table A.7**). We measured the Spearman's rank correlation of our machine-leaning-derived and Braun-Blanquet structural similarities, both computed from ASP fingerprints, for every compound in the RIKEN high-confidence set (**Figure A.6A**). The median of the resulting distribution of rank-based correlations was 0.31, suggesting that these two approaches produce substantially different rankings on the most similar compounds. Thus, using our proposed machine-leaning-based approach will likely produce substantially different results than the traditional measures of structural similarity in virtual screening settings.



Functional analogue for NPD2186	Chemical-genetic similarity	Structural similarity (ASP/Braun-Blanquet)	Predicted similarity (machine-learning-derived)	Rank of the compound pair in the predicted rank list using VS-SVM (Table S7)
NPD2366	0.894	0.065	0.716	1
NPD4288	0.401	0.061	0.700	2
NPD3792	0.822	0.048	0.685	6
NPD3088	0.832	0.068	0.676	8

Figure 2.8. Reciprocal evaluation of prediction performance of structural vs. functional similarity and machine-learning-based virtual screening of a target (e.g., NPD2186 from the RIKEN high-confidence set). Using the **(A)** RIKEN and **(B)** NCI/NIH/GSK high-confidence sets, we measured the abilities of structural and chemical-genetic similarities to reciprocally predict each other. The blue curve represents the performance of structural similarity in predicting chemical-genetic similarity, whereas the red curve represents the performance of chemical-genetic similarity in predicting structural similarity. **(C)** Our machine learning model retrieved biologically similar but structurally dissimilar compounds (determined by the ASP/Braun-Blanquet structural similarity measure) for NPD2186 from our RIKEN high-confidence set. The information table provides the chemical-genetic similarities, ASP/Braun-Blanquet structural similarities, and machine-leaning-derived predicted similarities for a few of the compounds at the top of the predicted ranked list that are functionally analogous to NPD2186. The highest predictive score generated by our machine

learning model was 0.716, retrieving NPD2366 as a functional analog of NPD2186. The rank of each compound pair comes from the table of all pairwise compound similarities ranked in the descending order of predicted machine-learning-derived similarities (**Table A.7**).

2.3.7 Evaluating the reciprocal predictive power of structural and functional similarities

The main focus of our study was to improve structural similarity measures by leveraging functional information from chemical-genetic interaction data. However, these data also offered an opportunity to explore the connection between structural similarity and similarity of functional impact on cellular function. Specifically, we were interested in the question of which was a stronger predictor: structural similarity predicting functional similarity, or functional similarity predicting structural similarity. To address this question, we carried out a reciprocal precision-recall analysis in which we switched the roles of structural and functional similarities and compared the relative strength of each for predicting the other. Specifically, we assessed the predictive power of the ASP/Braun-Blanquet-derived structural similarity against a standard of chemical-genetic similarity as described above. Then, we reversed the analysis, evaluated the potential of chemical-genetic similarity to predict the ASP/Braun-Blanquet-derived structural similarity, and compared the resulting performances. This analysis revealed that structural similarity exhibited substantially higher power in predicting chemical-genetic similarity than did chemical-genetic similarity for predicting similar structures (**Figure A.8A–B**). This result likely reflects the fact that compounds with very different chemical structures can result in a similar functional impact (e.g., targeting different subunits of the same protein complex

or different members of the same pathway will result in a similar phenotype). On the other hand, compounds with very similar structures are highly likely to exhibit similar biological activity and unlikely to possess extremely divergent biological activities. Interestingly, the gap between structural prediction of function vs. functional prediction of structure was substantially larger for the compounds in the RIKEN high-confidence set (**Figure A.8A–B**). This wider gap likely reflects the fact that the RIKEN high-confidence set contains many more local clusters of chemical compounds with highly related structures, whereas the NCI/NIH/GSK high-confidence set is inherently composed of several sub-collections with relatively few structurally related compounds, which will result in a more one-to-one correspondence between structural and functional profiles.

2.4 Conclusion

The drug discovery process benefits from improvements in our ability to link the structure of chemical compounds to their biological function. The chemical informatics community has proposed a wide range of molecular fingerprints and similarity coefficients for ligand-based virtual screening, where SPP has been the basis for ranking compounds with similar biological activity to a target molecule based on chemical structures. However, to date, the research community has lacked a systematic benchmark for biological activity that covers a broad range of protein targets to assess the performance of different molecular fingerprints and similarity coefficients. We used a large set of chemical-genetic interaction data from *S. cerevisiae* that we previously published, covering 13,431 compounds from the RIKEN NPDepo and several NCI/NIH/GSK compound collections as the standard for the

biological activity of our compounds. Using these chemical-genetic interaction profiles as a functional standard for our compounds, we systematically benchmarked 11 different molecular fingerprints and 13 different similarity coefficients.

We found that the ASP fingerprints combined with the Braun-Blanquet similarity coefficient was the superior structural similarity measure for prioritizing compounds with similar biological activity to a target molecule. The ASP fingerprints encode all shortest paths between atoms obtained through an exhaustive depth-first search of a compound's molecular graph (up to a predefined describing depth), and the Braun-Blanquet coefficient represents the fraction of features in common between two molecules to the total number of features presented in the larger molecule. We determined that the performance of our ASP/Braun-Blanquet structural similarity measure would not substantially benefit from the high describing depths of ASP fingerprints beyond a depth of 8. Our results suggest that the ASP/Braun-Blanquet similarity measure can be used with confidence as a replacement for the ECFP/Tanimoto similarity measure, which has been one of the most commonly used structural similarity measures for ligand-based virtual screening.

Moreover, we developed a machine learning model based on support vector machines that boosted the predictive power of several fingerprints up to fivefold, with the degree of improvement dependent on the functional diversity of the compound collections on which the machine learning model was applied. The compound collections with low functional diversity highly benefited from our machine learning model, where our machine-leaning-based approach generated a ranked list of functionally analogous

compounds that was entirely different from the ranked list generated by the ASP/Braun-Blanquet structural similarity measure for a target molecule.

One interesting future direction of our work would be to extend this analysis beyond the molecular fingerprints used in this study (e.g., considering computed physicochemical properties). Chemical-genetic interaction profiles may also yield important information on the relevance of other features for predicting functional similarity of compounds.

Overall, our high-dimensional chemical-genetic interaction data provide a powerful resource for connecting compound structure to function. We expect that the specific lessons learned here regarding the relative strengths of different molecular fingerprints, as well as the general approach of integration of the chemical-genetic interaction data with structural information for improving the structure-based prediction of biological activity, will be of use in other virtual screening contexts.

2.5 Materials and methods

2.5.1 Data collections

We previously published chemical-genetic interaction profiles for 13,431 uniquely-named compounds from several diverse compound collections and isolated a subset of these compounds that exhibited high-confidence predictions of perturbed biological processes based on integration of these profiles with genetic interaction profiles (73). We used two independent compound collections for which we had high-confidence biological processes: the RIKEN high-confidence set (826 compounds) and the NCI/NIH/GSK high-confidence set (659 compounds). The RIKEN high-confidence set is a subset of the RIKEN

Natural Product Depository (NPDepo) collection and is composed largely of purified natural products or natural product derivatives. The NCI/NIH/GSK high-confidence set is itself a diverse set of several sub-collections: four compound collections from the National Cancer Institute's Open Chemical Repository (natural products, approved oncology drugs, and structural and mechanistic diversity sets), one compound collection from the National Institutes of Health Small Molecule Repository with a history of use in human clinical trials, and the GSK kinase inhibitor compound collection. More detailed descriptions of these two compound collections are available in the "Methods – Description of compound collections" section of our recent publication (73).

2.5.2 Structural description of compounds

We used the jCompoundMapper tool (128) and RDKit toolkit (version 2020.09.4) (129) to describe our compounds in 11 different molecular fingerprint spaces (**Table 2.1**) using the structural information of our compounds (Supporting Structure Data Files and Supporting SMILES and InChI). Except for MACCS and RDKit molecular fingerprints, we generated all other molecular fingerprints using the jCompoundMapper tool. These molecular structural descriptions covered several topological aspects of a molecular graph, including depth-first search fingerprints, atom pair fingerprints, radial atom environment fingerprints, extended connectivity fingerprints, and pharmacophore fingerprints.

2.5.3 Establishing the gold standard for biological activity

We selected the 10% most similar compound pairs in chemical-genetic interaction profiles and labeled those pairs as our gold standard for true positive compound pairs. We previously published a systematic evaluation of several different profile similarity measures for genetic interaction networks (163) and more recently published our functional annotations of chemical libraries across diverse biological processes (73). These analyses suggested that the cosine similarity coefficient was a reasonable metric for measuring genetic and chemical-genetic similarities. Therefore, we used the cosine coefficient for computing our gold standard biological activity. Specifically, in the latter study, we annotated our RIKEN and NCI/NIH/GSK high-confidence sets to 17 different functional neighborhoods (e.g., DNA replication and repair, glycosylation, vesicle traffic, mitosis and chromosome segregation, etc.) using chemical-genetic interaction profiles (Supplementary Dataset 19 of our aforementioned study) (73). At a range of different cutoffs on the compound functional similarity score (2%, 5%, and 10%), we computed the distribution of our compounds selected across the 17 different neighborhoods. This analysis demonstrated that, at a cutoff of 2%, we observed a strong enrichment for the compounds mapped to the neighborhood 4 (N4 in Figure S7), corresponding to mitosis and chromosome segregation. However, at more relaxed cutoffs (e.g., 5% or 10%), the distribution of compounds across neighborhoods was relatively similar to the background distribution of our compounds, suggesting that we covered a broader representation of the compounds. This analytical method led us to selecting a higher cutoff (10%) on the pairwise functional similarity score to ensure a broad representation of the function of our compounds.

We need to note that the relative performance of our structural similarity measures and machine learning models for more stringent cutoffs (e.g., 5%) would remain consistent with our results for the top 10% of the functional similarity gold standard (**Figure A.8**).

2.5.4 Designing the support vector machine learning pipeline

We proposed SVR models (126,127) for predicting the functional similarity of two compounds based on their chemical structures. We used LibSVM (164), a popular open source support vector machine library developed at National Taiwan University, for implementing our models, and bootstrapping (162) for generating our training and test data. To generate our training and test sets, we randomly drew N (the total number of compounds in a collection) samples with replacement uniformly from a collection, which assigned ~0.632N unique compounds to the training set and the rest to the test set. To reduce the dimensionality of our structural spaces, we employed supervised principal component analysis (161) using chemical-genetic interaction profiles (from the training data only) as labels. We normalized each structural vector that described a compound in the low-dimensional space by its Euclidean length and multiplied each pair of the normalized vectors (both from the training set or both from the test set) in an element-wise manner to create a new space of structural vectors, called "pairwise structural vectors" (**Figure A.4A**), for the representation of compound pairs. We devised a machine learning pipeline (**Figure 2.5A**) to predict chemical-genetic similarities for the test data using pairwise structural vectors and pairwise chemical-genetic similarities (from the training data only). We used RBF kernels to build epsilon support vector regression (ϵ-SVR) models in the pairwise

structural space and evaluated the average performance of our pipeline across 100-200 bootstraps (**Figure A.4B**). To measure the prediction performance of the pipeline on held-out test data, we used the average model output over all the bootstraps for which a given example was in the test set (bagging). We sampled 200 bootstraps for the RIKEN high-confidence set, 200 bootstraps for the NCI/NIH/GSK high-confidence set, and 100 bootstraps for the combined collection of these two sets. We evaluated the prediction performance of our machine learning pipeline using PR curves (**Figure 2.5B–E**). Selecting the 10% most similar compound pairs based on the (cosine) similarity of their chemical-genetic interaction profiles and labeling those pairs as our gold standard for true positive compound pairs, we translated our predicted pairwise compound similarities into a binary evaluation using PR curves after we ranked our predicted similarities in the descending order against our binarized gold standard for true positive compound pairs (**Figure A.7**, where the Class column indicates our binarized gold standard for functional similarity).

To facilitate the use of our supervised similarity prediction pipeline, we provide the current version of our MATLAB source codes, which can be used for learning a new compound collection. The source codes and our data for the machine learning prediction pipeline are freely available at https://github.com/csbio/VS-SVM.

2.5.5 Estimating the diversity of compound collections

We defined functional and structural diversity measures for compound collections from the results of a recursive, divisive clustering algorithm based on either functional (chemical-genetic) or structural similarity, respectively. More specifically, we assigned all

the compounds in a collection to a single cluster and split up the cluster recursively to form smaller clusters of more similar compounds. At any step of recursion, we determined the cluster with the lowest average within-cluster pairwise chemical-genetic similarity (for computing functional diversity) or structural similarity (for computing structural diversity) and divided the cluster into two new clusters using the K-means (K=2) or K-medoids (K=2) clustering algorithms, respectively. We terminated the recursion when at least two new clusters would exceed our predefined hard threshold for the maximum average between-cluster pairwise chemical-genetic similarity (cosine chemical-genetic similarity of 0.3) or structural similarity (ASP/Braun-Blanquet similarity of 0.3). We repeated this algorithm 1000 times for functional diversity and 100 times for structural diversity. We computed the mean diversity score as the average exponentiation of the Shannon entropy indices over all instances: $\mathrm{D} = \mathrm{mean}\left(2^{(-\sum_i p_i \log_2(p_i))}\right)$, where p_i represented the proportional abundance of the compounds in the i^{th} cluster of the final clustering.

3 Target Identification of Bioactive Compounds Using a High-Resolution, Integrated Chemical-Genetic Approach

This research work is the continuation of the work described in the research article "Functional annotation of chemical libraries across diverse biological processes", which we published at Nature Chemical Biology, Volume 13, Issue 9, 2017. This thesis chapter represents an initial draft of a manuscript that is in preparation for journal submission at the time of this writing. Other contributors to this research work are Sheena C. Li, Mami Yoshimura, Luis A. Vega Isuhuaylas, Chad L. Myers, Yoko Yashiroda, Minoru Yoshida, and Charles Boone.

S.C.L., H.S. (this dissertation's author), C.L.M., Y.Y., Minoru Yoshida, and C.B. conceived this project and established the methodologies. S.C.L. developed experimental hypotheses and validated the results including compound–gene targets. S.C.L. and Mami Yoshimura performed chemical-genetic interaction screens and validation experiments. H.S. designed all computational frameworks for data analyses presented in this chapter. L.A.V.I. refined data analytical methods and provided additional computational support for the manuscript in preparation. C.L.M., Y.Y., Minoru Yoshida, and C.B. supervised the project and acquired funding. S.C.L. and H.S. wrote the initial drafts of the manuscript to be reviewed by other authors.

Some of the presented texts, figures (**Figure 3.3** and **Figure 3.4**), and tables (**Table 3.1**) were prepared by S.C.L as part of the draft for the manuscript in preparation. The discussions on the gene targets of RP-3-161 and NPE71 were completed as part of this

research effort but were published separately in R. Persaud, *et al.* Clionamines stimulate autophagy, inhibit Mycobacterium tuberculosis survival in macrophages, and target Pik1. Cell Chemical Biology, 29, 870-882.e11 (2022).

3.1 Abstract

Identifying the targets of chemical compounds is crucial for finding new drug leads and tool compounds but remains exceedingly difficult. Genome-scale chemical-genetic analysis monitors the effects of bioactive compounds on the fitness of specific mutants and provides a systematic method for determining the molecular mechanism of a compound. We developed a screening pipeline for high-resolution target identification of compounds in *Saccharomyces cerevisiae.* We constructed barcoded, drug-sensitized mutant libraries, which enabled pooled microculture experiments for measuring the mutant fitness in response to compound exposure and the rapid assembly of chemical-genetic interactions using next generation sequencing. We generated nearly 5,000 chemical-genetic interaction profiles across four essential and non-essential gene mutant collections and prioritized target predictions using a quantitative measure that combined information across these complementary chemical-genetic screens. A subset of the predicted targets was confirmed using a sequence analysis of spontaneously generated resistant mutants, chemical-genetic interaction analysis in a CRISPR-modified human cell line, or biochemical validation of target pathways. Overall, we identified several chemical compounds with therapeutically relevant cellular targets.

3.2 Introduction

Currently, a glut of compound libraries that are awaiting functional characterization and vying for relevance exist. The ability to provide functional information about these compound libraries and elucidating the direct targets of bioactive compounds is vital for identification of potential drug leads and chemical probes. Despite large efforts invested in drug discovery, relatively few new drugs with novel mechanisms of action have been developed (165). Commonly used discovery approaches involve screening for the agents that bind known protein targets or inhibit disease-related pathways, which leads to a bias towards discovery of additional molecules for the same targets and a scarcity of the agents with novel mechanisms of action (165). In this study, we present an unbiased, systematic approach for target identification of compounds based on a whole-cell screening approach that uses chemical-genetic interaction methods. Such approaches are valuable because they identify bioavailable molecules and maintain biological contexts, compared to *in vitro* assays that may reveal direct protein–inhibitor relationships but do not consider the cellular uptake, the presence or absence of important cofactors, or the transient interactions that may allosterically modulate the intracellular activity of an inhibitor.

A chemical-genetic interaction profile annotates systematic genetic perturbations that lead to increased sensitivity or resistance upon exposure to a bioactive compound of interest and provides physiological information about the mode of action of the compound (67). *S. cerevisiae* is an ideal model system for probing chemical-genetic interactions because this model organism is a powerhouse for genetic analysis with a plethora of functional genomics tools. Genome-wide gene deletion analysis revealed ~1,000 essential

genes required for viability and led to the generation of ~5,000 non-essential haploid deletion mutants (121). These mutants enabled synthetic genetic analysis on a global scale, where a query mutation is crossed into a collection of mutants to identify inviable double mutants (65). The genes that belong to the same cellular pathway share similar genetic interactions. When genes are placed on a network based on the genetic interaction profile similarity, the relative locations of the genes correspond to the "biological neighborhoods" enriched for specific cellular processes (73). Because a mutation that compromises the function of a gene provides a model for a compound that inhibits the function of the gene's product, the synthetic lethal profile of the target of a compound should match the chemical-genetic interaction profile of the compound (67,73). In other words, genetic interactions are functionally analogous to chemical-genetic interactions for a compound, and therefore the global genetic interaction map can serve as a key for linking compounds to their targets. For example, the chemical-genetic interaction profile of Tunicamycin, a drug that blocks asparagine-linked glycosylation by directly inhibiting ALG7, which transfers Glc-Nac-P to Dol-P in the first step of the dolichol pathway, highly resembles the genetic interaction profiles of the genes that are part of the glycosylation neighborhood of the network.

Essential genes can be probed in chemical genomics approaches as heterozygous diploid (HET), conditional temperature sensitive (TS), or overexpression mutants, while non-essential genes can be evaluated as haploid or homozygous gene deletion mutants. The chemical-genetic interactions involving essential genes are particularly important because compounds that are capable of inhibiting growth are likely to have targets that are encoded by essential genes, which also tend to be highly conserved across diverse species (95). The

direct target of a compound can be revealed by interactions with essential gene mutants at altered gene dosage. Compound hypersensitivity due to reduced gene dosage in a diploid with a heterozygous deletion, or resistance to a compound due to increased gene dosage in an overexpression strain, can point to a direct compound–target pair. In fact, essential genes are densely connected hubs in the genetic interaction network and provide rich information about the functional relationships between the well-conserved processes that are absolutely required for life.

We developed a high-resolution framework for rapidly identifying the direct gene targets of bioactive compounds. We constructed several DNA-barcoded mutant collections in a drug hypersensitive genetic background that interrogated both the essential and non-essential genes in *S. cerevisiae*. We probed several large compound libraries with known structures, which were of high quality and were freely available. We used a parallel barcode sequencing protocol to construct high-resolution chemical-genetic interaction profiles for compounds of interest (73). DNA-barcoded mutants were grown in a pooled format; each pooled microculture was exposed to a unique compound. The sensitivity or resistance of mutants to compound treatments was quantified by multiplexed Illumina sequencing of mutant-specific barcodes. We generated 4,942 chemical-genetic interaction profiles from four different mutant collections spanning both essential and non-essential yeast genes. We developed a new quantitative measure that integrated complementary chemical-genetic information and prioritized prominent compound–gene pairs for further experimental analysis. We used this computational pipeline to systematically identify the direct gene targets for bioactive compounds from diverse compound libraries. Validation using drug

resistant mutant analysis, CRISPR chemical-genetic screens in a human HAP1 cell line, and biochemical assays led to the identification of several direct targets for compounds.

3.3 Target identification pipeline

To establish a high-resolution target identification approach, we constructed novel mutant collections for gene-level target identification of compounds in *S. cerevisiae*. Each mutant collection was DNA-barcoded and constructed in a drug hypersensitive genetic background, where a gene of interest was deleted for constructing the whole genome (WG) non-essential gene deletion collections, thermally perturbed for constructing TS essential gene collections, reduced in the copy number for constructing HET essential gene deletion collections, or overexpressed for constructing MoBY 2-micron essential gene collections. Chemical compound libraries were screened against these mutant collections in a pooled microculture format, and the relative abundance of a mutant was quantified by multiplexed barcode sequencing. We used computational and analytical methods for eliminating non-specific technical effects (75), quantifying chemical-genetic interactions, and identifying highly prioritized compound–target pairs.

3.3.1 Assembly of whole genome, temperature sensitive, heterozygous diploid, and MoBY overexpression profiles for compounds from diverse libraries

We screened several compound collections that contained natural products, their derivatives, and synthetic compounds. The RIKEN Natural Product Depository (NPDepo) is comprised of roughly ~30,000 natural products and their derivatives. In a previous study

(73), we developed a high throughput screening approach for generating chemical-genetic interaction profiles using a miniaturized collection of diagnostic yeast mutants (72,166). This was amenable for screening natural products that were available in limited amounts. We selected ~1,500 compounds from RIKEN NPDepo that had high-confidence biological process target predictions derived from our previous work (73) or had potent bioactivity. We also screened 268 highly bioactive chemical compounds from the University of Tokyo Drug Discovery Initiative Core Library, which is comprised of 9,600 diverse compounds, and 216 compounds from a large compound library of natural products from the University of British Columbia. In total, we generated 4,942 chemical-genetic interaction profiles (**Table 3.1**).

For each compound collection, we generated four different classes of data: 1. HET data: the heterozygous diploid gene deletion data, where we measured the sensitivity of a diploid cell to a compound by deleting one of the two copies of an essential gene (the data were collected for 973 essential genes and 1,793 compounds). 2. TS data: the temperature-sensitive chemical-genetic interaction data, where we measured the sensitivity of a haploid cell to the temperature-sensitive mutant of an essential gene (these data were collected for 1,154 essential genes and 1,764 compounds). 3. WG data: the whole genome chemical-genetic interaction profiles in a drug-sensitive genetic background that included all non-essential genetic mutants (these data were collected for 3,500 genes and 1,148 compounds). 4. MoBY data: the overexpressed chemical-genetic data, where we measured the sensitivity of a cell to a compound by overexpression (increasing dosage) of one of the two copies of an essential gene (these data were collected for 1,110 essential genes and 237 compounds).

We obtained HET, TS, and MoBY profiling data at multiple dosages, while the WG data was measured at only a single dose. To control the performance of our screening pools, we determined some gold-standard compounds to be included in each screen.

Table 3.1. Data generation using several diverse compound libraries. Each number represents the number of compounds from a compound library screened against a class of genetic mutants.

	HET	TS	MoBY	WG	Total
RIKEN NPDepo high-confidence set	922	908	54	880	2,764
RIKEN NPDepo bioactive set	387	387	N/A	N/A	774
University of Tokyo DDI core library	268	268	183	268	987
University of British Columbia natural products library	216	201	N/A	N/A	417
Total					4,942

3.3.2 Deprioritizing noisy data screens

We observed for some compounds that their chemical-genetic interaction profiles screened at a specific concentration inhibited growth of the yeast mutant pools by more than 95%. This pattern led to uninformative profiles with half of the interactions clustered close to a high chemical-genetic interaction score limit and the other half clustered close to a low chemical-genetic interaction score limit, leading to a “bimodal” distribution of chemical-genetic interaction scores for a compound. We used a computational bimodality test to identify these profiles and immediately filter such screens out of the dataset.

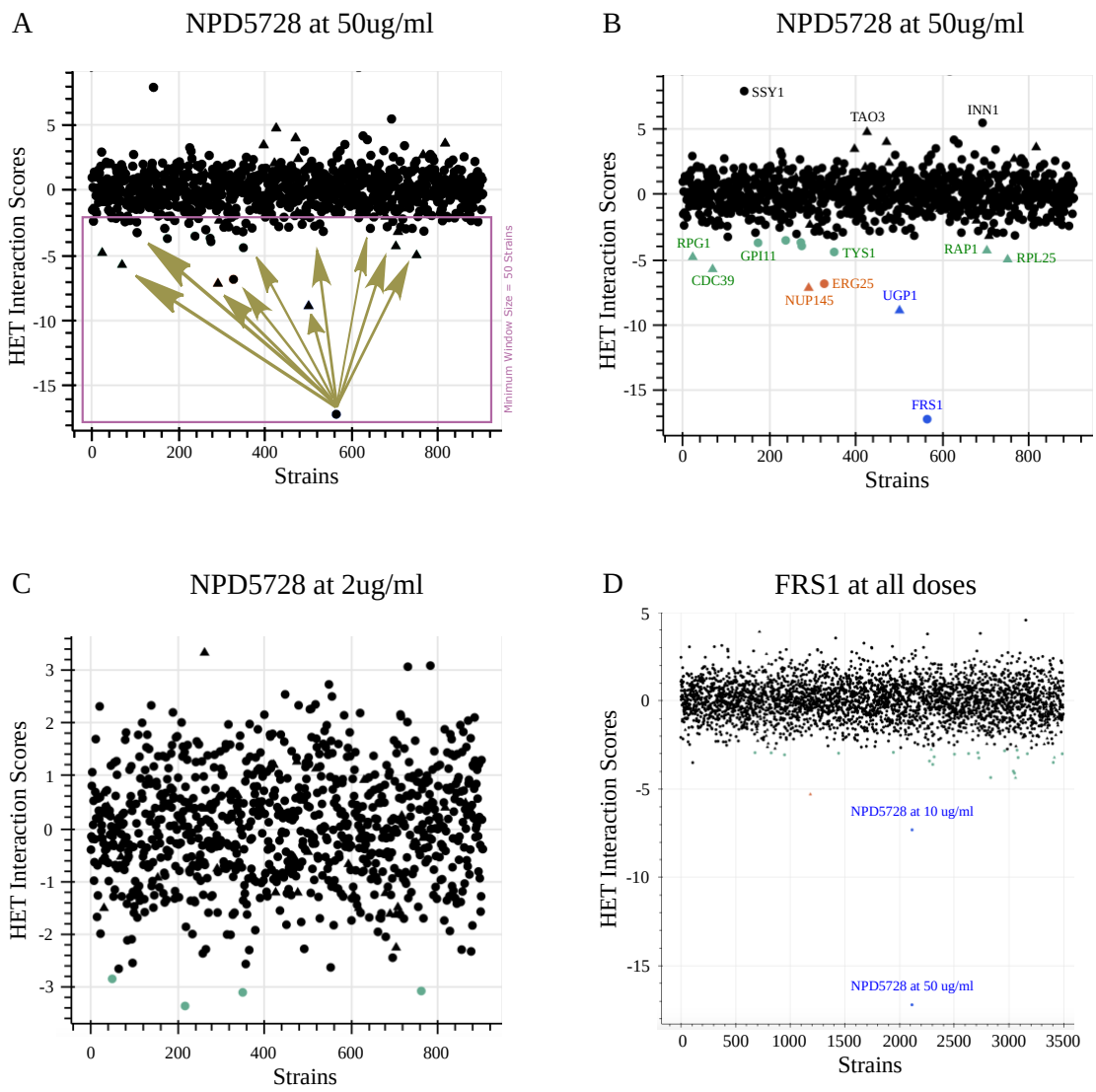


Figure 3.1. Removing noisy screens and demonstrating outlier analysis concepts. **(A)** Minimum window score for the NPD5728 HET screen at 50 ug/ml, where we measured the mean Euclidean distance of the mutant that had the minimum HET score with 50 (window size) other mutants that had the most negative HET scores. We computed a minimum window score of ~13.64 for this NPD5728 screen, while the cutoff for the removal of noisy screens was set at the minimum window score < 1.18. **(B)** Compound-wise outlier analysis for the NPD5728 HET screen at 50 ug/ml. We highlighted the mutants that fell into the outlier, strong outlier, and extreme outlier regions by green, orange, and blue colors, respectively. We displayed frequent flier genes using triangles. **(C)**

A HET screen for NPD5728 at 2 ug/ml with more noise than signals. Although we computed a low minimum window score of ~1.32 for this screen, the screen still survived in our analysis as its score was barely above the cutoff of 1.18. Such screens were removed by manual inspections around the predefined cutoff. **(D)** Gene-wise outlier analysis for FRS1 genetic mutants at all three doses of our HET screens. FRS1 mutants stood out in the drug-sensitive background caused by NPD5728 at 50 ug/ml and 10 ug/ml substantially, making FRS1 the potential gene target of NPD5728.

We developed a quantitative score, called "minimum window", for each screen's data to capture the level of noise primarily on the negative interaction end of the screen because negative chemical-genetic interactions contained most of the information for the annotation and target identification of compounds (73). We measured a minimum window score as the mean Euclidean distance of the mutant having the minimum chemical-genetic interaction with all other mutants falling into a predefined window of the mutants with low interaction values (**Figure 3.1A**). We predefined the size of this window to contain ~50 genetic mutants with low chemical-genetic interaction values. The screens that gained high minimum window scores demonstrated strong biological signals, whereas the screens with scores near zero were considered noisy data screens and were therefore removed from any further analysis.

Analogously, we developed a maximum window score for every chemical-genetic interaction screen to deprioritize or remove the screens with so much noise on the positive interaction end of the screen. The maximum window scores were particularly useful when we analyzed the effects of positive chemical-genetic interactions to gain more knowledge on a putative gene target inferred from the analyses of negative interactions.

3.3.3 Outlier analysis for prioritizing gene targets

We developed a computational framework that detected outliers at three different levels to identify the strongest chemical-genetic interactions in every screen. We used Tukey's test (167) to determine the outliers and measure the strength of an outlier in each screen based on the interquartile range of the distribution of the mutants. If Q_1 and Q_3 were the lower and the upper quartiles of a screen respectively, we defined an outlier to be any observation outside the following interval: $[Q_1 - K\Delta Q, Q_3 + K\Delta Q]$ where $\Delta Q = Q_3 - Q_1$ is the interquartile range of the screen and K is a non-negative constant. We called a mutant an "outlier" if the mutant fell outside the interval for K = 1.5 (**Figure 3.1B**). We extended this terminology to "strong outliers" and "extreme outliers" if a genetic mutant fell outside the interval for K = 3.0 and K = 4.0, respectively (**Figure 3.1B**). Since negative interactions contained the most information for predicting the gene target of a bioactive compound (73), we searched for three different levels of strength of outliers in $[Q_1 - K\Delta Q, 0)$ after we removed noisy screens using minimum window scores.

We developed our outlier-based analyses in two different senses: compound-wise and gene-wise. The compound-wise outlier analysis was centered around a compound to determine the genes that dropped in their interaction values in presence of the compound (**Figure 3.1B–C**). However, in the gene-wise outlier analysis, we observed interactions for every gene and measured in presence of what compounds the interaction corresponding to a gene dropped out (**Figure 3.1D**).

Since chemical-genetic interaction screens were measured for each compound (i.e., each pooled microculture was exposed to a unique compound), not a gene, we required that

the chemical-genetic interaction scores of each gene corresponding to different compounds and at different dosages to be normalized for the gene-wise outlier analysis. For the purpose of this normalization step, we computed the z-scores of the genes in every data distribution for a compound and used these z-scores as the normalized interaction scores of the genes for gene-wise outlier analyses. This data normalization ensured that the chemical-genetic interaction scores obtained for a compound were statistically interpretable at the gene level.

3.3.4 Detecting genes exhibiting frequent drop-outs across many compounds

We observed that a number of mutants were hypersensitive (having low interaction values) in exposure to many compounds in heterozygous gene deletion mutant collections. The genes associated to such genetic mutants were called “frequent flier” genes. Similarly, we defined frequent flier compounds as the ones to which many genes were hypersensitive. Because such HET interactions were unspecific at the gene or the compound level, or both, we used quantitative measures to define frequent fliers. We then deprioritized frequent flier genes and frequent flier compounds in our target identification analyses. We labeled a gene as a frequent flier if the percentage of the unique compounds that had the gene of interest as an outlier in the compound-wise sense exceeded a cutoff measured by the corresponding distribution (predefined cutoff = mean of the distribution + 2 standard deviations, **Figure 3.2A**). For example, we measured that POL2 and RAP1 genes were identified as outliers by ~30% and ~21% of the compounds in our RIKEN high-confidence set, making these two genes pleiotropic and highly unspecific for target identification purposes (**Table B.1**). Similarly, we classified a compound as a frequent flier if the percentage of the unique genes

that had the compound of interest as an outlier in the gene-wise outlier analysis exceeded a predefined cutoff determined by the corresponding distribution (predefined cutoff = mean of the distribution + 2 standard deviations, **Figure 3.2B**). A frequent flier compound was seen as a compound that targeted many genes and thus was not valuable for specific target identifications (**Table B.1**). Such a compound has a likelihood of introducing many side-effects, decreasing its potential value for drug discovery efforts.

3.3.5 Specificity analysis in gene target predictions

To identify specific gene targets for compounds, we developed compound-wise and gene-wise specificity scores. We measured, using our compound-wise outlier analysis, the number of more sensitive compounds (MSC), and using our gene-wise outlier analysis, the number of more sensitive genes (MSG) to promote the compound–gene target pairs highly specific in a particular pool. MSC referred to the total number of the unique compounds for which the z-score of the gene of interest in the compound-wise outlier analysis was less than or equal to the z-score of that gene for a given compound. This measure counted how many other compounds in a collection targeted a gene of interest with stronger z-scores in the compound-wise sense and hence controlled how specific a compound–gene interaction for a given compound in the compound-wise sense was.

Similarly, in the gene-wise outlier analysis, we defined MSG as the total number of unique genes for which the z-score of the compound of interest was less than or equal to the z-score of the same compound for a given gene. MSG represented a measure of the specificity of a gene against all compounds in a set using the gene-wise outlier analysis.

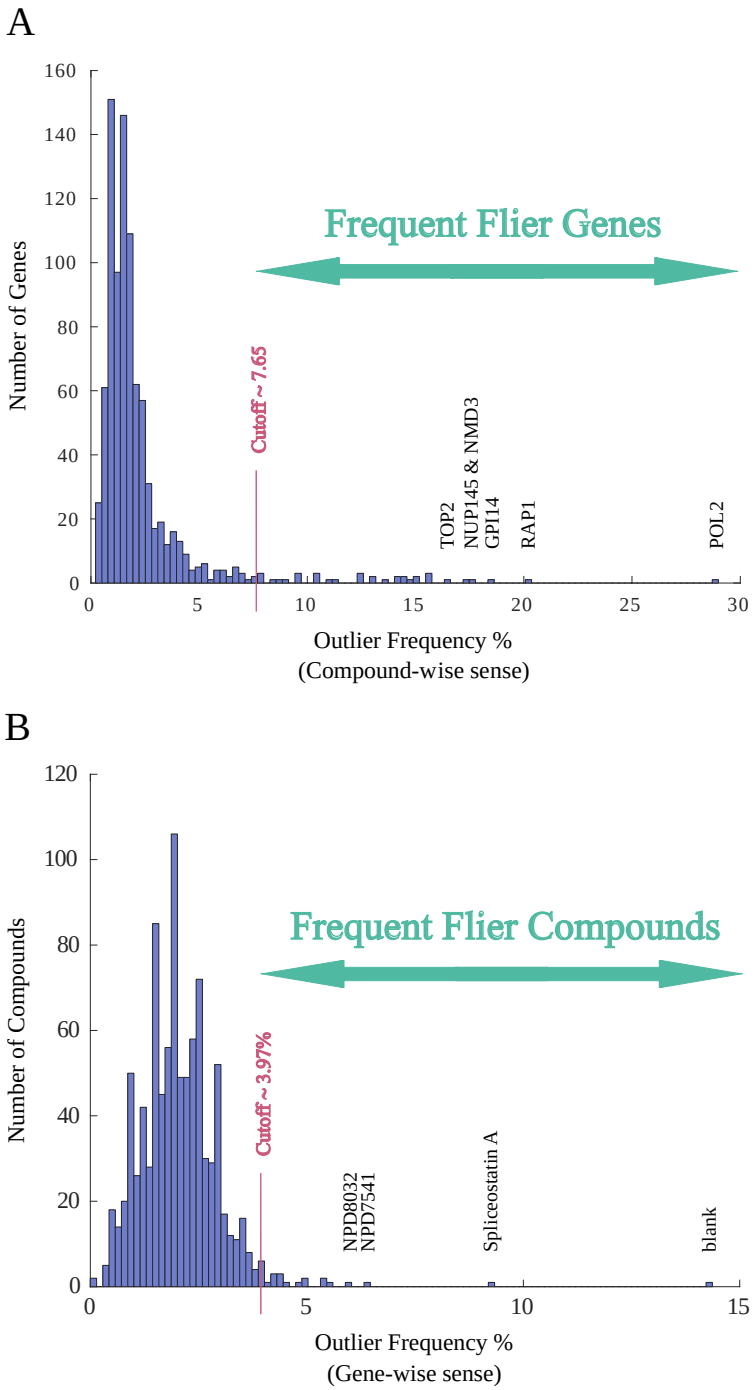


Figure 3.2. Frequent flier analysis. **(A)** Frequent flier genes for the HET data of our RIKEN high-confidence set. For every genetic mutant, we measured the percentage of the unique compounds that found the mutant in their outlier region of their compound-wise outlier analysis. We isolated

all the genes that exhibited outlier frequency > 7.65% (measured as mean + 2 standard deviations) in the compound-wise sense as frequent flier genes. **(B)** Frequent flier compounds for the HET data of our RIKEN high-confidence set. Similarly, for every compound, we measured the percentage of the unique genes that found the compound in their outlier region of their gene-wise outlier analysis. We labeled all the compounds that had outlier frequency > 3.97% (measured as mean + 2 standard deviations) in the gene-wise sense as frequent flier compounds.

3.3.6 Target scores for identifying specific, strong gene targets

We identified potential gene targets of compounds by integrating information from three distinct sources of chemical genomics. The first data source relied on comparisons between chemical-genetic and genetic interaction profiles (**Figure 3.3A** and **Figure 3.4**) using information from a diagnostic collection of non-essential mutants (**Figure 3.3B**) that we previously described (67,73) as well as novel drug hypersensitive mutant collections that spanned non-essential (WG data) and essential (TS data) genes (**Figure 3.3C**). From this analysis, statistically significant biological process target predictions were derived, and the processes required for buffering against the activity of a compound were identified.

The second chemical genomic approach utilized drug induced haploinsufficiency, where a heterozygous diploid mutant (HET data) that contained half of the normal copy of the gene target of a compound would become more sensitive to that compound compared to other heterozygous mutants (**Figure 3.3D-E**). This method reveals the precise essential gene target for a compound. For example, Tunicamycin induces haploinsufficiency in the diploids that have one copy of the *ALG7* gene, which encodes an enzyme that catalyzes the first step of asparagine-linked glycosylation and is directly inhibited by Tunicamycin (95).

We validated HET interactions using a library of the mutants that overexpressed essential yeast genes on a high copy plasmid, where overexpression of the gene target of a compound should rescue toxicity of that compound (**Figure 3.3G**).

The third source of chemical genomics was through identifying the genetic loci that conferred resistance against a drug (**Figure 3.3F**). Spontaneous resistant mutants against compounds are useful for target identification because resistant mutations often arise in the specific gene that encodes the target of a compound. For example, point mutations in *FKS1*, a glucan synthase inhibited by the antifungal drug Caspofungin, confer resistance to this drug (168).

We devised a ranking system to identify high priority gene target predictions for compounds by combining information from these three sources of chemical genomics. Our ranking system considered the specificity and strength of interactions as well as the overlap between known gene ontology (GO) annotations of the putative gene targets and predicted target bioprocesses. We computed a "target score" for every compound–gene pair using the information in compound-wise and gene-wise senses. We filtered and processed our raw data scores (see previous subsections) to generate a unified target score that combined all the chemical-genetic interaction information that reflected every aspect of gene target prediction for a compound.

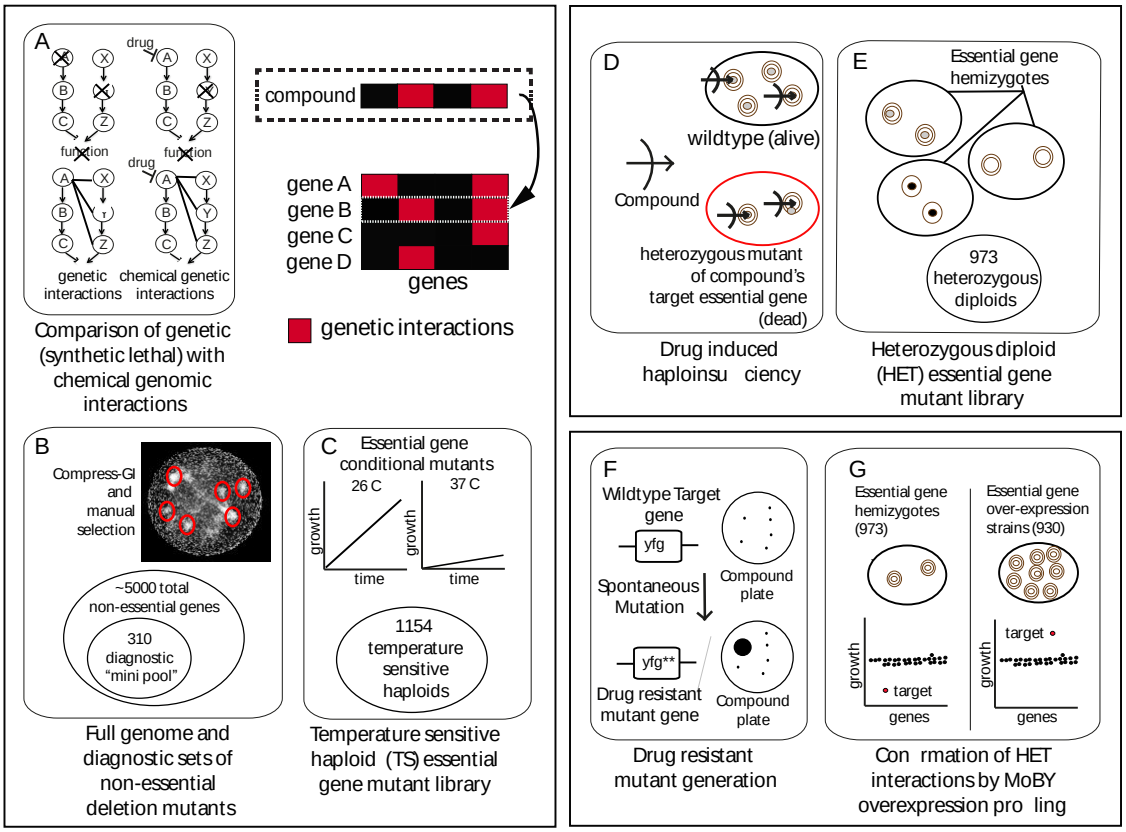


Figure 3.3. Integration of information from three distinct chemical genomic approaches for target identification of a bioactive compound. **(A)** Compound inhibition phenocopies a deletion of the gene target of a compound. The synthetic lethal profile of the gene target of a compound should match the chemical-genetic interaction profile of the compound. **(B)** Diagnostic or miniaturized mutant collection, which was selected using the Compress-GI algorithm (72,166) from the set of viable yeast gene deletion mutants and was predictive for all major biological processes in yeast. **(C)** The chemical-genetic interaction profiles for the TS data were constructed by screening a temperature-sensitive haploid library for essential genes to reveal target pathways. **(D, E)** Screening heterozygous mutants for compound-induced haploinsufficiency predicted precise gene targets. **(F)** Identifying the genetic loci, and **(G)** MoBY overexpression strains that conferred resistance to a compound confirmed the gene target prediction for the compound.

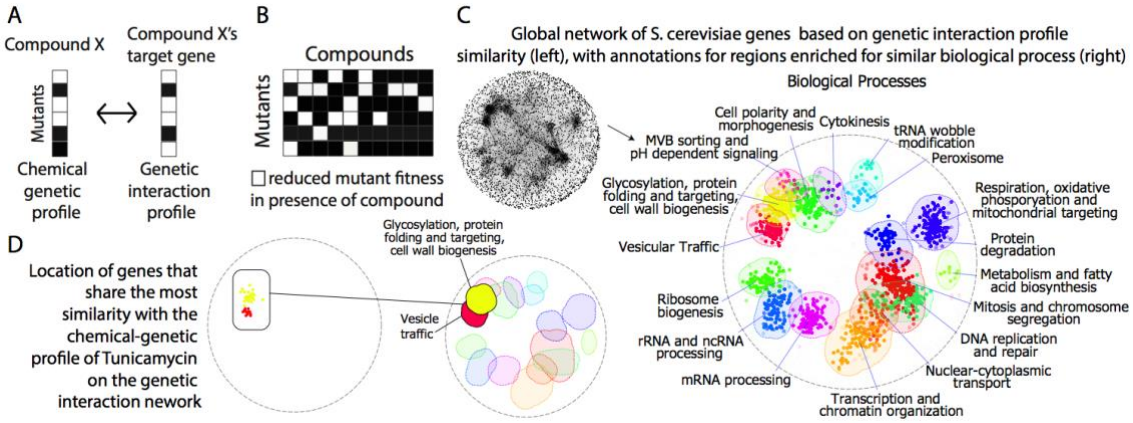


Figure 3.4. Comparison of chemical-genetic interaction profiles with genetic interaction profiles enabled compound–target identification. **(A)** The chemical-genetic interaction profile of compound X should be similar to the genetic interaction of the gene target of compound X. **(B)** Information from a matrix of chemical-genetic interaction profiles can be compared to **(C)** the global genetic interaction network of *S. cerevisiae* (adopted from (65)). **(D)** Genes with high profile similarity to Tunicamycin map to the glycosylation and protein folding neighborhoods of the genetic interaction network in *S. cerevisiae*.

Since large values of MSC and MSG indicated non-specificity of target predictions for a compound, we first filtered out all the interactions that exhibited MSC > a predefined cutoff in the compound-wise sense and MSG > a predefined cutoff in the gene-wise sense. This filtering step resulted in a shorter and more informative data table for every compound and every gene. For every compound, we then computed the overlap of the data table of the compound with the data tables of the genes from which the data table of the compound was constituted. If the compound did not exist in the data table of the gene, the gene was removed from the data table of the compound. We then integrated the data of the data tables

of the constituting genes into the data table of the compound. The result of this process was a smaller, prioritized data table for the compound, which hereafter we call the "top data table" of the compound, where both MSC and MSG were smaller than or equal to the predefined cutoffs. Similarly, we generated a "top data table" for every gene in the gene-wise sense, where both MSC and MSG measures were smaller than or equal to their predefined cutoffs. Using the cutoff = 20 for both MSC and MSG measures, we generated a top data table for every compound in the compound-wise and every gene in the gene-wise sense (**Table B.2**). We ranked the entities of a top data table based on the following gene target prediction criteria:

- **Specificity Score (SS):** SS = ((1 – (MSC – 1) / cutoff) + (1 – (MSG – 1) / cutoff)) / 2, where cutoff = 20. The specificity score for the frequent flier genes and frequent flier compounds equaled zero, by definition.

- **Height Score (HS):** HS = (|compound-wise IQR-score| / max (|compound-wise IQR-score|) + |gene-wise IQR-score| / max (|gene-wise IQR-score|)) / 2.

- **Reproducibility Score (RS):** We measured this score using a few distinct score levels {0, 0.5, 1} based on the number of different doses that we observed in a top data table for a given compound or gene. We broke down this score for a compound–gene pair observed in a top data table as follows:

If we performed data screening at three different doses, but only one dose stood out in the top data table, then RS = 0. If two or all the three doses of the interactions corresponding to a compound–gene pair appeared in the top data table, then RS = 0.5 or RS = 1, respectively.

If we screened our data at two different doses and observed only one dose or both doses in the top data table for a compound–gene pair, then RS = 0 or RS = 1, respectively.

For all data screenings at only one dose, RS = 0.

- **Dose Score (DS):** If a compound targeted a gene in the heterozygous interaction (HET) data, we anticipated that the heterozygous interaction scores would drop by increasing the dose. We manifested this phenomenon by a general definition for the dose score where DS = (Spearman's rank correlation coefficient of the vector of heterozygous interaction scores and the corresponding vector of doses) * (–1) + 1) / 2. We also required that a compound–gene pair to fulfil the criterion of being at least an outlier in both compound-wise and gene-wise senses in order to gain DS > 0.

  We observed in our heterozygous interaction (HET) data that some interactions did not follow the aforementioned criterion for increasing doses, but those

interactions were strong or even extremely strong at different doses. Therefore, we added the following correction to the dose score in order to rescue those interactions that fell into the region of strong or extreme outliers either in the compound-wise or in the gene-wise sense. This dose scoring correction ignored the use of Spearman's rank correlation coefficient for computing the dose score and relied solely on the position of heterozygous interactions in the strong and extreme regions of outliers: If a given interaction was a strong outlier in either the compound- or the gene-wise sense, the interaction received 0.25 for each region. If an interaction was an extreme outlier in either the compound- or the gene-wise sense, the interaction received 0.5 for each region. For instance, if an interaction demonstrated to be an extreme outlier in the compound-wise sense and a strong outlier in the gene-wise sense, DS = 0.75.

We included the above metrics in measuring a target score corresponding to every compound-gene interaction. A target score was computed for every gene entry of a top data table in the compound-wise sense and for every compound entry of a top data table in the gene-wise sense. We defined a simplified form of a target score as follows:

$$\text{Target Score} = \alpha\,\text{SS} + \beta\,\text{HS} + \gamma\,\text{RS} + \delta\,\text{DS}$$

$$\text{Where } \alpha, \beta, \gamma, \delta > 0 \text{ and } \alpha + \beta + \gamma + \delta = 1$$

$$\text{In our analysis, we set } \alpha = \beta = 0.35 \text{ and } \gamma = \delta = 0.15$$

Although we described target scores based on heterozygous diploid mutant screens as our gold standard for target identifications, we also executed our computational ranking framework for temperature-sensitive and overexpression interactions (two other sources of essential gene interactions) to generate complementary scores for target identifications.

3.3.7 Driver scores of bioprocess target predictions for target identification

We previously developed the CG-TARGET pipeline that facilitated the annotation of compounds to their bioprocess target predictions using chemical-genetic interactions in yeast (74). CG-TARGET identified the driver genes that were the members of a bioprocess that led to its prediction. To identify the driver genes, CG-TARGET analytically generated a driver score for every gene in the chemical-genetic interaction pool – the higher a driver score, the more reliable the corresponding bioprocess target prediction for the compound.

We used the driver scores computed by the CG-TARGET pipeline as an additional layer of prioritization to promote the compound–gene pairs with both high target and high driver scores. We generated these driver scores for all chemical-genetic interaction pools including the diagnostic pool of genetic mutants. Each compound–gene pair in our analysis used the corresponding driver score obtained from this diagnostic pool of chemical-genetic interaction mutants as an additional prioritization layer.

3.3.8 Bioprocess target prediction agreements across different mutant pools

To predict a specific target for a compound at the gene level more accurately, we determined if the gene target predictions for a compound also agreed at the bioprocess level across mutant collections. We measured the overlap between the process target predictions that resulted from the diagnostic mutant collection and the ones that resulted from either the whole genome or the temperature-sensitive mutant collection. This agreement criterion was designed to be lenient to provide a general agreement of bioprocess predictions related to a compound–gene pair and required that only one of the (three) available diagnostic pool data replicates agreed with either the whole-genome data or one of the (three) temperature-sensitive data replicated at the bioprocess prediction level. We developed a pipeline for evaluating these bioprocess target prediction agreements across mutant pools (**Figure 3.5**). First, for every compound exposed to the diagnostic mutant collection, we generated all bioprocess predictions using CG-TARGET (74). Then, we isolated a subset of these bioprocess predictions that also existed for the whole genome or temperature-sensitive mutant collection as provided by the CG-TARGET pipeline. For each surviving bioprocess prediction, we collected all the overlapping genes that drove this bioprocess prediction in our two different data pools using the CG-TARGET lists of driver genes. Since we tested a larger set of essential genes against a compound in the heterozygous diploid collection (~973 essential genes for our RIKEN high-confidence set), we systematically added all missing essential genes, using Gene Ontology (GO) annotations, to the set of overlapping driver genes to ensure that we would not miss any potential gene targets.

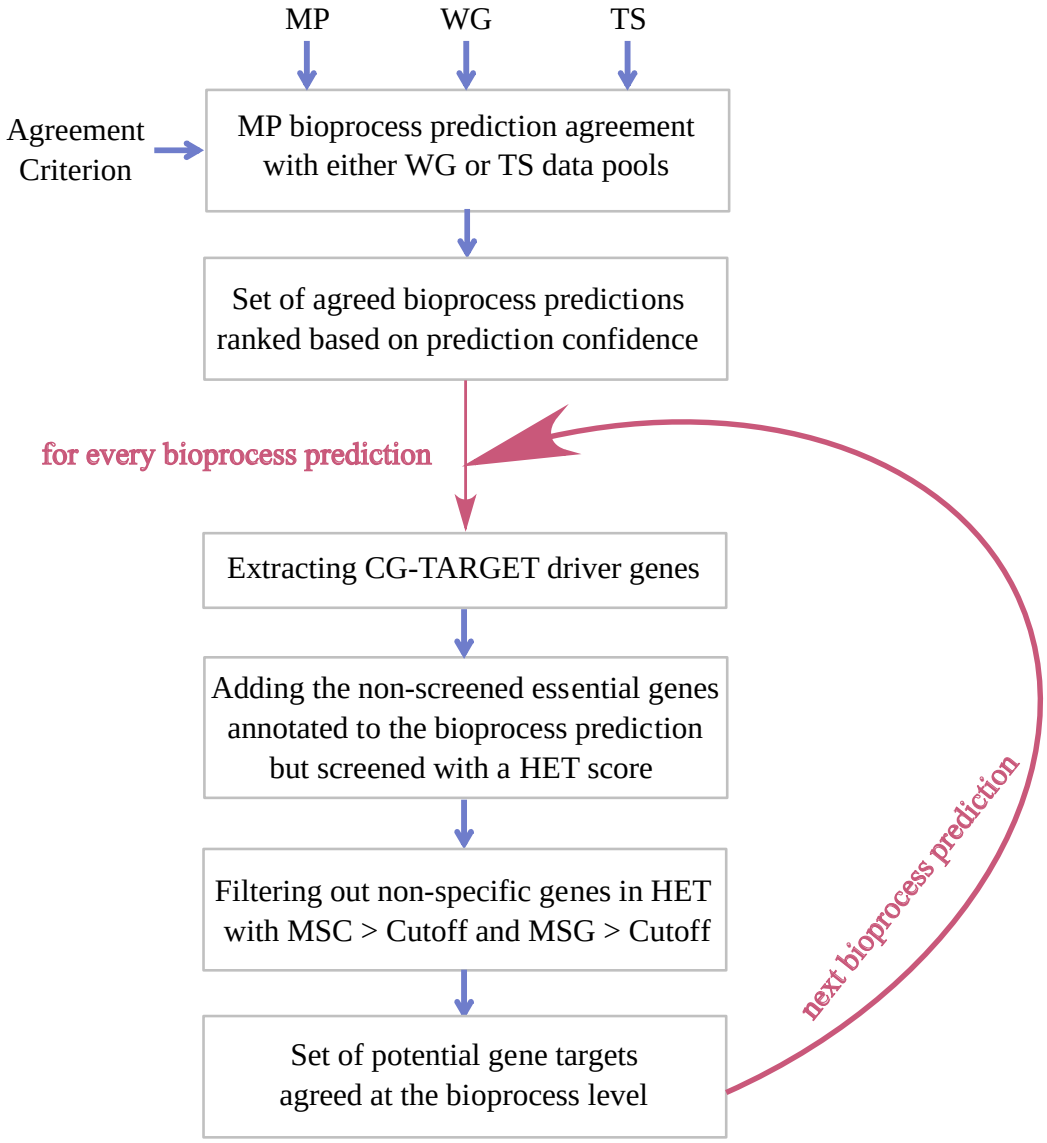


Figure 3.5. Pipeline for bioprocess target prediction agreements across data pools. The agreement criterion required that the bioprocess predictions from at least one of the (three) replicates of the diagnostic mutant pool to agree with the predictions obtained from either the only whole genome or at least one of the (three) temperature-sensitive chemical-genetic interaction replicates. All these bioprocess target predictions were generated using CG-TARGET (74). After we generated a ranked list of agreed bioprocess target predictions for a compound (ranked using CG-TARGET confidence scores), we performed the rest of the pipeline for every one of these predictions. For each bioprocess

prediction, we collected the driver genes from all the agreed pools and added to this list the essential genes that were annotated to the bioprocess prediction but were not present in our chemical-genetic interaction data. We filtered these genes for specificity using predefined cutoffs on MSC and MSG measurements and generated a set of potential gene targets that agreed at the bioprocess level. We repeated this cycle for all the agreed bioprocess target predictions, separately. Here, MP represents the diagnostic or miniaturized pool of mutants, WG represents the whole genome mutant pool, and TS represents the temperature-sensitive mutant pool.

3.3.9 Structural similarity measurements

In the process of identifying compounds that might have targeted a given gene, we required that the compounds emerging to the top data table of the gene to be structurally distinct. This criterion helped us find molecules with distinct scaffolds targeting a gene. In order to meet the need for identifying structurally distinct compounds targeting a gene, we added all pairwise structural similarities of the compounds that targeted the gene of interest to the top data table of the gene using APS molecular fingerprints (depth 8) and the Braun-Blanquet similarity coefficient (169) (**Table B.2** – Structural Similarity spreadsheet).

3.4 Results and Discussion

3.4.1 Novel mutant reagent sets for high resolution target identification in *Saccharomyces cerevisiae*

Since a subset of genes were sufficient to capture functionally informative genetic interactions across the genome (72,166), we previously designed a miniaturized, diagnostic collection of 310 non-essential gene deletion mutants (73). This gene subset was optimized

for gene-similarity-based target prediction across the entire set of genetic interaction query strains. To circumvent the highly efficient drug efflux system of yeast, the collection was constructed in a drug hypersensitive genetic background. This included deletions of *PDR1* and *PDR3*, which are transcription factors that regulate the yeast pleiotropic drug response, and a deletion of *SNQ2*, a multidrug transporter. Each mutant in the diagnostic collection was tagged with a DNA barcode, which enabled highly multiplexed barcode profiling and rapid prediction of high confidence target bioprocesses for compounds from several diverse compound libraries (73,170).

In our previous study (73), we characterized the functional behavior of compounds at the bioprocess level and annotated large libraries of our compounds to their bioprocess predictions with high confidence using the CG-TARGET pipeline (74). However, for many cases of target identification, we need to identify the target of a compound at the protein level. For such cases, we are usually interested in narrowing down the target of a compound to a specific protein. We refer to the gene target of a compound, instead of its target at the protein level, here, since we measure similarity to genetic effects.

To find the direct targets for compounds using higher resolution interaction profiles with richer information, we leveraged essential gene reagent sets available in *S. cerevisiae* and constructed new drug hypersensitive mutant collections in the same *pdr1Δpdr3Δsnq2Δ* genetic background. The sets included a temperature sensitive mutant collection for ~1,154 essential genes, a heterozygous diploid collection for ~973 essential genes, and a collection of genetic mutants that overexpressed for ~1,110 essential genes on high copy plasmids (**Table 3.1**). We also constructed a haploid deletion mutant collection for all ~3,500 non-

essential genes to provide complementary information to the compound screening that we previously conducted using only the diagnostic collection of the yeast genes. Gold-standard compounds were included in each screen in order to evaluate the screening performance. We observed that certain DNA-barcodes were under-represented in our pooled sequencing experiments, while others were over-represented. To prevent technical effects due to these differences in barcode performance, we selected the barcodes that were amplified and were sufficiently detectable in sequencing reactions for each mutant collection.

3.4.2 Examples of compound–gene target validations

We used a high-throughput chemical-genetic interaction platform in *S. cerevisiae* to measure the effects of compound exposures on the fitness of hundreds of mutants across the entire yeast genome. We, then, developed a computational framework that combined the information from all various types of these chemical-genetic interaction experiments to promote the compound–gene pairs where the compound of interest could potentially target the gene. We used our target scores (see 3.3.6) to generate a ranked list of compound–gene pairs in decreasing order of their priority for experimental investigation. We generated such ranked priority lists in both compound-wise and gene-wise senses. In this subsection, we briefly discuss three validated examples of compound-gene targets that were prioritized in using our proposed target scores.

- **Clionamines targets PIK1**

Tuberculosis (TB) is a lethal bacterial infectious disease caused by Mycobacterium tuberculosis (Mtb). Though Mtb pathogens can attack any organ in the body, these bacteria usually attack the lungs and cause deadly respiratory infections. TB is a world's leading cause of death according to the World Health Organization (WHO). In 2021, an estimated 10.6 million new cases of TB infections were diagnosed, and an estimated 1.6 million patients died from TB (171). WHO also estimated that about 18-21% of the patients who previously received treatments for TB developed multidrug-resistant/rifampicin-resistant (MDR/RR) TB (172). Multidrug-resistant TB (MDR-TB) is defined as resistance to both isoniazid and rifampicin, the two most effective first-line drugs for TB treatment. MDR-TB requires treatments with second-line drugs, which are usually less effective with more side-effects. Moreover, a large number of infected people are usually asymptomatic at the initial stages of illness but develop the TB disease with a deadly outcome years later (latent TB) (173). Hence, we need to employ novel approaches to discover new therapeutic drugs for addressing the MDR-TB and latent TB infections that resist existing antimycobacterial chemotherapy regimens (174). One solution is to develop host-directed therapeutics (HDT) aiming at activating the xenophagy (selective autophagy) component of the innate immune system (175). In a recent publication (176), our colleagues at the University of Toronto and the University of British Columbia validated Clionamines, a marine natural product, which was synthetically produced as RP-3-161, as a potential second-line drug for TB treatments through stimulating autophagy. The RP-3-161 structure was the most effective autophagy-stimulating synthetic analog to the South African sponge *Cliona celata* of clionamines.

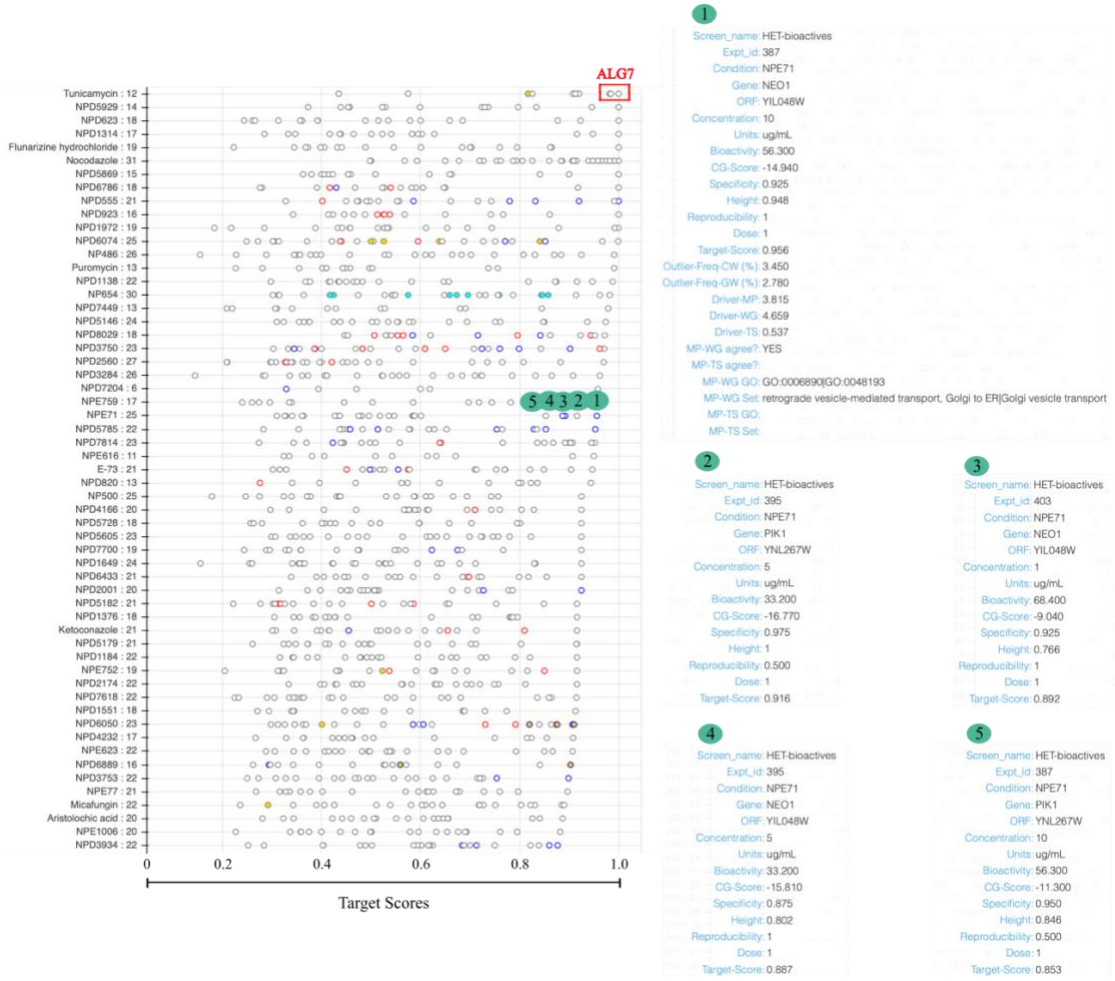


Figure 3.6. Priority gene targets for NPE71 from the RIKEN bioactive set. The top section of the HET-based priority plots for the combined collection of the RIKEN high-confidence and bioactive sets is represented. Our ranking system assigned high target scores to NEO1 and PIK1 as potential targets of NPE71. Blue circles show that at least one process prediction across two different mutant pools or the replicates of only one existing pool agreed (see 3.3.8). For two distinct mutant pools, the process predictions agreed for at least one replicate from each pool. If only one mutant pool existed, the process predictions agreed in at least two and the majority of replicates (if majority > 2 replicates). We showed this agreement in node hovers by YES along with the top three agreed bioprocess predictions (if available). Similarly, red circles show that the process predictions across two different mutant pools or the replicates of only one existing mutant pool agreed; however, their

corresponding chemical-genetic interactions were not available. The gold filling color represent that the highest driver score belonged to the diagnostic mutant pool. A hard cutoff for considering a node filled with a gold color was applied. This hard cutoff was driver score > 10 for the RIKEN high-confidence and bioactive sets. The cyan and chartreuse filling colors represent that the highest driver score belonged to the whole-genome or temperature-sensitive pool, respectively. The same cutoffs as above were considered for filling a node colored with cyan or chartreuse.

Our computational framework prioritized NEO1 and PIK1 as the potential targets of NPE71 and RP-3-161 based on target score measurements (**Figure 3.6** and **Figure 3.7**). Hence, our computational framework suggested that RP-3-161 could directly target PIK1 or modulate another protein or pathway that would depend on normal phosphatidylinositol-4-phosphate (PI4P) levels (176). A recent study of the chemogenomic fitness signatures in yeast indicated that compounds with cationic and amphiphilic properties (CADs) induced NEO1 haploinsufficiency through NEO1 and NEO1-PIK1 signatures (177). Therefore, our colleagues at the University of Toronto and the University of British Columbia also studied NPE71, another CAD-like compound, with the NEO1-PIK1 haploinsufficiency signatures prioritized by our computational pipeline. Our colleagues related the behavior of RP-3-161 to that of the CAD-like compounds using chemical-genetic interaction profiles. They found that RP-3-161 inhibited the yeast PI4 kinase *in vivo* as compound exposure led to decreased levels of trans-Golgi PI4P, and that NPE71 treatment led to a similar reduction in trans-Golgi PI4P levels too (176). The experimental validations proved the efficacy of RP-3-161 for the inhibition of Mtb survival in macrophages and identified PI4KB, a human homolog of PIK1, as a molecular target for the HDT treatments of the TB disease (176).

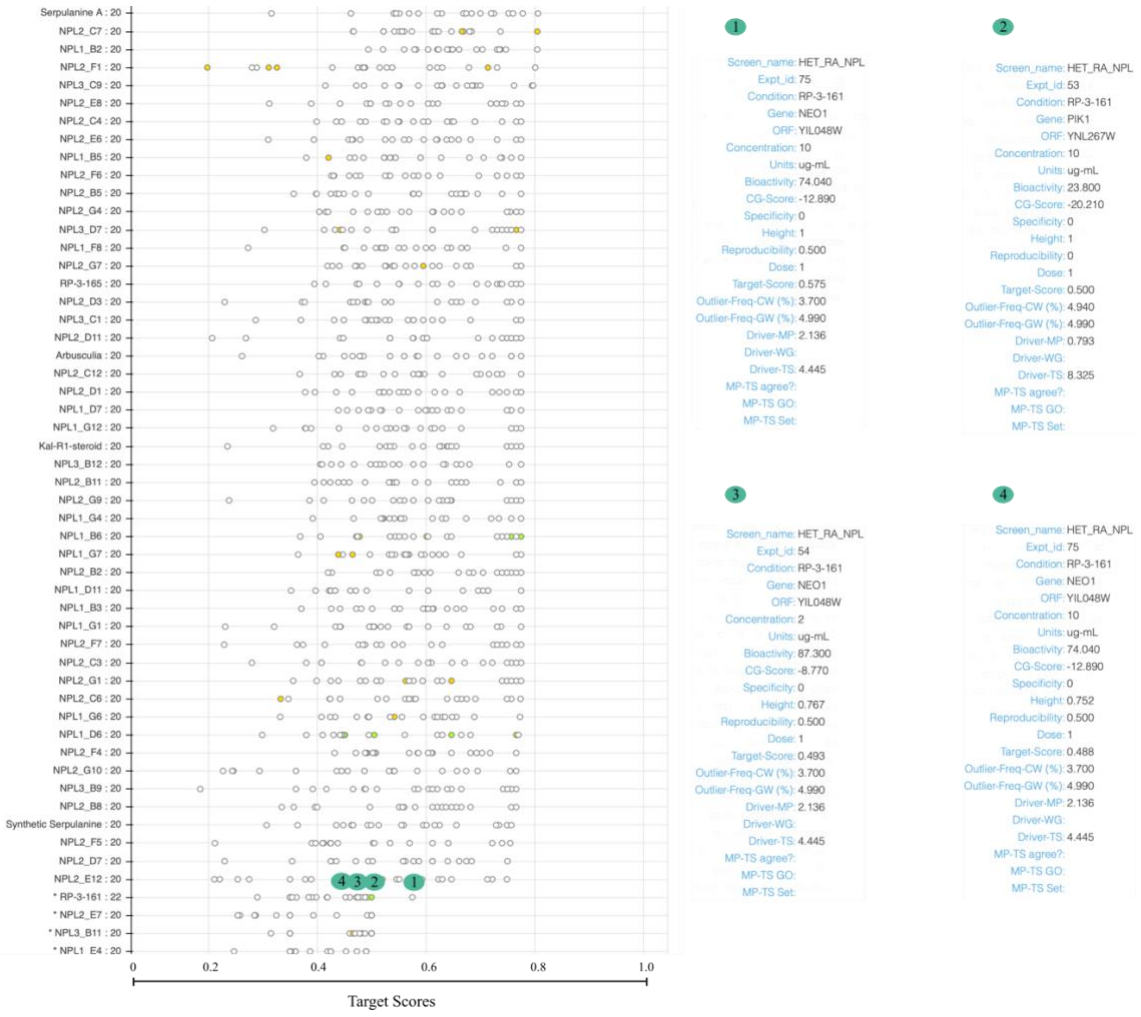


Figure 3.7. Priority gene targets for RP-3-161 from the University of British Columbia natural products library. Part of the HET-based priority plots for the University of British Columbia natural products library is represented. Our ranking system assigned high target scores to NEO1 and PIK1 as potential targets of RP-3-161. This compound is marked with * as a frequent flier compound, and thus, our ranking system considered 0 for the specificity of such compound–gene pairs. For small compound collections such as the University of British Columbia natural products library, we investigated all the existing compounds regardless of their specificity scores. Blue circles show that at least one process prediction across two different mutant pools or the replicates of only one existing pool agreed (see 3.3.8). For two distinct pools, the process predictions agreed for at least

one replicate from each pool. If only one pool existed, the process predictions agreed in at least two and the majority of pool replicates (if majority > 2 replicates). Similarly, red circles indicate that the process predictions across two different mutant pools or the replicates of only one existing mutant pool agreed, but their corresponding chemical-genetic interactions were not available. As the node hovers indicate, we did not find any agreement for the bioprocess predictions related to RP-3-161. The gold filling color represent that the highest driver score belonged to the diagnostic mutant pool. A hard cutoff for considering a node filled with the gold color was applied. This cutoff was driver score > 5 for the University of British Columbia natural products library. The cyan and chartreuse filling colors represent that the highest driver score belonged to the whole-genome or temperature-sensitive pool, respectively. The same cutoffs as above were considered for filling a node colored with cyan or chartreuse.

- **NPD6433 targets FAS1**

Infections caused by fungal pathogens are considered a major threat to public health (178–180). Each year, fungal diseases affect over a billion and kill more than 1.5 million people globally, mainly in immunocompromised individuals such as organ transplantation, cancer, AIDS, and asthma patients (181,182). Currently, only four classes of systematically acting antifungal drugs are available– polyenes, azoles, allylamines, and echinocandins (183,184). Antifungal pathogens are becoming increasingly resistant to these antifungal drugs, while entirely emerging pathogens such as the yeast *Candida auris* exhibit even multiple drug resistance (185). Additionally, rapid and sensitive diagnostic approaches do not exist for most fungal pathogens, and for those pathogens with existing diagnostics, the current diagnostics are not widely available or affordable globally. Consequently, delays

in establishing a definitively true diagnosis in critically ill patients often lead to poor patient outcomes (179). Therefore, development of antifungal drugs with broad-spectrum activity against prominent human pathogens is highly needed for treating fungal diseases.

Since robust inhibition of FAS1 is fungicidal (186,187), if we find compounds that target the activity of fungal fatty acid synthases (FAS), we will potentially discover a novel antifungal agent. Our computational framework prioritized FAS1 as the potential target of NPD6433 from the RIKEN high-confidence set using target score measurements (**Figure 3.8**). Our ranking framework used heterozygous diploid mutant collections to prioritize the compound–gene pairs where the compound potentially targeted the gene. In this ranking system, FAS1 stood out as the potential target of NPD6433 in two (out of three) available dosages of 0.5 μg/ML and 1 μg/ML (**Figure 3.8**). In addition, our prioritization framework pushed this compound–gene pair to the top of the comprehensive list of our potential compound–gene targets, which indicated the high priority of this candidate for further experimental evaluations. Therefore, our colleagues at the University of Toronto evaluated the NPD6433–FAS1 pair in several experimental settings and confirmed that NPD6433 inhibited FAS1. Chemical-genetic and biochemical approaches in *C. albicans* and the yeast *S. cerevisiae* indicated that NPD6433 targets the essential process of fatty acid biosynthesis in fungi by inhibiting the activity of fungal fatty acid synthase. In our experimental studies, NPD6433 was identified as a potent, broad-spectrum antifungal with fungicidal activity across all four prominent human fungal pathogens, *C. albicans*, *C. auris*, *C. glabrata*, and *C. neoformans*.

Figure 3.8. Priority gene targets for NPD6433 from the RIKEN bioactive set and NPD5728 from the RIKEN high-confidence set. The top section of the HET-based priority plots for the combined collection of the RIKEN high-confidence and bioactive sets is shown. Our ranking system assigned high target scores to FAS1 and FRS1 as potential targets of NPD6433 and NPD5728, respectively. Blue circles indicate that at least one process prediction across two different mutant pools or the replicates of only one existing mutant pool agreed (see 3.3.8). For two distinct pools, the process predictions agreed for at least one replicate from each pool. If only one mutant pool existed, the process predictions agreed in at least two and the majority of replicates (if majority > 2 replicates). Similarly, red circles indicate that process predictions across two different pools or the replicates of only one existing pool agreed; however, their corresponding chemical-genetic interactions were not available. The gold filling color shows that the highest driver score belonged to the diagnostic

mutant pool. A hard cutoff for considering a node filled with a gold color was applied. This cutoff was driver score > 10 for the RIKEN high-confidence and bioactive sets. The cyan and chartreuse filling colors represent that the highest driver score belonged to the whole-genome or temperature-sensitive pool, respectively. The same cutoffs as above were considered for filling a node colored with cyan or chartreuse.

- **NPD5728 targets FRS1**

Our computational framework prioritized FRS1 as the potential target of NPD5728 from the RIKEN high-confidence set using our target score measurements (**Figure 3.8**). This compound induced haploinsufficiency in diploids with one copy of FRS1 at multiple doses. Experimental evaluations based on a ribosome profiling assay confirmed FRS1 as the putative target of NPD5728 (data not shown). Experimental studies of NPD5728–FRS1 is still ongoing.

3.5 Conclusion

Finding the direct targets of chemical compounds is an important step in developing novel biopharmaceutical agents for diseases. We developed a computational framework that combined information from various sources of chemical-genetic interaction screens to generate a unified "target score" that prioritized the compound–gene pairs with the highest likelihood for the compound to target the gene. This ranking system integrated signals from heterozygous diploid and temperature-sensitive mutants for essential genes, whole-genome mutants, and a miniaturized collection of the diagnostic mutants in *S. cerevisiae*. The target scores accurately investigated the signals from heterozygous diploid mutant collections as

haploinsufficiency profiling could be a major source for target identification. This profiling runs under the principle that reducing the copy number of the gene that encodes the target of a compound or is involved in a pathway that relates to the target of the compound will result in hypersensitivity to the compound.

Our studies showed that the compounds RP-3-161 and NPE71 targeted NEO1 and PIK1 and inhibited the survival of Mtb in macrophages, providing a potential therapeutic solution for TB treatment. We also found that the compound NPD6433 targeted FAS1 in fungi and exhibited fungicidal activity against major human fungal pathogens. Our studies suggest that the proposed computational framework is fundamental in identifying the direct targets of compounds by speeding up the selection process for experimental evaluations of the most promising compound–gene pairs that may contribute to the development of novel drugs.

4 Target Identification of Bioactive Compounds Using a Network Diffusion Approach

This research work is the continuation of the research article "Functional annotation of chemical libraries across diverse biological processes", which we previously published at Nature Chemical Biology, Volume 13, Issue 9, 2017. We designed the research project in this chapter based on the results published in this article and used heterozygous diploid mutant collections for our evaluation purposes. This thesis chapter represents a draft of a manuscript that is in preparation for submission to a journal at the time of writing. Other contributors to this research project are Sheena C. Li, Zhijian Li, Charles Boone, and Chad L. Myers.

H.S. (dissertation's author), and C.L.M. conceived this project and established the methodologies. H.S. performed data analyses and designed the computational frameworks and evaluation methods presented in this chapter. S.C.L. and Z.L. generated heterozygous diploid mutant collections for evaluation purposes. C.B. and C.L.M. supervised the project and acquired funding. H.S. wrote the manuscript draft to be reviewed by other authors.

4.1 Abstract

Recent advances in high throughput chemical-genetic interaction screening technology demand the development of scalable computational methods for prioritizing candidate gene targets of bioactive compounds. We designed, developed, and validated a computational framework, called Precise Identification of Network-based Targets (PINT), which was inspired from the mathematical approaches for the detection of rumor sources in social and viral networks. The PINT framework models potential gene targets as rumor sources that spread rumors across modules of related genes using an underlying functional network of genetic interactions as a diffusion network. We used a large library of chemical-genetic interactions from the yeast *Saccharomyces cerevisiae* for compounds from the RIKEN Natural Product Depository and several NCI/NIH/GSK compound collections to validate PINT. This target prediction and prioritization pipeline provides a foundation for identifying the gene targets of bioactive compounds from chemical-genetic interaction data and facilitates the discovery of novel gene targets for uncharacterized compounds.

4.2 Introduction

Recent technological improvements in high-throughput screening approaches have encouraged the creation and expansion of diverse compound libraries that await functional characterization. Characterizing these ever-expanding libraires of small molecules is vital for the discovery of potential drug leads and chemical probes. However, current approaches primarily focus on the high-throughput screening of compound libraries for the agents that bind known protein targets or inhibit disease-related pathways (165). Therefore, despite substantial efforts in drug discovery, the number of novel drugs remains low, and the drugs with novel mechanisms of action are scarce (165). Hence, establishing new approaches that facilitate the rapid identification of direct targets of compounds using large-scale functional data for these compounds is inevitable.

Chemical genomic approaches offer a valuable new source of data for elucidating the biological function of bioactive compounds. The yeast *Saccharomyces cerevisiae* is a well-characterized eukaryotic system for which ~5000 viable deletion mutants have been identified (121). Interrogating these deletion mutants with a bioactive compound generates a chemical-genetic interaction profile that contains the degree of sensitivity or resistance of each genetic mutant to the compound (66,67). This quantitative and high-dimensional representation of biological function can be interpreted using a global compendium of genetic interaction profiles (64,65), which were systematically mapped through pairwise, double-knockout screens in *S. cerevisiae*. Specifically, the chemical-genetic interaction profile of a compound mirrors the genetic interaction profile obtained through the genetic knockout of the biomolecular target (71). These chemical-genetic interaction profiles can

be used as a gold standard for evaluating the biological activity of compounds, even for the compounds where the actual biological gene targets are not yet known, since these profiles accurately reflect compound function.

We recently generated, annotated, and published a large compendium of more than 13,000 chemical-genetic interaction profiles obtained in *S. cerevisiae* (73), providing a unique basis for exploring the function, and more specifically, the gene targets of chemical compounds. Our compounds were from the RIKEN Natural Product Depository (NPDepo) and several NCI, NIH, and GlaxoSmithKline (GSK) compound collections. We annotated these compounds to their most likely bioprocess target predictions using the CG-TARGET tool (74). For each compound, the CG-TARGET tool generated a ranked list of bioprocess target predictions for the compound. However, for many target identification purposes in drug discovery, we still need to identify the target of a compound at the protein level. We refer to the gene target of a compound, instead of its target at the protein level, throughout this chapter since we measure similarity to genetic effects.

We downloaded the significant bioprocess predictions (p-value<0.05) generated by the CG-TARGET tool for a compound from the MOSAIC database (76). We collected all the genes associated to these bioprocess target predictions by the CG-TARGET pipeline to construct a gene set. We overlaid this gene set, as a set of the driver genes of the significant bioprocess target predictions of a compound, on the global network of genetic interactions for *S. cerevisiae* (64,65) to make a rumor graph for the compound. Inspired by rumor source detection approaches in social and viral networks, we analyzed the constructed rumor graph and modeled the gene target of the compound as the rumor source that spread the inhibitory

signals of the compound across pathways supported by a global functional network of yeast genes. Therefore, the task of identifying the gene target of a compound was translated to finding the rumor source that infected the gene set across the rumor graph (**Figure 4.1**). To detect this rumor source for a compound with the PINT computational framework, we used as input a set of genes as potential gene targets and computed rumor centrality scores (188), where the gene with the highest rumor centrality score was highly likely to be the rumor source and, therefore, the gene target of the compound (**Figure C.1**). We evaluated PINT in several computational settings and concluded that the PINT model can more precisely identify the gene target of a compound in contrast to previous computational pipelines such as CG-TARGET.

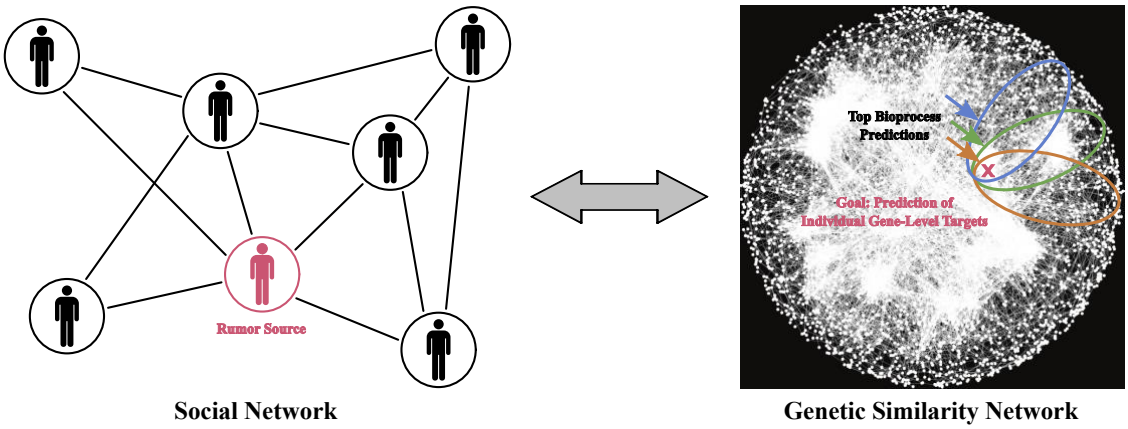


Figure 4.1. Modeling gene targets of compounds as rumor sources using a global genetic similarity network. The rumor graph was comprised of all the driver genes generated for the significant bioprocess target predictions linked to a compound (p-value<0.05) using the CG-TARGET tool. These bioprocess predictions were downloaded from the MOSAIC database (mosaic.cs.umn.edu) (76). The edge connections among the genes in the set of combined driver genes were defined using

an underlying network of genetic similarities to form a rumor graph. Similar to social and viral networks, where the rumor source is responsible for infecting individuals, the gene target of a compound was modeled to spread inhibitory signals across the rumor graph.

4.3 Results and Discussion

4.3.1 Modeling the rumor graph for a compound

We constructed a binarized network of functional similarities among the genes in *S. cerevisiae* by thresholding similarity in the global compendium of genetic interactions (64,65). Using a similarity threshold of 0.1, we constructed a genetic similarity network of approximately 565,000 edges, which corresponds to a density of ~2%. For each compound, we used this binarized network as the underlying functional network for connecting the genes with the most similar genetic interaction profiles. The rumor graph for a compound was a subset of the genes in this network and was composed of the sets of driver genes that we obtained from the significant bioprocess target predictions (p-value<0.05) generated by the CG-TARGET tool. On average, a compound from our studied compound libraries had ~22.2 significant bioprocess predictions, where each predicted bioprocess had ~16.7 genes (many of these predicted bioprocesses overlap in terms of their member genes). The rumor graph associated to a compound was constructed, on average, from ~68.1 unique genes as nodes. If a gene were a direct target of a compound, we modeled the effect of the gene as the rumor source that infected the rumor graph (**Figure 4.1**).

We modeled the spread of the rumor in the underlying similarity network based on the classical susceptible, infected, and recovered/removed (SIR) model, and then obtained

an estimator for the rumor source (or analogously, the gene target) that assigned to each node (or gene) of the rumor graph a likelihood for spreading the rumor, which we hereafter call “rumor centrality (R)” (188). This quantity is the maximum likelihood estimator for the rumor source– the maximal rumor centrality for a gene corresponds to the maximum likelihood of the gene to be the target of the compound for which we created the rumor graph. For every compound, we generated a rumor graph that contained the driver genes from all significant bioprocesses predicted by the CG-TARGET tool (p-value<0.05; data downloaded from the MOSAIC database (76)) and computed a ranked list of potential gene targets based on rumor centrality scores.

Drug-induced haploinsufficiency occurs when reducing the dosage of an essential gene from two copies to one in diploid cells results in a heterozygote that displays increased drug sensitivity compared with wild-type strains (95). This method can precisely reveal the essential gene targeted by a compound. The intuition behind using heterozygous diploid mutants for identifying gene targets of compounds is that, if a gene is essential for growth, the heterozygous diploid mutant of the gene whose product is targeted by a compound is sensitive to the compound. We generated a collection of heterozygous diploid mutants in *S. cerevisiae* for our compounds, which we call HET profiles, as the gold standard for the direct gene target identification of compounds. The importance of HET profiles for target identification purposes is described in Chapter 3 in details.

We developed the PINT framework for target identification of compounds using a diffusion network of global genetic similarities and evaluated the ability of this framework to prioritize candidate gene targets using three different approaches as follows: evaluating

the PINT framework against random rumor graphs, against CG-TARGET driver scores, and against HET dropout scores. The results of these three approaches are discussed in the following sections.

4.3.2 Evaluating the PINT framework against random rumor graphs

We generated chemical-genetic interaction profiles using a miniaturized collection of the diagnostic mutants in *S. cerevisiae* (~300 genetic mutants), which were predictive of all major yeast bioprocesses (72,166). We published these chemical-genetic interaction profiles for 13,431 compounds from several diverse compound collections and identified a subset of these compounds (1,504 compounds) that were predicted with high-confidence to perturb predefined biological processes based on integrating these profiles with genetic interaction profiles (73). We used these high-confidence compounds and predicted the gene target with the highest likelihood for each compound using the PINT approach. This high-confidence set was composed of two major subcollections: RIKEN high-confidence set (846 compounds) and NCI/NIH/GSK high-confidence set (659 compounds).

For every compound in the RIKEN or NCI/NIH/GSK high-confidence set, we used the significant bioprocess target predictions (p-value<0.05) generated by the CG-TARGET tool (74) and combined the associated driver genes generated by the CG-TARGET method. We used a global genetic similarity network (64,65) as an underlying functional network to construct a rumor graph for the compound. The PINT framework received the bioprocess target predictions annotated to a compound and computed a ranked list of rumor centrality scores for the compound (**Figure 4.2A**).

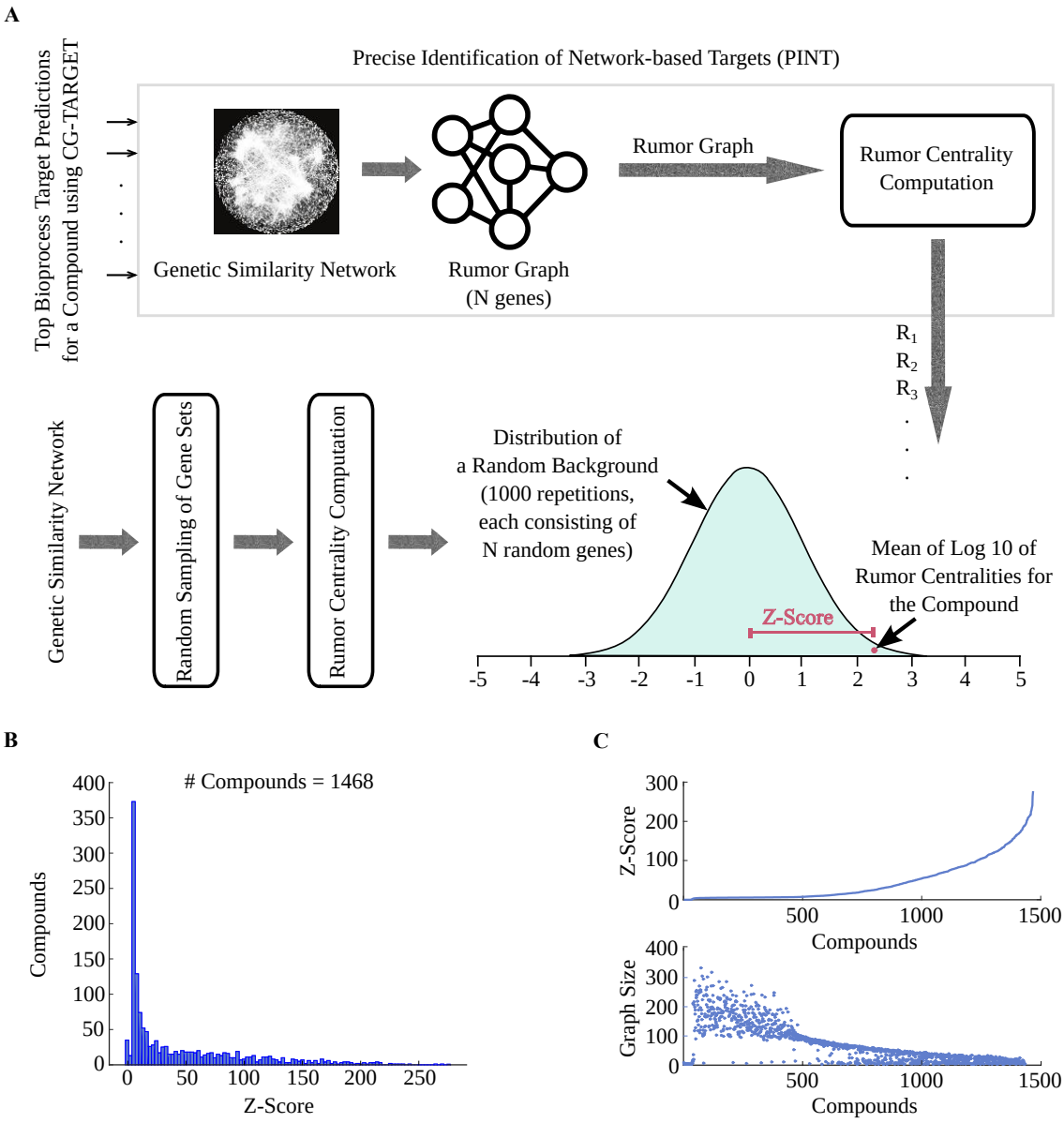


Figure 4.2. Evaluating the PINT framework against random rumor graphs. **(A)** For each compound, we constructed a rumor graph using the CG-TARGET driver genes of significant bioprocess target predictions (p-value<0.05) and overlaying those genes on the global genetic similarity network. The PINT framework generated rumor centrality scores for this rumor graph. We contrasted these rumor centrality scores to a distribution of such scores from random graphs and rejected the null hypothesis that the rumor centrality scores generated by PINT were random scores. (B) The null

hypothesis was strongly rejected by 1,427 compounds with z-score > 3, which accounted for ~97% of the total compounds. (C) The z-scores associated to the compounds were ranked in ascending order. The large rumor graphs generally received smaller z-scores, compared to the full range of z-scores.

As the first evaluation approach, PINT rumor centrality scores were compared against a null distribution of gene scores for each compound to calculate a z-score for that compound. To generate this null distribution, we sampled a thousand random gene sets, each as the same size as the input rumor graph of the compound, chosen from the global genetic similarity network. Using each one of these random gene samples, we constructed a random rumor graph and computed rumor centralities corresponding to this rumor graph. We then computed the z-score of the mean of rumor centralities (on a logarithmic scale) given the distribution of a thousand random means corresponding to random rumor graphs. This evaluation step was designed to test whether rumor centrality scores were significantly higher for actual CG-TARGET bioprocess target predictions than they were for randomly selected gene sets of the same size.

For 1,468 total compounds in our evaluation set for which we computed these z-scores, we found that 1,427 compounds (~97%) had a z-score > 3 (**Figure 4.2B** and **Table S1**). This strong rejection of the null hypothesis of random rumor graphs for a large number of our compounds confirmed that our rumor centrality scores were different from the scores coming from randomly selected gene sets. We ranked these scores in ascending order and observed that larger rumor graphs received smaller z-scores (**Figure 4.2C** and **Table S1**). For all these larger rumor graphs with one exception, we still had a z-score > 3. However,

these smaller z-scores, compared to the full range of z-scores, indicated that enlarging the rumor graphs using large bioprocess target predictions or a large number of bioprocess predictions generally weakened the rejection of the null hypothesis. Hence, we recommend trimming large rumor graphs by using only the very topmost significant bioprocess target predictions associated to a compound.

4.3.3 Exploring the correlation of rumor centrality and CG-TARGET driver scores

As another evaluation of the PINT approach, we tested the extent to which the rumor centrality score for each compound correlated with the original driver scores. The CG-TARGET driver genes were used as the basis for defining the initial rumor graphs in PINT, but the actual driver scores, which capture the direct correlation between that gene's interaction profile and the compound's chemical-genetic interaction profile, were not used as input for PINT. We only considered rumor centrality scores amongst the genes in the driver set and computed the Pearson correlation coefficient between rumor centrality and CG-TARGET driver scores (**Figure 4.3**). The distribution of correlation coefficients was significantly shifted towards positive correlations, with an average correlation of ~0.23 for the subcollection of 1,396 compounds for which we had profiles from both gene target identification frameworks (**Figure 4.3**). This provides further evidence that PINT's ranking of candidate gene targets is distinguishing closely related vs. more distantly related genes even within the target bioprocess.

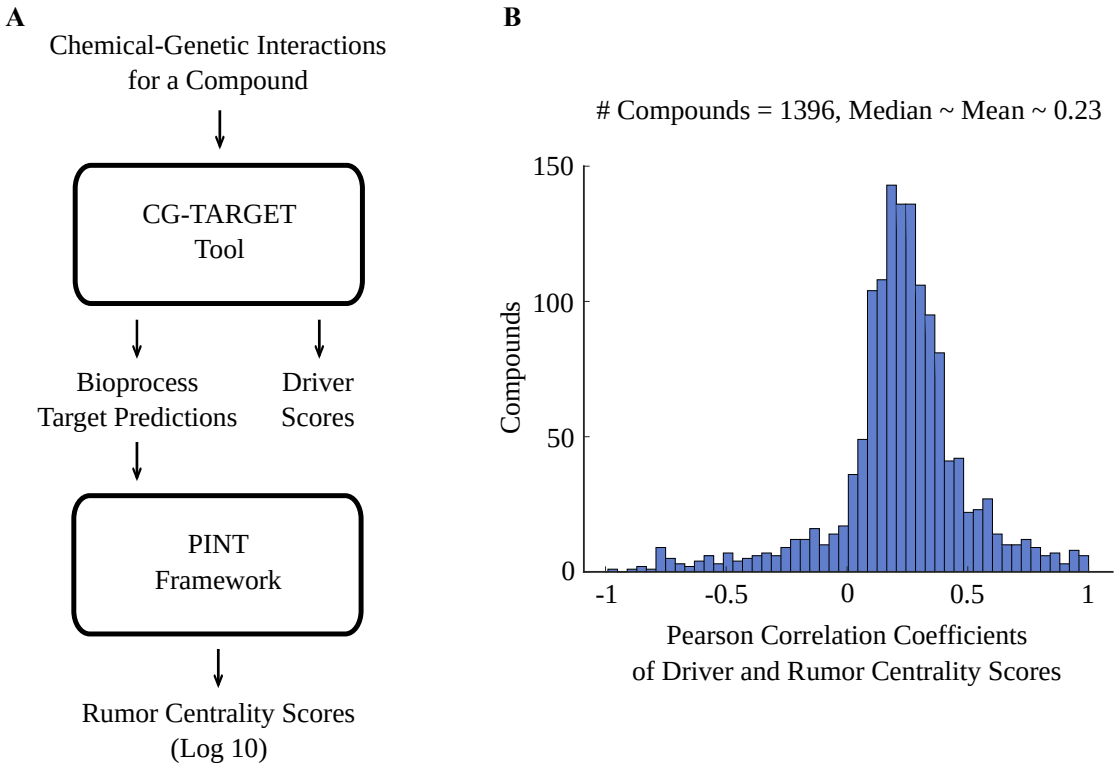


Figure 4.3. Correlation analysis of rumor centrality scores with CG-TARGET driver scores. **(A)** General flow of gene target identification using CG-TARGET and PINT, where PINT generated a ranked list of rumor centralities for gene target identification of compounds. (B) Distribution of the Pearson correlation coefficients of driver and rumor centrality scores.

4.3.4 Evaluating gene targets using heterozygous diploid mutants

Chemical-genetic interactions with essential genes play an important role in target identification of compounds because compounds capable of inhibiting growth are likely to have targets encoded by essential genes, which also tend to be highly conserved across a range of species. Essential genes are densely connected hubs in the genetic interaction network and provide rich information about the functional relationships between the well-conserved processes that are required for life. The direct gene target of a compound can be

revealed by interactions with essential genetic mutants at altered gene dosages, and the compound hypersensitivity due to reduced gene dosage in a diploid with a heterozygous deletion can point to a direct gene target for a compound (95). For example, Tunicamycin induces haploinsufficiency in diploids that have one copy of the *ALG7* gene, which encodes an enzyme that catalyzes the first step of asparagine-linked glycosylation and is directly inhibited by Tunicamycin (95). We probed essential genes by reducing the copy number to form a heterozygous diploid mutant collection in *S. cerevisiae*. We generated the HET data for ~973 essential genes and used these data to evaluate rumor centralities generated by PINT. Although we ran PINT for ~1500 compounds, only ~700 compounds had HET profiles (**Figure 4.4**).

We developed a randomization model (**Figure 4.4A**) for evaluating the strength of rumor centralities in predicting HET dropouts, under the assumption that the strongest HET scores are indicative of the true gene target (**Figure C.2**). We generated a vector of ranks that mapped genes with high rumor centrality scores to their position in the ranked list of HET scores sorted in ascending order. Assuming that the genes with top rumor centrality scores were the genes with a higher probability of being the target, we expected that top rumor centralities would receive lower mapping ranks. Therefore, we considered only top rumor centralities of a compound to be mapped to their corresponding HET scores (e.g., we used only top three rumor centralities in **Figure 4.4B)**. The mean of these top mapping ranks was used as an indicator of the strength of connection from rumor centralities to HET data, with lower means indicating a stronger connection between the rumor centrality and the HET data.

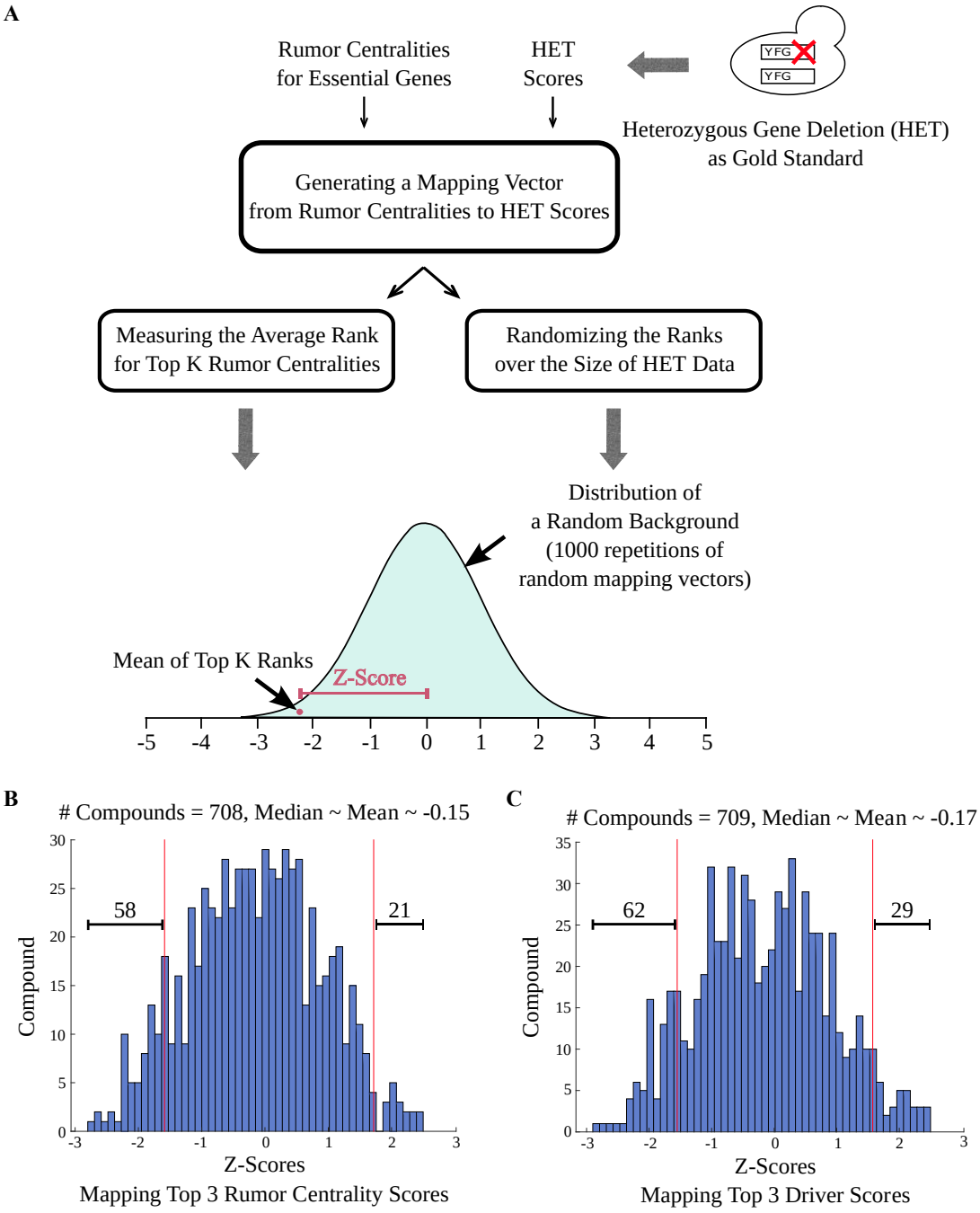


Figure 4.4. Evaluation of PINT gene targets using HET data. **(A)** Randomization model. For each compound, we ranked the rumor centrality scores of essential genes in ascending order against the ranked list of HET scores in descending order. These rumor centrality scores represented a subset

of the rumor nodes for the compound, while the ranked list of HET scores covered all the essential genes for which we had a HET score (~973 essential genes). We mapped top K rumor centrality scores to their corresponding HET scores (**Figure C.2**) and measured the average mapping rank for these top K rumor centralities. We, then, generated 1000 random mapping vectors of size K in [1, size of HET data] and computed the average for each random vector in order to generate a random distribution. We used this random distribution to compute a z-score for the compound. **(B)** Distribution of z-scores for K=3. **(C)** The same analysis as (B) except that we used CG-TARGET driver scores in place of rumor centrality scores.

For each compound, we generated a random distribution of the mapping ranks and computed a z-score for the compound using this distribution (**Figure 4.4A**). We observed that the distribution of these z-scores when we used only the top three rumor centralities in the mapping system was shifted toward negative values, reflecting a tendency for high-scoring HET mutants to fall near the top of the PINT ranked candidate list (**Figure 4.4B**). The ratio of the left tail (p-value<0.05 for the compound) to the right tail (p-value<0.05 for the compound) of this z-score distribution was ~2.8, indicating that our rumor centrality scores contained signals from HET data, which was used as the gold standard for the target identification of compound. Notably, when we replaced rumor centrality scores with the CG-TARGET driver scores in this analysis, we gained a smaller ratio of ~2.1, implying that rumor centralities provided a stronger connection to HET data than using the driver scores directly (**Figure 4.4C**).

4.3.5 Example compound case: Benomyl

Benomyl is a systematic benzimidazole fungicide widely used in agriculture. This drug exerts its toxicity by inhibiting microtubule assembly, which causes interference with a number of cellular processes. As a drug belonging to benzimidazole fungicides, Benomyl causes the blockage of nuclear division and disruption of mitosis and meiosis in cells (189). Benomyl binds to β-tubulin, which is encoded in *S. cerevisiae* by the TUB2 gene (189).

For Benomyl, we ranked the CG-TARGET driver scores against rumor centrality scores and noticed that the TUB3 gene received the highest scores in both spaces (**Figure 4.5**). TUB2 was not generated by CG-TARGET as a driver gene for Benomyl because we did not have TUB2 mutants in our chemical-genetic screening pools; as a result, TUB2 was not an input to PINT for constructing rumor graphs. However, TUB3 binds TUB2 to form tubulin dimers, which polymerize to form microtubules. Hence, TUB3 was about as close as to the actual target of Benomyl in our experiments. The high scores of TUB3 generated by both the CG-TARGET and PINT frameworks implied that both the frameworks pointed at the right target for Benomyl. This analysis showed that the PINT framework added more precision to the target identification process initiated by the CG-TARGET driver genes.

For this analysis, the total number of driver genes for which PINT generated rumor centrality scores was 214, and the Pearson and Spearman's rank correlations for this gene set were computed 0.43 (p-value = 3.0×10^{-11}) and 0.43 (p-value = 4.1×10^{-11}), respectively (**Figure 4.5**). These results indicated that the PINT framework was able to prioritize TUB3 as the target of Benomyl among the total of 214 candidate genes, and that the ranked lists generated by PINT and CG-TARGET were distinct although correlated.

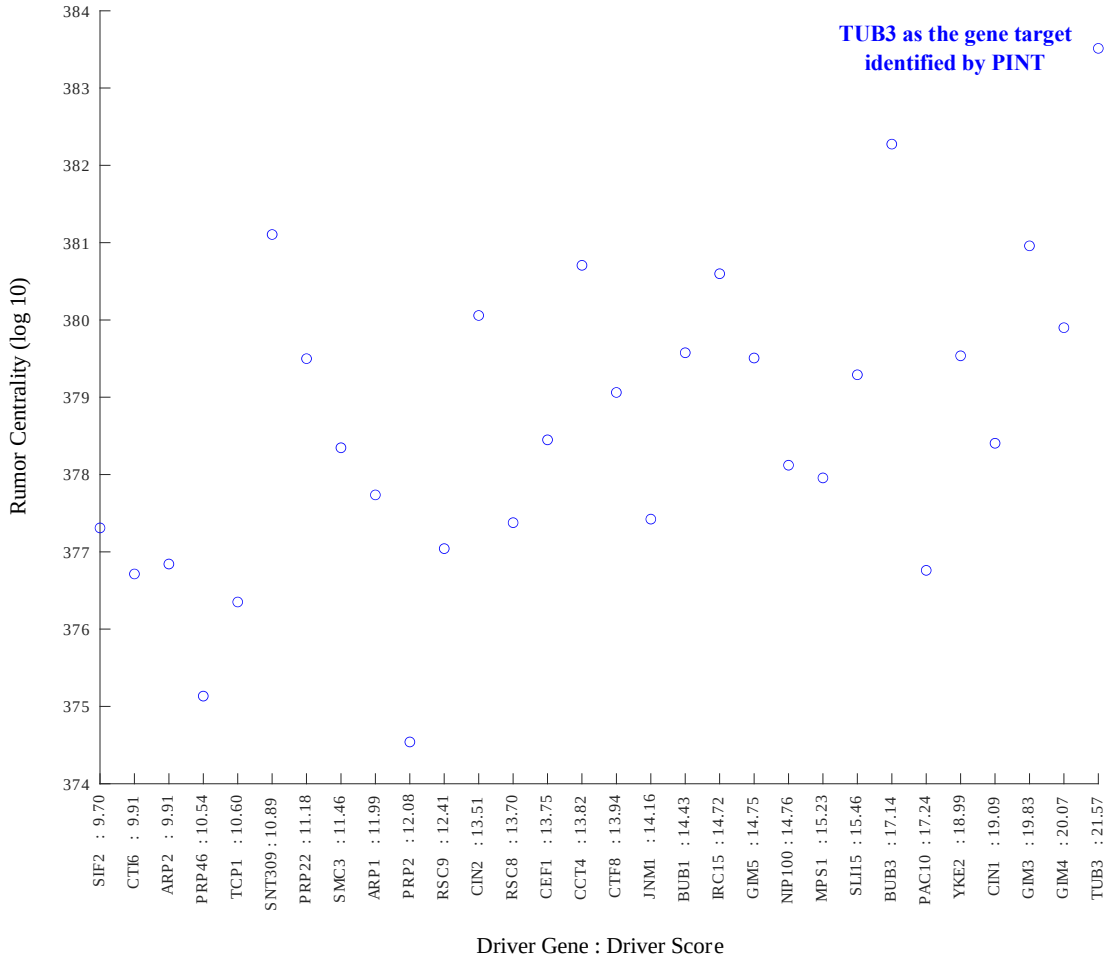


Figure 4.5. Target identification for Benomyl using PINT. The x-axis represents the top 30 driver genes generated by the CG-TARGET tool ranked in ascending order of their driver scores. If a driver gene contributed to different bioprocess target predictions, the mean of its driver scores was used. The y-axis represents the rumor centrality scores generated by PINT in a logarithmic scale. CG-TARGET identified these genes along with 184 other genes (not shown here) as the drivers of the bioprocess target predictions for Benomyl. The top three predicted bioprocesses were tubulin complex assembly, microtubule-based process, and microtubule cytoskeleton organization. PINT identified TUB3 as the specific gene target of Benomyl among 214 candidate genes by assigning the highest rumor centrality score of ~$10^{383.5}$ to TUB3.

4.4 Conclusion

We developed the PINT framework for target identification of compounds using a functional genetic network. We translated the task of identifying the target of a compound to that of finding the rumor source in social or viral networks. We connected the genes with the greatest potential of being the target for a compound by overlaying those genes on the genetic similarity network and creating rumor graphs. We selected such genes from the set of driver genes produced by the CG-TARGET method for the topmost bioprocess target predictions annotated to the compound. Using such a rumor graph, we generated a rumor centrality score for each gene, where the score reflected the likelihood of the gene to be the target of the compound. We compared our results with a random background and strongly rejected the null hypothesis that rumor centralities for the actual rumor graph would follow the same distribution as the rumor centralities for randomly generated rumor graphs. We used HET data as the gold standard for direct target identification and assessed the strength of connection from rumor centralities to the HET data using a mapping system from rumor centrality scores to the HET scores. We found that rumor centralities yielded a stronger connection to the HET profiles of our compounds than using CG-TARGET driver scores directly. Overall, the PINT target identification framework provides a strong computational basis for integrating functional networks to prioritize gene targets from chemical genomic data.

5 Conclusion and Future Work

The discovery and development of new drugs that demonstrate desired biochemical behavior against a target are challenging and expensive. One common paradigm to discover new drugs in shorter time and at lower cost is to search in structural compound libraries for the candidates that are structurally similar to a query drug or well-characterized compound because structurally similar molecules will more likely possess similar biological activities. Using this strategy, researchers in the field of drug discovery can identify a reasonably short ranked list of candidate compounds, from freely available or inexpensive structural databases, with high structural, and therefore functional, similarity to a query compound. This prioritized list of candidate compounds narrows down the number of compounds for experimental study in biochemical laboratory settings. Using large collections of chemical-genetic interactions from *Saccharomyces cerevisiae* as a systematic, unbiased proxy for the biological activity of our compounds, we developed two computational frameworks for predicting the function of compounds using chemical structures.

Systematic benchmarking framework: In a systematic manner, we benchmarked 11 different molecular fingerprint encodings, each combined with 13 different similarity coefficients. We found that All-Shortest Path fingerprints paired with the Braun-Blanquet similarity coefficient provided superior prediction performance being robust across several compound collections. Prior to this study, the Tanimoto/Jaccard similarity coefficient was commonly used for the similarity measurement of fingerprints, and ECPF fingerprints were widely used for describing molecular graphs of compounds in the two-dimensional space.

Machine learning framework: We developed a machine learning pipeline based on support vector machine models to improve the functional prediction of our compounds from chemical structures. Our learning pipeline provided up to fivefold improvement for a collection of natural products compared to our predictions using the ASP/Braun-Blanquet similarity measure. We observed that compound collections with low functional diversity would substantially benefit from our machine learning framework and that our machine learning framework was able to identify the functionally analogous compounds that were structurally distinct based on ASP/Braun-Blanquet similarity scores. We provided this framework freely available to researchers at https://github.com/csbio/VS-SVM.

Target identification of compounds is another challenging problem in the field of drug discovery and design. Genome-wide chemical-genetic interaction analysis monitors the effect of bioactive compounds on the fitness of specific genetic mutants and provides an unbiased method for determining the molecular mechanism of action for a compound. Using large collections of chemical-genetic interactions in *S. cerevisiae*, we addressed the problem of gene target identification for compounds through two different computational frameworks.

Computational framework for target identification of bioactive compounds using a high-resolution, integrated yeast chemical-genetic approach: We developed a pipeline that integrated information from several modalities of chemical-genetic interaction screens and generated a unified target score for every compound–gene pair; the higher the target score, the higher the likelihood of the compound to target the gene of interest. These target scores accounted for signal specificity and strength of chemical-genetic interactions. Our ranking

system computed the regions of overlap for the predicted bioprocesses from orthologous reagent sets as well. In generating target scores and organizing promising compound–gene pairs for experimental evaluation, we used information from the essential and non-essential genes in *S. cerevisiae*. The essential gene mutant libraries were comprised of heterozygous diploid gene deletion mutants, temperature-sensitive chemical-genetic interaction mutants, and overexpression strains, while the non-essential gene mutant libraries were comprised of the diagnostic set and the complete set of viable haploid deletion mutants in yeast.

Target identification of bioactive compounds using a diffusion network of yeast chemical-genetic interactions: We developed a computational pipeline inspired from the approaches for the detection of rumor sources in social and viral networks to facilitate the elucidation of the target genes of chemical compounds. This pipeline modeled target genes of compounds as rumor sources that spread compound inhibitory signals across pathways using an underlying functional network of genetic interactions as a diffusion network. We used a large set of chemical-genetic interactions to evaluate our models in computational settings.

Overall, these computational pipelines provide a strong foundation for identifying compounds with certain cellular functionality or compounds that target certain genes. However, an interested researcher may extend the scope of any of these computational frameworks. For the benchmarking system, one may describe compounds in other two- or three-dimensional spaces or may explore other possibilities of molecular graph similarity measurements rather than similarity coefficients. Certainly, chemical-genetic interaction profiles of compounds can still play a major role as the gold functional standard in such a

system, and we believe a more comprehensive benchmarking system may elucidate novel metrics for chemical similarity measurements. For the machine learning pipeline, future work could focus on more sophisticated machine learning models such as deep neural network models with the aim of achieving larger performance for functional predictions or discovering novel scaffolds. For our integrated target prediction pipeline, a researcher may combine every informative signal from the existing pools of chemical-genetic interactions or other related signals to refine the target score measure. Although our current ranking system combines signals of all different pools to prioritize the most promising compound–gene pairs for experimental evaluations, our target score primarily relies on the chemical-genetic interactions related to essential genes. For the diffusion and network-based target prediction pipeline, a researcher may improve this pipeline by accurately tuning the sizes of the rumor graphs for compounds, refining the gene sets used for constructing the rumor graphs of compounds, incorporating meaningful weights to the edges of the rumor graphs, incorporating other related biochemical measures to the rumor graphs, and evaluating gene target predictions in experimental settings. In general, the computational frameworks that we developed in this dissertation have substantially enhanced functional genomics-based drug discovery approaches.

References

1. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **22**, 151–185 (2003).
   DOI: 10.1016/S0167-6296(02)00126-1.

2. Adams, C. P. & Brantner, V. V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? *Health Affairs* **25**, 420–428 (2006).
   DOI: 10.1377/hlthaff.25.2.420.

3. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33 (2016).
   DOI: 10.1016/j.jhealeco.2016.01.012.

4. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844 (2020).
   DOI: 10.1001/jama.2020.1166.

5. Mestre-Ferrandiz, J., Sussex, J. & Towse, A. *The R&D Cost of A New Medicine*. (Office of Health Economics, London, UK, 2012).
   URL: https://www.ohe.org/publications/rd-cost-new-medicine

6. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
   DOI: 10.1093/biostatistics/kxx069.

7. Issa, A. *et al.* Drug Withdrawals in the United States: A Systematic Review of the Evidence and Analysis of Trends. *Current Drug Safety* **2**, 177–185 (2007). DOI: 10.2174/157488607781668855.

8. Qureshi, Z. P., Seoane-Vazquez, E., Rodriguez-Monguio, R., Stevenson, K. B. & Szeinbach, S. L. Market withdrawal of new molecular entities approved in the United States from 1980 to 2009. *Pharmacoepidemiology and Drug Safety* **20**, 772–777 (2011). DOI: 10.1002/pds.2155.

9. Onakpoya, I. J., Heneghan, C. J. & Aronson, J. K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine* **14**, 10 (2016). DOI: 10.1186/s12916-016-0553-2.

10. Neu, H. C. The Crisis in Antibiotic Resistance. *Science* **257**, 1064–1073 (1992). DOI: 10.1126/science.257.5073.1064.

11. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews* **74**, 417–433 (2010). DOI: 10.1128/MMBR.00016-10.

12. Laxminarayan, R. *et al.* Antibiotic resistance—the need for global solutions. *The Lancet Infectious Diseases* **13**, 1057–1098 (2013). DOI: 10.1016/S1473-3099(13)70318-9.

13. Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P & T : a peer-reviewed journal for formulary management* **40**, 277–83 (2015). PMID: 25859123, PMCID: PMC4378521.

14. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology* **13**, 42–51 (2015).
DOI: 10.1038/nrmicro3380.

15. Murray, C. J. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
DOI: 10.1016/S0140-6736(21)02724-0.

16. Perlin, D. S., Rautemaa-Richardson, R. & Alastruey-Izquierdo, A. The global problem of antifungal resistance: prevalence, mechanisms, and management. *The Lancet Infectious Diseases* **17**, e383–e392 (2017).
DOI: 10.1016/S1473-3099(17)30316-X.

17. Fisher, M. C., Hawkins, N. J., Sanglard, D. & Gurr, S. J. Worldwide emergence of resistance to antifungal drugs challenges human health and food security. *Science* **360**, 739–742 (2018).
DOI: 10.1126/science.aap7999.

18. Revie, N. M., Iyer, K. R., Robbins, N. & Cowen, L. E. Antifungal drug resistance: evolution, mechanisms and impact. *Current Opinion in Microbiology* **45**, 70–76 (2018).
DOI: 10.1016/j.mib.2018.02.005.

19. Berman, J. & Krysan, D. J. Drug resistance and tolerance in fungi. *Nature Reviews Microbiology* **18**, 319–331 (2020).
DOI: 10.1038/s41579-019-0322-2.

20. Lee, Y., Puumala, E., Robbins, N. & Cowen, L. E. Antifungal Drug Resistance: Molecular Mechanisms in Candida albicans and Beyond. *Chemical Reviews* **121**, 3390–3411 (2021).
DOI: 10.1021/acs.chemrev.0c00199.

21. Moscona, A. Global Transmission of Oseltamivir-Resistant Influenza. *New England Journal of Medicine* **360**, 953–956 (2009).
DOI: 10.1056/NEJMp0900648.

22. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science* **328**, 1272–1275 (2010).
DOI: 10.1126/science.1187816.

23. Strasfeld, L. & Chou, S. Antiviral Drug Resistance: Mechanisms and Clinical Implications. *Infectious Disease Clinics of North America* **24**, 809–833 (2010).
DOI: 10.1016/j.idc.2010.07.001.

24. Mason, S., Devincenzo, J. P., Toovey, S., Wu, J. Z. & Whitley, R. J. Comparison of antiviral resistance across acute and chronic viral infections. *Antiviral Research* **158**, 103–112 (2018).
DOI: 10.1016/j.antiviral.2018.07.020.

25. Vitiello, A. Sars-Cov-2 and risk of antiviral drug resistance. *Irish Journal of Medical Science* **191**, 2367–2368 (2022).
DOI: 10.1007/s11845-021-02820-y.

26. Kruh, G. D. Introduction to resistance to anticancer agents. *Oncogene* **22**, 7262–7264 (2003).
DOI: 10.1038/sj.onc.1206932.

27. Cao, X., Hou, J., An, Q., Assaraf, Y. G. & Wang, X. Towards the overcoming of anticancer drug resistance mediated by p53 mutations. *Drug Resistance Updates* **49**, 100671 (2020).
DOI: 10.1016/j.drup.2019.100671.

28. Nussinov, R., Tsai, C.-J. & Jang, H. Anticancer drug resistance: An update and perspective. *Drug Resistance Updates* **59**, 100796 (2021).
DOI: 10.1016/j.drup.2021.100796.

29. Antimicrobial Resistance. *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance (2021).

30. Turner, N. C. & Reis-Filho, J. S. Genetic heterogeneity and cancer drug resistance. *The Lancet Oncology* **13**, e178–e185 (2012).
DOI: 10.1016/S1470-2045(11)70335-7.

31. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **13**, 714–726 (2013).
DOI: 10.1038/nrc3599.

32. Housman, G. *et al.* Drug Resistance in Cancer: An Overview. *Cancers* **6**, 1769–1792 (2014).
DOI: 10.3390/cancers6031769.

33. Nikolaou, M., Pavlopoulou, A., Georgakilas, A. G. & Kyrodimos, E. The challenge of drug resistance in cancer treatment: a current overview. *Clinical & Experimental Metastasis* **35**, 309–318 (2018).
DOI: 10.1007/s10585-018-9903-0.

34. Arora, A. & Scholar, E. M. Role of Tyrosine Kinase Inhibitors in Cancer Therapy. *Journal of Pharmacology and Experimental Therapeutics* **315**, 971–979 (2005). DOI: 10.1124/jpet.105.084145.

35. Barker, A. J. *et al.* Studies leading to the identification of ZD1839 (iressa^{TM}): an orally active, selective epidermal growth factor receptor tyrosine kinase inhibitor targeted to the treatment of cancer. *Bioorganic & Medicinal Chemistry Letters* **11**, 1911–1914 (2001). DOI: 10.1016/S0960-894X(01)00344-4.

36. Capdeville, R., Buchdunger, E., Zimmermann, J. & Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nature Reviews Drug Discovery* **1**, 493–502 (2002). DOI: 10.1038/nrd839.

37. Wilhelm, S. *et al.* Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nature Reviews Drug Discovery* **5**, 835–844 (2006). DOI: 10.1038/nrd2130.

38. Chow, L. Q. M. & Eckhardt, S. G. Sunitinib: From Rational Design to Clinical Efficacy. *Journal of Clinical Oncology* **25**, 884–896 (2007). DOI: 10.1200/JCO.2006.06.3602.

39. von Itzstein, M. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363**, 418–423 (1993). DOI: 10.1038/363418a0.

40. De Clercq, E. Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery* **1**, 13–25 (2002). DOI: 10.1038/nrd703.

41. Dorr, P. *et al.* Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity. *Antimicrobial Agents and Chemotherapy* **49**, 4721–4732 (2005).
DOI: 10.1128/aac.49.11.4721-4732.2005.

42. MacArthur, R. D. & Novak, R. M. Maraviroc: The First of a New Class of Antiretroviral Agents. *Clinical Infectious Diseases* **47**, 236–241 (2008).
DOI: 10.1086/589289.

43. Hazuda, D. J. *et al.* Inhibitors of Strand Transfer That Prevent Integration and Inhibit HIV-1 Replication in Cells. *Science* **287**, 646–650 (2000).
DOI: 10.1126/science.287.5453.646.

44. Pommier, Y., Johnson, A. A. & Marchand, C. Integrase inhibitors to treat HIV/Aids. *Nature Reviews Drug Discovery* **4**, 236–248 (2005).
DOI: 10.1038/nrd1660.

45. Summa, V. *et al.* Discovery of Raltegravir, a Potent, Selective Orally Bioavailable HIV-Integrase Inhibitor for the Treatment of HIV-AIDS Infection. *Journal of Medicinal Chemistry* **51**, 5843–5855 (2008).
DOI: 10.1021/jm800245z.

46. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery* **16**, 531–543 (2017).
DOI: 10.1038/nrd.2017.111.

47. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nature Reviews Drug Discovery* **10**, 507–519 (2011).
DOI: 10.1038/nrd3480.

48. Ford, C., Zurenko, G. & Barbachyn, M. The Discovery of Linezolid, the First Oxazolidinone Antibacterial Agent. *Current Drug Target -Infectious Disorders* **1**, 181–199 (2001).
DOI: 10.2174/1568005014606099.

49. Marks, P. A. & Breslow, R. Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. *Nature Biotechnology* **25**, 84–90 (2007).
DOI: 10.1038/nbt1272.

50. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**, 680–686 (1997).
DOI: 10.1126/science.278.5338.680.

51. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868 (1998).
DOI: 10.1073/pnas.95.25.14863.

52. Cho, R. J. *et al.* A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell* **2**, 65–73 (1998).
DOI: 10.1016/S1097-2765(00)80114-8.

53. Gasch, A. P. *et al.* Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell* **11**, 4241–4257 (2000).
DOI: 10.1091/mbc.11.12.4241.

54. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
DOI: 10.1038/35015701.

55. Hughes, T. R. *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell* **102**, 109–126 (2000).
DOI: 10.1016/S0092-8674(00)00015-5.

56. Baryshnikova, A. *et al.* Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Methods in enzymology* **470**, 145–79 (2010).
DOI: 10.1016/S0076-6879(10)70007-0.

57. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**, 855–867 (2008).
DOI: 10.1038/nrg2452.

58. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics* **8**, 437–449 (2007).
DOI: 10.1038/nrg2085.

59. Tong, A. H. Y. *et al.* Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science* **294**, 2364–2368 (2001).
DOI: 10.1126/science.1065810.

60. Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B. & Boone, C. Systematic Mapping of Genetic Interaction Networks. *Annual Review of Genetics* **43**, 601–625 (2009).
DOI: 10.1146/annurev.genet.39.073003.114751.

61. Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods* **7**, 1017–1024 (2010).
DOI: 10.1038/nmeth.1534.

62. van Leeuwen, J. *et al.* Exploring genetic suppression interactions on a global scale. *Science* **354**, (2016).
DOI: 10.1126/science.aag0839.

63. Tong, A. H. Y. & Boone, C. Synthetic Genetic Array Analysis in (ed. Xiao, W.) Saccharomyces cerevisiae. in *Yeast Protocols, Methods in Molecular Biology, vol 313*, 171–191 (Humana Press, Totowa, NJ, 2006).
DOI: 10.1385/1-59259-958-3:171, Print ISBN: 978-1-58829-437-1, Online ISBN: 978-1-59259-958-5.

64. Costanzo, M. *et al.* The Genetic Landscape of a Cell. *Science* **327**, 425–431 (2010).
DOI: 10.1126/science.1180823.

65. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, 1381 (2016).
DOI: 10.1126/science.aaf1420.

66. Giaever, G. *et al.* Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proceedings of the National Academy of Sciences* **101**, 793–798 (2004).
DOI: 10.1073/pnas.0307490100.

67. Parsons, A. B. *et al.* Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast. *Cell* **126**, 611–625 (2006).
DOI: 10.1016/j.cell.2006.06.040.

68. Pierce, S. E., Davis, R. W., Nislow, C. & Giaever, G. Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. *Nature Protocols* **2**, 2958–2974 (2007).
DOI: 10.1038/nprot.2007.427.

69. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Research* **19**, 1836–1842 (2009).
DOI: 10.1101/gr.093955.109.

70. Smith, A. M. *et al.* Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research* **38**, e142–e142 (2010).
DOI: 10.1093/nar/gkq368.

71. Parsons, A. B. *et al.* Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature Biotechnology* **22**, 62–69 (2004).
DOI: 10.1038/nbt919.

72. Deshpande, R. *et al.* Efficient strategies for screening large-scale genetic interaction networks. *bioRxiv* (2017).
DOI: 10.1101/159632.

73. Piotrowski, J. S. *et al.* Functional annotation of chemical libraries across diverse biological processes. *Nature Chemical Biology* **13**, 982–993 (2017).
DOI: 10.1038/nchembio.2436.

74. Simpkins, S. W. *et al.* Predicting bioprocess targets of chemical compounds through integration of chemical-genetic and genetic interactions. *PLOS Computational Biology* **14**, e1006532 (2018).
DOI: 10.1371/journal.pcbi.1006532.

75. Simpkins, S. W. *et al.* Using BEAN-counter to quantify genetic interactions from multiplexed barcode sequencing experiments. *Nature Protocols* **14**, 415–440 (2019). DOI: 10.1038/s41596-018-0099-1.

76. Nelson, J. *et al.* MOSAIC: a chemical-genetic interaction data repository and web resource for exploring chemical modes of action. *Bioinformatics* **34**, 1251–1252 (2018). DOI: 10.1093/bioinformatics/btx732.

77. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today* **3**, 160–178 (1998). DOI: 10.1016/S1359-6446(97)01163-X.

78. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004). DOI: 10.1038/nature03197.

79. McInnes, C. Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology* **11**, 494–502 (2007). DOI: 10.1016/j.cbpa.2007.08.033.

80. Lavecchia, A. & Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry* **20**, 2839–2860 (2013). DOI: 10.2174/09298673113209990001.

81. Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **7**, 1047–1055 (2002). DOI: 10.1016/S1359-6446(02)02483-2.

82. Cheng, T., Li, Q., Zhou, Z., Wang, Y. & Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal* **14**, 133–141 (2012).
DOI: 10.1208/s12248-012-9322-0.

83. Lionta, E., Spyrou, G., Vassilatis, D. & Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry* **14**, 1923–1938 (2014).
DOI: 10.2174/1568026614666140929124445.

84. Stahura, F. L. & Bajorath, J. New Methodologies for Ligand-Based Virtual Screening. *Current Pharmaceutical Design* **11**, 1189–1202 (2005).
DOI: 10.2174/1381612053507549.

85. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **38**, 983–996 (1998).
DOI: 10.1021/ci9800211.

86. Irwin, J. J. & Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177–82.
DOI: 10.1021/ci049714+.

87. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, S55–S57 (2018).
DOI: 10.1038/d41586-018-05267-x.

88. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **23**, 1241–1250 (2018).
DOI: 10.1016/j.drudis.2018.01.039.

89. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **19**, 353–364 (2020).
DOI: 10.1038/s41573-019-0050-3.

90. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
DOI: 10.1038/s41586-021-03819-2.

91. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59**, 3370–3388 (2019).
DOI: 10.1021/acs.jcim.9b00237.

92. Ciallella, H. L. & Zhu, H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chemical Research in Toxicology* **32**, 536–547 (2019).
DOI: 10.1021/acs.chemrestox.8b00393.

93. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).
DOI: 10.1016/j.cell.2020.01.021.

94. Jin, W. *et al.* Deep learning identifies synergistic drug combinations for treating COVID-19. *Proceedings of the National Academy of Sciences* **118**, (2021).
DOI: 10.1073/pnas.2105070118.

95. Giaever, G. *et al.* Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics* **21**, 278–283 (1999).
DOI: 10.1038/6791.

96. Sopko, R. *et al.* Mapping Pathways and Phenotypes by Systematic Gene Overexpression. *Molecular Cell* **21**, 319–330 (2006).
DOI: 10.1016/j.molcel.2005.12.011.

97. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* **11**, 191–200 (2012).
DOI: 10.1038/nrd3681.

98. Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery* **8**, 959–968 (2009).
DOI: 10.1038/nrd2961.

99. Johnson, M. A. & Maggiora, G. M. *Concepts and applications of molecular similarity*. (Wiley, 1990).
isbn-10: 0471621757, isbn-13: 978-0-471-62175-1.

100. Bajorath, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery* **1**, 882–894 (2002).
DOI: 10.1038/nrd941.

101. Tanrikulu, Y., Krüger, B. & Proschak, E. The holistic integration of virtual screening in drug discovery. *Drug Discovery Today* **18**, 358–364 (2013).
DOI: 10.1016/j.drudis.2013.01.007.

102. Martin, Y. C., Kofron, J. L. & Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *Journal of Medicinal Chemistry* **45**, 4350–4358 (2002).
DOI: 10.1021/jm020155c.

103. Eckert, H. & Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **12**, 225–233 (2007).
DOI: 10.1016/j.drudis.2007.01.011.

104. Maggiora, G. M. On Outliers and Activity Cliffs–Why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **46**, 1535 (2006).
DOI: 10.1021/ci060117s.

105. Bajorath, J. *et al.* Navigating structure-activity landscapes. *Drug Discovery Today* **14**, 698–705 (2009).
DOI: 10.1016/j.drudis.2009.04.003.

106. Muratov, E. N. *et al.* QSAR without borders. *Chemical Society Reviews* **49**, 3525–3564 (2020).
DOI: 10.1039/D0CS00098A.

107. Bender, A. & Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry* **2**, 3204 (2004).
DOI: 10.1039/b409813g.

108. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **57**, 3186–3204 (2014).
DOI: 10.1021/jm401411z.

109. Willett, P. Similarity Searching Using 2D Structural Fingerprints. in (ed. Bajorath, J.) *Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology, vol 672*, 133–158 (Humana Press, Totowa, NJ, 2010).
DOI: 10.1007/978-1-60761-839-3_5, Print ISBN: 978-1-60761-838-6, Online ISBN: 978-1-60761-839-3.

110. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling* **29**, 157–170 (2010). DOI: 10.1016/j.jmgm.2010.05.008.

111. Stumpfe, D. & Bajorath, J. Similarity searching. *WIREs Computational Molecular Science* **1**, 260–282 (2011). DOI: 10.1002/wcms.23.

112. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015). DOI: 10.1016/j.ymeth.2014.08.005.

113. Muegge, I. & Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery* **11**, 137–148 (2016). DOI: 10.1517/17460441.2016.1117070.

114. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **11**, 1046–1053 (2006). DOI: 10.1016/j.drudis.2006.10.005.

115. Todeschini, R. *et al.* Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *Journal of Chemical Information and Modeling* **52**, 2884–2901 (2012). DOI: 10.1021/ci300261r.

116. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, 20 (2015). DOI: 10.1186/s13321-015-0069-3.

117. Rácz, A., Bajusz, D. & Héberger, K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of Cheminformatics* **10**, 48 (2018). DOI: 10.1186/s13321-018-0302-y.

118. Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspectives in Drug Discovery and Design* **7**–**8**, 65–84 (1996). DOI: 10.1007/BF03380182.

119. Dixon, S. L. & Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *Journal of Medicinal Chemistry* **42**, 2887–2900 (1999). DOI: 10.1021/jm980708c.

120. Holliday, J. D., Salim, N., Whittle, M. & Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *Journal of Chemical Information and Computer Sciences* **43**, 819–828 (2003). DOI: 10.1021/ci034001x.

121. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002). DOI: 10.1038/nature00935.

122. Hert, J. *et al.* New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *Journal of Chemical Information and Modeling* **46**, 462–470 (2006). DOI: 10.1021/ci050348j.

123. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **23**, 1538–1546 (2018). DOI: 10.1016/j.drudis.2018.05.010.

124. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **18**, 463–477 (2019).
DOI: 10.1038/s41573-019-0024-5.

125. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials* **18**, 435–441 (2019).
DOI: 10.1038/s41563-019-0338-z.

126. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).
DOI: 10.1007/BF00994018.

127. Vapnik, V. N. The Nature of Statistical Learning Theory. (Springer New York, NY, 1999).
DOI: 10.1007/978-1-4757-3264-1, Hardcover ISBN: 978-0-387-98780-4, Softcover ISBN: 978-1-4419-3160-3 (2010), eBook ISBN: 978-1-4757-3264-1 (2013).

128. Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N. & Zell, A. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. *Journal of Cheminformatics* **3**, 3 (2011).
DOI: 10.1186/1758-2946-3-3.

129. Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org (2006).

130. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Modeling* **25**, 64–73 (1985).
DOI: 10.1021/ci00046a002.

131. Borgwardt, K. M. & Kriegel, H. Shortest-Path Kernels on Graphs. in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 74–81 (2005).
DOI: 10.1109/ICDM.2005.132.

132. Ralaivola, L., Swamidass, S. J., Saigo, H. & Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **18**, 1093–1110 (2005).
DOI: 10.1016/j.neunet.2005.07.009.

133. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).
DOI: 10.1021/ci100050t.

134. Xing, L. & Glen, R. C. Novel Methods for the Prediction of logP, pKa, and logD. *Journal of Chemical Information and Computer Sciences* **42**, 796–805 (2002).
DOI: 10.1021/ci010315d.

135. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **42**, 1273–1280 (2002).
DOI: 10.1021/ci010132r.

136. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. 'Scaffold-Hopping' by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie - International Edition* **38**, 2894–2896 (1999).
DOI: 10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F

137. Daylight Chemical Information Systems, Inc. http://www.daylight.com.

138. Raymond, J. W. & Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design* **16**, 59–71 (2002).
DOI: 10.1023/A:1016387816342.

139. Braun-Blanquet, J. *Plant sociology. The study of plant communities*. (McGraw-Hill Book Co., Inc., New York and London, 1932).

140. Zimmerman, B. L. Audio strip transects. in (eds. Heyer, R. W., Donnelly, M. A., McDiarmid, R. W., Hayek, L.-A. C. & Foster, M. S.) *Measuring and monitoring biological diversity: Standard methods for amphibians,* 92-96 (Smithsonian Institution Press, Washington, DC, 1994).
ISBN-10: 1560982845, ISBN-13: 978-1560982845.

141. Driver, H. E. & Kroeber, A. L. Quantitative expression of cultural relationship. *University of California Publications in American Archaeology and Ethnology, vol 31,* 211−256 (University of California Press, Berkeley, CA, 1932).
URL: https://digitalassets.lib.berkeley.edu/anthpubs/ucb/text/ucp031-005.pdf

142. Ochiai, A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries* **22**, 526–530 (1957).

143. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945).
DOI: 10.2307/1932409.

144. Sorensen, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons. *Kongelige Danske Viden-skabernes Selskabs Biologiske Skrifter* **5**, 1–34 (1948).

145. Kulczynski, S. Die Pflanzenassociationen der Pienenen. *Bulletin international de l'Académie polonaise des sciences et des lettres. Classe des sciences mathématiques et naturelles. Série B: Sciences naturelles. Supplément II.* **2**, 57–203 (1927).

146. McConnaughey, B. H. The Determination and Analysis of Plankton Communities. *Marine Research in Indonesia, Special No* 1–40 (1964).

147. Russell, P. F. & Rao, T. R. On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras. **3**, 153–178 (1940).

148. Simpson, G. G. Mammals and the nature of continents. *American Journal of Science* **241**, 1–31 (1943).
DOI: 10.2475/ajs.241.1.1.

149. Sokal, R. R. & Sneath, P. H. A. Principles of numerical taxonomy. *San Francisco: W. H. Friedman and Company* (1963).

150. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles* **37**, 547–579 (1901).

151. Tanimoto, T. T. *Elementary Mathematical theory of Classification and Prediction. IBM Corporation*. (1958).

152. Tulloss, R. E. Assessment of Similarity Indices for Undesirable Properties and a New Tripartite Similarity Index Based on Cost Functions. in (eds. Palm, M. E. & Chapela, I. H.) *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*, 122–143 (Parkway Publishers, Inc., 1997).

153. Tversky, A. Features of similarity. *Psychological Review* **84**, 327–352 (1977). DOI: 10.1037/0033-295X.84.4.327.

154. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5**, 107–113 (1965). DOI: 10.1021/c160017a018.

155. Bender, A. *et al.* Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs 9, 199–204 (2006). PMID: 16523386

156. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* **5**, 26 (2013). DOI: 10.1186/1758-2946-5-26.

157. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics* **8**, 36 (2016). DOI: 10.1186/s13321-016-0148-0.

158. Héberger, K. Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry* **29**, 101–109 (2010). DOI: 10.1016/j.trac.2009.09.009.

159. Kollár-Hunek, K. & Héberger, K. Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemometrics and Intelligent Laboratory Systems* **127**, 139–146 (2013).
DOI: 10.1016/j.chemolab.2013.06.007.

160. Héberger, K. & Kollár-Hunek, K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *Journal of Chemometrics* **25**, 151–158 (2011).
DOI: 10.1002/cem.1320.

161. Barshan, E., Ghodsi, A., Azimifar, Z. & Zolghadri Jahromi, M. Supervised principal component analysis Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition* **44**, 1357–1371 (2011).
DOI: 10.1016/j.patcog.2010.12.015.

162. Efron, B. & Tibshirani, R. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association* **92**, 548–560 (1997).
DOI: 10.1080/01621459.1997.10474007.

163. Deshpande, R., VanderSluis, B. & Myers, C. L. Comparison of Profile Similarity Measures for Genetic Interaction Networks. *PLoS ONE* **8**, e68664 (2013).
DOI: 10.1371/journal.pone.0068664.

164. Chang, C.-C. & Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 27 (2011).
DOI: 10.1145/1961189.1961199.

165. Roemer, T. & Boone, C. Systems-level antimicrobial drug and drug synergy discovery. *Nature Chemical Biology* **9**, 222–231 (2013).
DOI: 10.1038/nchembio.1205.

166. Deshpande, R., Nelson, J., Myers, C. L. & Simpkins, S. W. Screening of large-scale genetic interaction networks. *US Patent 11017880B2* (2017).

167. Tukey, J. W. *Exploratory Data Analysis*. (Addison-Wesley, 1977).
DOI: 10.1002/bimj.4710230408

168. Douglas, C. M. *et al.* The Saccharomyces cerevisiae FKS1 (ETG1) gene encodes an integral membrane protein which is a subunit of 1,3-beta-D-glucan synthase. *Proceedings of the National Academy of Sciences* **91**, 12907–12911 (1994).
DOI: 10.1073/pnas.91.26.12907.

169. Safizadeh, H. *et al.* Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical–Genetic Interactions. *Journal of Chemical Information and Modeling* **61**, 4156–4172 (2021).
DOI: 10.1021/acs.jcim.0c00993.

170. Zhou, F. *et al.* Integrating yeast chemical genomics and mammalian cell pathway analysis. *Acta Pharmacologica Sinica* **40**, 1245–1255 (2019).
DOI: 10.1038/s41401-019-0231-y.

171. Bagcchi, S. WHO's Global Tuberculosis Report 2022. *The Lancet Microbe* **4**, e20 (2023).
DOI: 10.1016/S2666-5247(22)00359-7.

172. Global Tuberculosis Report. *World Health Organization*. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (2022).

173. Chee, C. B. E., Reves, R., Zhang, Y. & Belknap, R. Latent tuberculosis infection: Opportunities and challenges. *Respirology* **23**, 893–900 (2018).
DOI: 10.1111/resp.13346.

174. Young, C., Walzl, G. & Du Plessis, N. Therapeutic host-directed strategies to improve outcome in tuberculosis. *Mucosal Immunology* **13**, 190–204 (2020). DOI: 10.1038/s41385-019-0226-5.

175. Paik, S., Kim, J. K., Chung, C. & Jo, E.-K. Autophagy: A new strategy for host-directed therapy of tuberculosis. *Virulence* **10**, 448–459 (2019). DOI: 10.1080/21505594.2018.1536598.

176. Persaud, R. *et al.* Clionamines stimulate autophagy, inhibit Mycobacterium tuberculosis survival in macrophages, and target Pik1. *Cell Chemical Biology* **29**, 870-882.e11 (2022). DOI: 10.1016/j.chembiol.2021.07.017.

177. Lee, A. Y. *et al.* Mapping the Cellular Response to Small Molecules Using Chemogenomic Fitness Signatures. *Science* **344**, 208–211 (2014). DOI: 10.1126/science.1250217.

178. Almeida, F., Rodrigues, M. L. & Coelho, C. The Still Underestimated Problem of Fungal Diseases Worldwide. *Frontiers in Microbiology* **10**, (2019). DOI: 10.3389/fmicb.2019.00214.

179. Fisher, M. C. *et al.* Threats Posed by the Fungal Kingdom to Humans, Wildlife, and Agriculture. *mBio* **11**, e00449-20 (2020). DOI: 10.1128/mBio.00449-20.

180. Xu, J. Assessing global fungal threats to humans. *mLife* **1**, 223–240 (2022). DOI: 10.1002/mlf2.12036.

181. Bongomin, F., Gago, S., Oladele, R. & Denning, D. Global and Multi-National Prevalence of Fungal Diseases—Estimate Precision. *Journal of Fungi* **3**, 57 (2017). DOI: 10.3390/jof3040057.

182. Rayens, E. & Norris, K. A. Prevalence and Healthcare Burden of Fungal Infections in the United States, 2018. *Open Forum Infectious Diseases* **9**, 1–12 (2022). DOI: 10.1093/ofid/ofab593.

183. Dixon, D. M. & Walsh, T. J. Antifungal Agents. in (ed. Baron, S.) *Medical Microbiology*, Chapter 76 (University of Texas Medical Branch at Galveston, 1996).

184. Chen, S. C. A. & Sorrell, T. C. Antifungal agents. *Medical Journal of Australia* **187**, 404–409 (2007). DOI: 10.5694/j.1326-5377.2007.tb01313.x.

185. Rhodes, J. & Fisher, M. C. Global epidemiology of emerging Candida auris. *Current Opinion in Microbiology* **52**, 84–89 (2019). DOI: 10.1016/j.mib.2019.05.008.

186. Chayakulkeeree, M., Rude, T. H., Toffaletti, D. L. & Perfect, J. R. Fatty Acid Synthesis Is Essential for Survival of Cryptococcus neoformans and a Potential Fungicidal Target. *Antimicrobial Agents and Chemotherapy* **51**, 3537–3545 (2007). DOI: 10.1128/AAC.00442-07.

187. Iyer, K. R., Revie, N. M., Fu, C., Robbins, N. & Cowen, L. E. Treatment strategies for cryptococcal infection: challenges, advances and future outlook. *Nature Reviews Microbiology* **19**, 454–466 (2021). DOI: 10.1038/s41579-021-00511-0.

188. Shah, D. & Zaman, T. Rumors in a Network: Who's the Culprit? *IEEE Transactions on Information Theory* **57**, 5163–5181 (2011).
DOI: 10.1109/TIT.2011.2158885.

189. Davidse, L. C. Benzimidazole Fungicides: Mechanism of Action and Biological Impact. *Annual Review of Phytopathology* **24**, 43–65 (1986).
DOI: 10.1146/annurev.py.24.090186.000355.

Appendix A: Supporting Information for Chapter 2

Algorithm A.1. Algorithm for finding the best performing structural similarity measures

	Steps for the systematic benchmarking of structural similarity measures based on chemical-genetic interaction data	Results (RIKEN high-confidence set)
	Selecting the best-performing molecular fingerprints	
1	Generate molecular fingerprints for the compound collections for which we have chemical-genetic interaction profiles.	
2	Establish the binarized gold standard for biological activity (10% most similar compound pairs based on chemical-genetic cosine similarity).	
3	Compute precision for each structural similarity measure at several recall thresholds. (Each structural similarity measure is defined as one molecular fingerprint paired with one similarity coefficient.)	
4	Select up to the top 3 fingerprints based on the highest precision at each recall threshold (precision values within 10% of the maximum were considered equivalent).	Recall = 0.002: ASP, DFS, RAD2D Recall = 0.005: ASP, LSTAR, RAD2D
	Top-performing molecular fingerprints: ASP, LSTAR, and RAD2D	Recall = 0.02: ASP, LSTAR, RAD2D Reall = 0.05: ASP, LSTAR, RAD2D Recall = 0.2: LSTAR
	Selecting the best-performing similarity coefficients	
1	Remove the similarity coefficients for which the precision was < 90% of the maximum precision achieved at the majority of recall thresholds for the majority of molecular fingerprints.	Exclude: Dot-product, Euclidean, Russel/Rao, Simpson
2	Select the similarity coefficients that appear in the top 3 in terms of precision at all recall thresholds (based on the best-performing molecular fingerprint for that similarity coefficient).	Recall = 0.002: Braun-Blanquet, Tullos Recall = 0.005: Braun-Blanquet, Tullos
	Top-performing similarity coefficients: Braun-Blanquet and Tullos	Recall = 0.02: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos Recall = 0.05: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos Recall = 0.2: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos

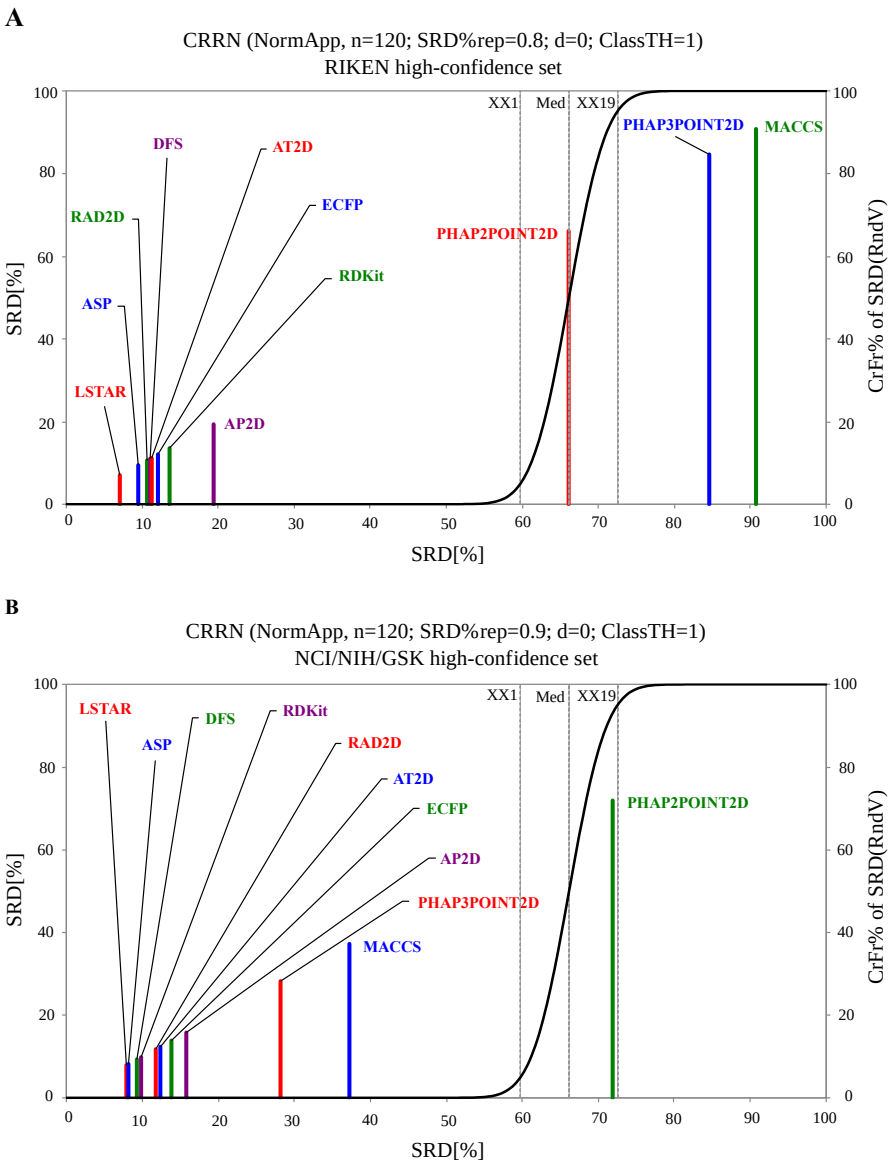


Figure A.1. SRD analysis for molecular fingerprints. Since our systematic benchmark compares many structural similarity measures/models, we used the SRD method/model comparison approach (72) as a multi-criteria decision making framework to ensure that we fairly compared the components of our similarity models. We downloaded, from http://aki.ttk.hu/srd, the Excel macro program file for the SRD analysis with ties (i.e., repeated observations) (73). We generated a matrix of our data that contained the precision at all predefined recall thresholds and the areas under the

ROC curves as two sets of criteria for evaluating the performance of our structural similarity models (120x11 data matrix by transposing and combining all spreadsheet tables from **Table A.1** or **Table A.2**; rows: 12 similarity coefficients measured at 9 recall thresholds and for the areas under the ROC curves; columns: 11 molecular fingerprints). We used only the fingerprints with a depth of 8 and the symmetric similarity coefficients (Tversky with $\alpha = 0.9$ was removed) for this SRD analysis. We used the maximum of the row values as the reference ranking because we searched for the molecular fingerprint with the highest performance. We validated the resulting sum of absolute ranking differences using the approximated normal distribution of random numbers. The Comparison of Ranks by Random Numbers (CRRN) validation (74) using **(A)** the RIKEN and **(B)** the NCI/NIH/GSK high-confidence sets confirmed the superiority of LSTAR and ASP fingerprints over all other molecular fingerprints. The colors were assigned to the bars on a rotational basis by the SRD program file.

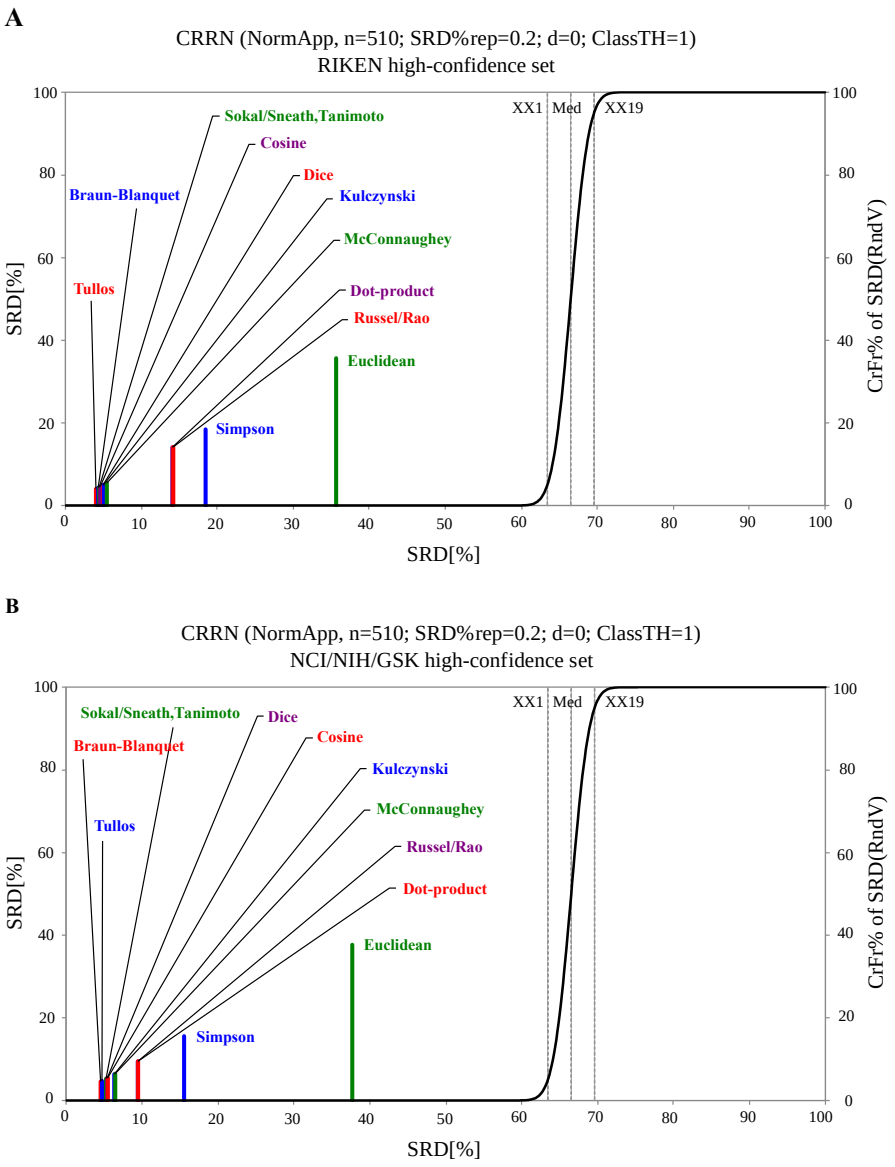


Figure A.2. SRD analysis for similarity coefficients. For each compound collection, we generated a matrix of our data (510x12 matrix by combining all spreadsheet tables from **Table A.1** or **Table A.2**; rows: 11 molecular fingerprints measured at all available depths, at 9 recall thresholds, and for the areas under the ROC curves; columns: 12 symmetric similarity coefficients). Since we included the Dice and Tanimoto similarity coefficients, which are two symmetric instances of the Tversky coefficient, we removed Tversky with $\alpha = 0.9$ from this SRD analysis. We used the

maximum of the row values as the reference ranking because we searched for the similarity coefficient with the highest performance. The CRRN validation for **(A)** the RIKEN and **(B)** the NCI/NIH/GSK high-confidence sets indicated the relative superiority of the Tullos and Braun-Blanquet coefficients over all other similarity coefficients. Although several similarity coefficients were competing for high performance based on the CRRN validations for our compound collections, the Dot-product, Euclidean, Russel/Rao, and Simpson similarity coefficients were distinctly the worst coefficients, whereas the Tullos and Braun-Blanquet similarity coefficients were the best ones. While the Tullos coefficient achieved comparable performance to the Braun-Blanquet coefficient, the simplicity of the Braun-Blanquet coefficient makes this coefficient preferable in most practical scenarios. We also repeated this SRD analysis for individual recall thresholds (51x12 data matrices), which again confirmed the superiority of the Braun-Blanquet and Tullos similarity coefficients over all other coefficients at our predefined recall thresholds (**Table A.1** or **Table A.2** – SRD analysis spreadsheets). The colors were assigned to the bars on a rotational basis by the SRD program file.

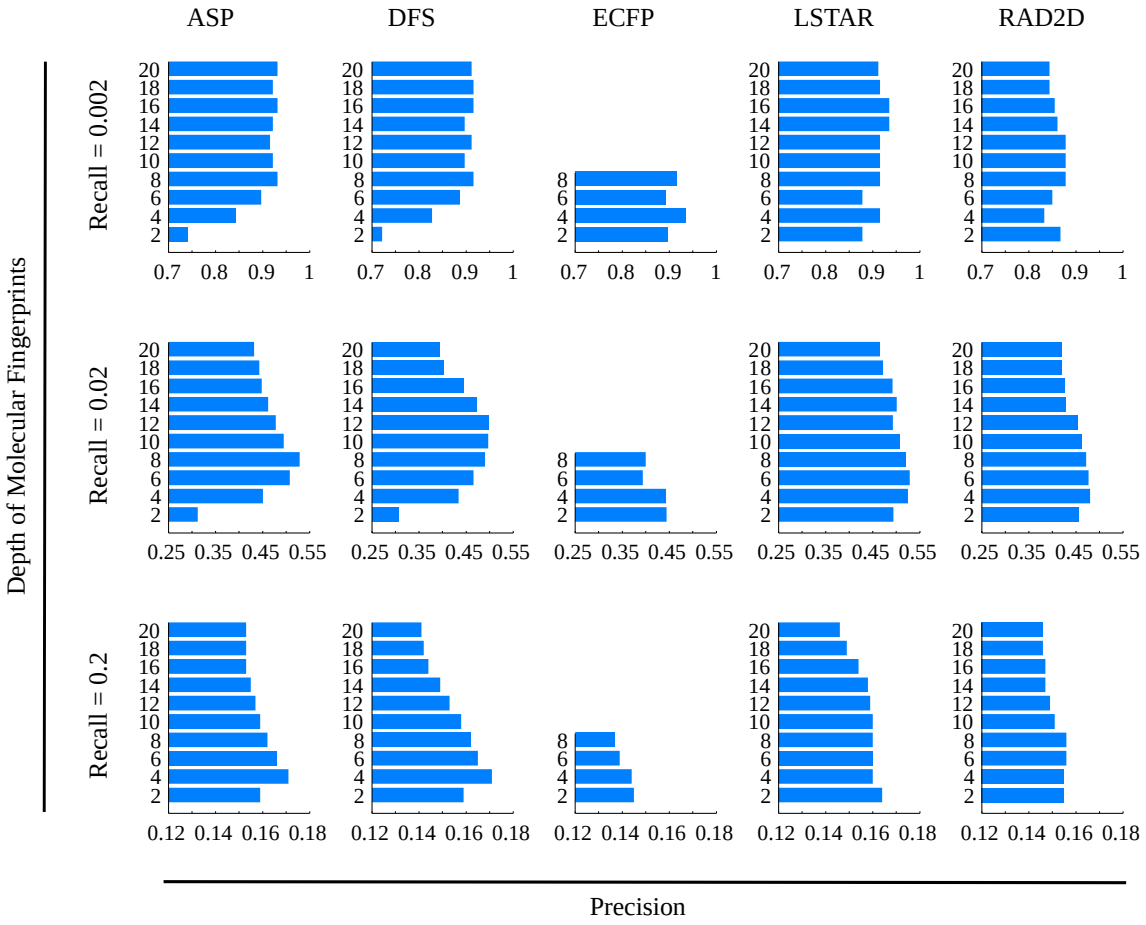
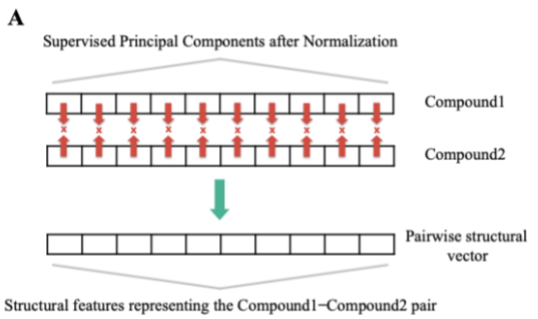


Figure A.3. Impact of the describing depth of molecular fingerprints on the NCI/NIH/GSK high-confidence set. We measured the precision of our prediction models at 10 molecular depths, ranging from 2 to 20, for five different molecular fingerprints. Similarities were calculated with the Braun-Blanquet similarity coefficient, and the precision at three different recall thresholds for each molecular depth is shown.



B

	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4	Mean
Pair 1	X	X		X	3 values
Pair 2	X	X	X	X	4 values
Pair 3			X		1 value
Pair 4	X	X			2 values
Pair 5			X	X	2 values

X: Both compounds in the pair are in the test set and consequently have a predicted chemical-genetic similarity in the bootstrap.

Figure A.4. Pairwise structural vectors and bootstrapping used by our machine learning pipeline. **(A)** We generated the pairwise structural features/vector for a compound pair by the element-wise multiplication of the normalized, low-dimensional structural vectors describing the compounds. **(B)** Bootstrap aggregating (bagging) example. For an example of five compound pairs and four bootstraps, we illustrate the bootstrap aggregating procedure: We averaged the predicted chemical-genetic similarities of a test pair (represented by X values in a row) over all bootstraps to generate the final prediction for the test pair. For example, compound pair 1 was a test pair in bootstraps 1, 2, and 4; therefore, we averaged these three predictions to form the final prediction for pair 1 (In bootstrap 3, compound pair 1 was either a training pair or an invalid pair for which one compound belonged to the training set and the other to the test set).

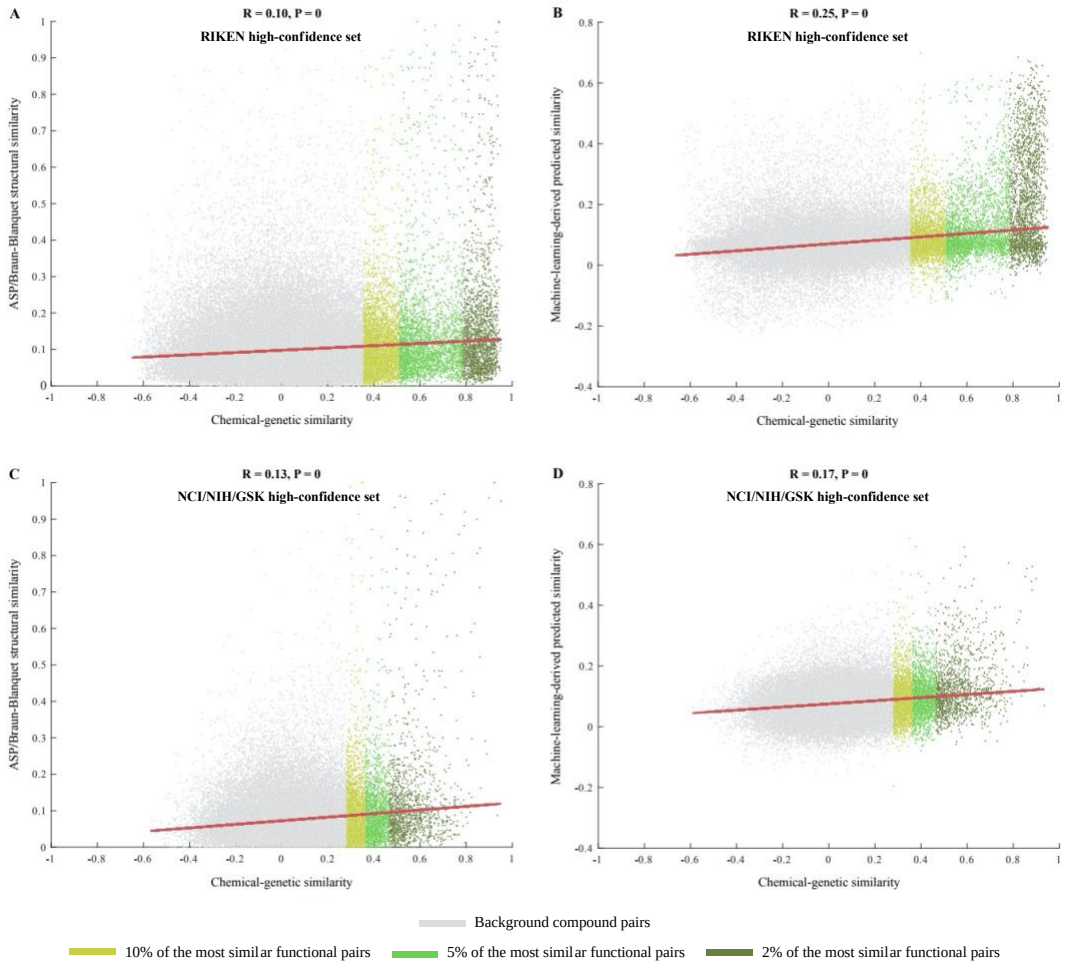


Figure A.5. Correlation analysis of the predicted structural similarities with the chemical-genetic similarities. Correlation of **(A)** the ASP/Braun-Blanquet-derived structural similarity and **(B)** the machine-learning-derived predicted similarity with chemical-genetic similarity for the RIKEN high-confidence set. Correlation of **(C)** the ASP/Braun-Blanquet-derived structural similarity and **(D)** the machine-learning-derived predicted similarity with chemical-genetic similarity for the NCI/NIH/GSK high-confidence set. We computed all machine-learning-derived predicted similarities from bootstrap aggregation, where we calculated the prediction for a compound pair as the mean of the model output over all the bootstraps for which the compound pair was in the test set (**Figure A.4B**). R and P are the Pearson correlation coefficient and the corresponding p-value, respectively.

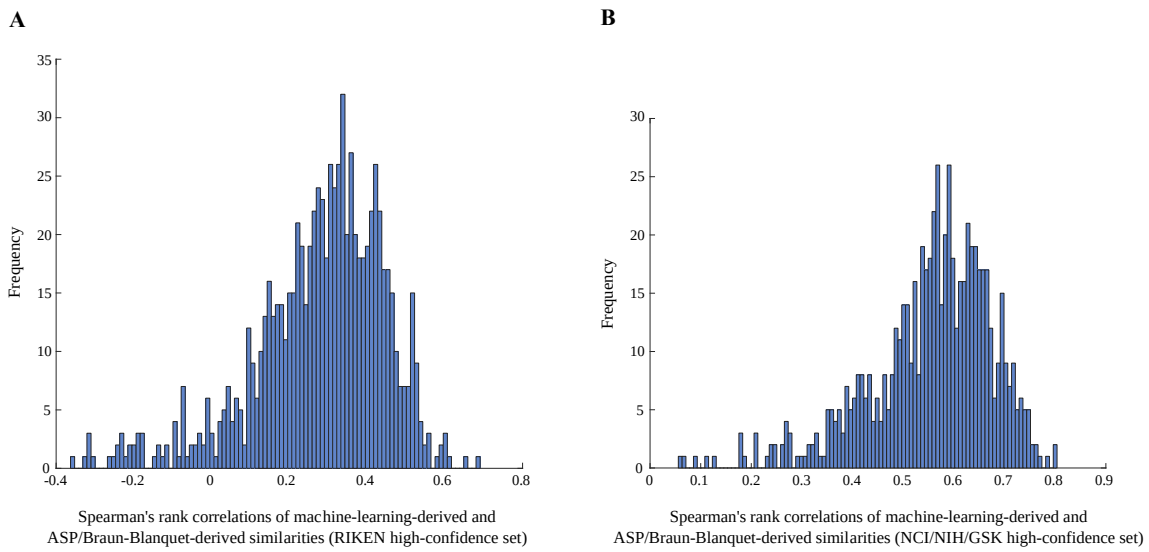


Figure A.6. Spearman's rank correlation distribution between the machine-learning-derived and the ASP/Braun-Blanquet-derived structural similarity predictions. We measured the Spearman's rank correlation distribution for **(A)** the RIKEN and **(B)** the NCI/NIH/GSK high-confidence sets. Each value represents the Spearman's rank correlation for a compound. To compute the correlation value for a compound, we measured the Spearman's rank correlation between two lists of predicted similarities, where one list was generated by our machine learning model and the other list by the Braun-Blanquet similarity coefficient, both using ASP fingerprints.

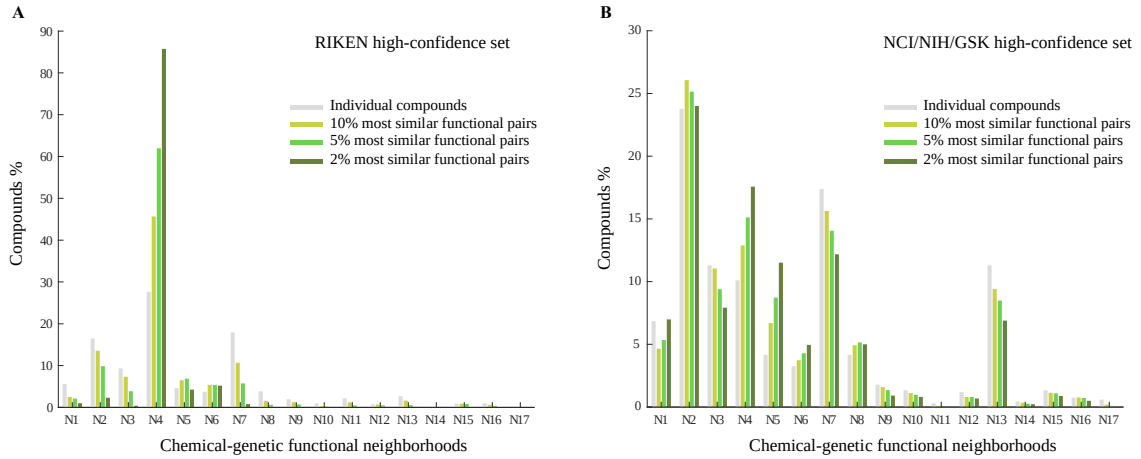


Figure A.7. Distribution of the predicted biological functions across 17 broad, previously defined functional neighborhoods (32,34) for **(A)** the RIKEN and **(B)** the NCI/NIH/GSK high-confidence sets. The gray bar heights represent the distribution of all the studied compounds in these sets, whereas the green bar heights represent the distribution of a subset of these compounds.

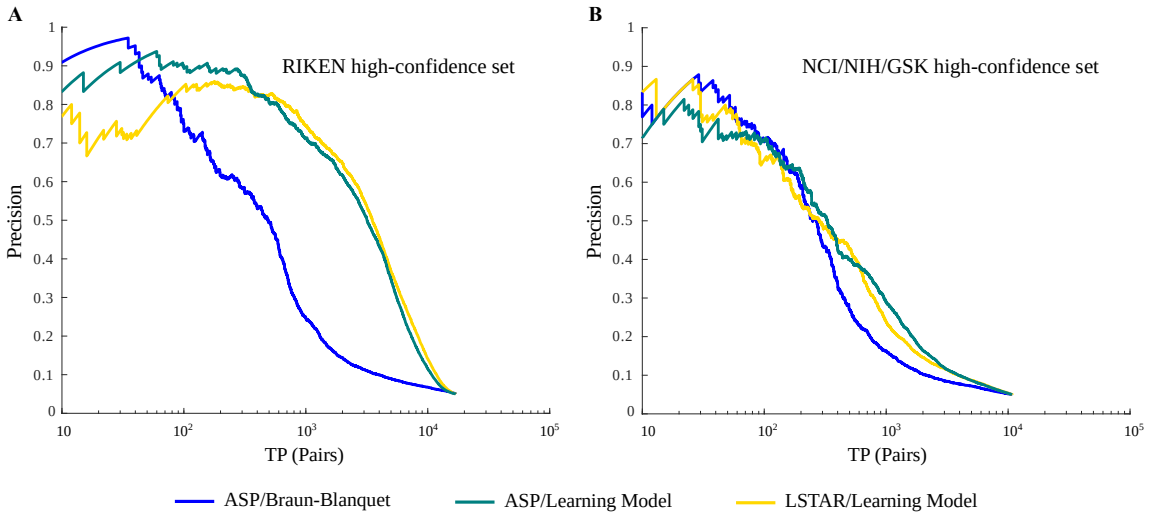


Figure A.8. Prediction performance of the machine learning models for the 5% cutoff (as a more stringent cutoff than 10%) on the functional similarity gold standard. **(A)** Model performance for the RIKEN high-confidence set. The blue PR curve represents the prediction performance gained by the best-performing structural similarity measure (ASP/Braun-Blanquet), whereas the teal and gold PR curves represent the performance of the machine learning models using ASP and LSTAR fingerprints, respectively. A prediction is considered a true positive if the compound pair is within the top 5% of functionally similar compound pairs using chemical-genetic interaction profiles. **(B)** Model performance for the NCI/NIH/GSK high-confidence set.

Table A.1. Precision at several recall thresholds for all the combinations of a molecular fingerprint and a similarity coefficients using the RIKEN high-confidence set (XLSX)

Table A.2. Precision at several recall thresholds for all the combinations of a molecular fingerprint and a similarity coefficients using the NCI/NIH/GSK high-confidence set (XLSX)

Table A.3. Functional clustering using the K-means algorithm and functional enrichment analysis using the hypergeometric test for the RIKEN high-confidence set (XLSX)

Table A.4. Functional clustering using the K-means algorithm and functional enrichment analysis using the hypergeometric test for the NCI/NIH/GSK high-confidence set (XLSX)

Table A.5. Structural clustering of our RIKEN and NCI/NIH/GSK high-confidence sets using the K-medoids algorithm (XLSX)

Table A.6. Ranked lists of compound pairs, selected from the top 1000 machine learning-derived true positive pairs, with similar chemical-genetic interaction profiles (all the compounds from the first functional cluster) but distinct chemical structures (as determined by the ASP/Braun-Blanquet structural similarity measure) for our RIKEN and NCI/NIH/GSK high-confidence sets (XLSX)

Table A.7. Comprehensive ranked lists of structural and chemical-genetic similarities for our RIKEN and NCI/NIH/GSK high-confidence sets (XLSX)

Structure Data Files – molecular structural information in SDF notations for our RIKEN and NCI/NIH/GSK high-confidence sets freely available in MOSAIC (http://mosaic.cs.umn.edu) (ZIP)

SMILES and InChI – molecular structural information in SMILES and InChI notations for our RIKEN and NCI/NIH/GSK freely available in MOSAIC (http://mosaic.cs.umn.edu) (XLSX)

Appendix B: Supporting Information for Chapter 3

Table B.1. Ranked lists of frequent flier genes and frequent flier compounds using the HET data for our RIKEN high-confidence set (XLSX)

Table B.2. Target score examples: data tables of the outlier analyses of NPD5728 and FRS1 after applying the specificity criteria of MSC ≤ 20 and MSG ≤ 20 and scoring these top data table entries using our target identification scoring system (XLSX)

Appendix C: Supporting Information for Chapter 4

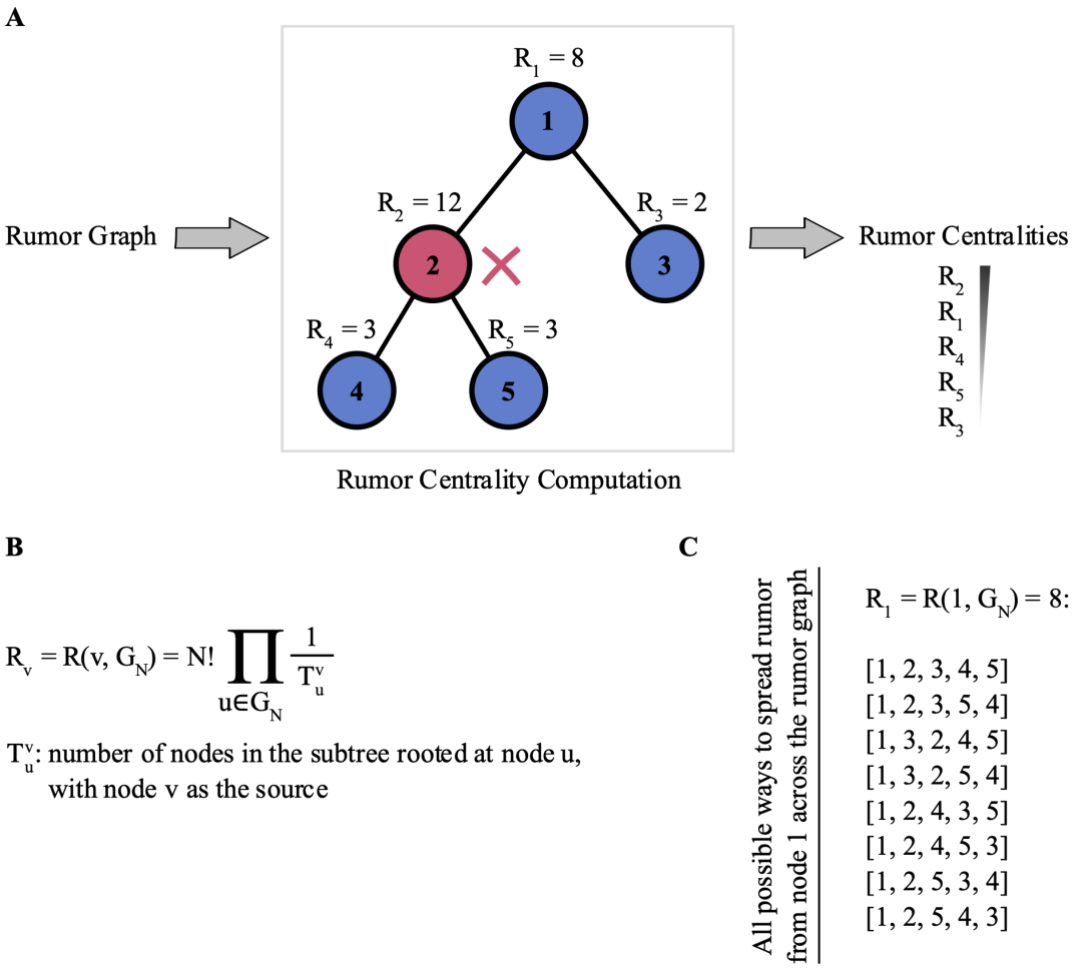


Figure C.1. Concept of the PINT framework. **(A)** For an example compound, PINT generated a rumor graph as the result of overlaying the driver genes of the compound's predicted bioprocesses on the genetic similarity network. PINT computed a rumor centrality score for each gene and ranked the genes in descending order of rumor centrality scores. In this example case, the rumor centrality for gene #2 (shown in red) is the highest score ($R_2 = 12$), indicating that gene #2 has the maximum likelihood to be the gene target of the compound. **(B)** Computation of the rumor centrality score for a gene (gene v) in the rumor graph. Refer to (188) for details. **(C)** All possible permutations that could cause the rumor graph infection with gene #1 as the rumor source. In this figure, each node represents a gene.

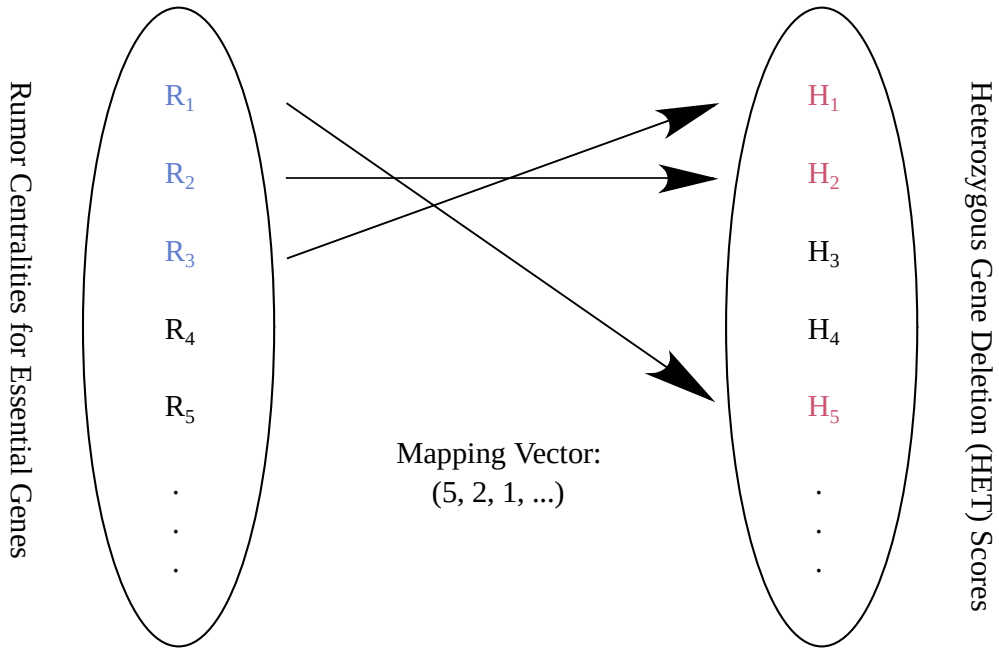


Figure C.2. Mapping approach for connecting rumor centralities with HET data. For all essential genes, we ranked the rumor centrality scores generated by the PINT framework in descending order and the HET scores in ascending order. We, then, mapped the rumor centrality score of each gene to its HET score to generate a mapping vector. We generated this mapping vector from the ranks of the HET data for the top K rumor centralities. In this example, the mapping vector will be (5,2,1) if K=3.