A Spectral-Geometric Framework for Non-Redundant Drug Target Selection via the Weighted Lovász Theta Function

Application to the *Mycobacterium tuberculosis* Interactome

A Dissertation Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE
in
MATHEMATICS

Submitted by

Vani Kumar
24/MSCMAT/12

Under the Supervision of
Prof. Sangita Kansal

Department of Applied Mathematics
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi - 110042

DECLARATION

I **Vani Kumar (24/MSCMAT/12)** hereby certify that the work which is being presented in this Dissertation II (MSMA204) entitled:

"A Spectral-Geometric Framework for Non-Redundant Drug Target Selection via the Weighted Lovász Theta Function: Application to the *Mycobacterium tuberculosis* Interactome"

in partial fulfillment of the requirements for the award of the Degree of Master of Science in Mathematics, submitted to the Department of Applied Mathematics, Delhi Technological University, is an authentic record of my own work carried out during the period from August 2025 to May 2026 under the supervision of **Prof. Sangita Kansal**. The matter presented in this dissertation has not been submitted by me for the award of any other degree or diploma of this or any other institute.

Vani Kumar
24/MSCMAT/12

Date:
Place: New Delhi

SUPERVISOR CERTIFICATE

I certify that **Vani Kumar (24/MSCMAT/12)** has carried out research work presented in this dissertation report entitled **"A Spectral-Geometric Framework for Non-Redundant Drug Target Selection via the Weighted Lovász Theta Function: Application to the *Mycobacterium tuberculosis* Interactome"** for the award of Master of Science in Mathematics from the Department of Applied Mathematics, Delhi Technological University, under my supervision. The contents of this dissertation are original and have not been submitted by the candidate for the award of any other degree or diploma of this or any other institute.

Prof. Sangita Kansal
Department of Applied Mathematics
Delhi Technological University

Date:

Place: New Delhi

ACKNOWLEDGEMENTS

I sincerely thank Prof. Sangita Kansal for her unwavering support, guidance, and encouragement throughout my research. Her expertise in graph theory has been instrumental in developing my understanding of these interdisciplinary subjects. The clarity of her mathematical exposition and thoughtful suggestions have significantly elevated the quality of this work.

I thank the Department of Applied Mathematics at Delhi Technological University for providing an excellent academic environment, access to vital resources, and a vibrant community of scholars. I am grateful for the collaborative atmosphere that fostered intellectual growth and interdisciplinary thinking throughout this program.

I am grateful to the broader research communities in both spectral graph theory and computational biology for making datasets, software tools, and theoretical frameworks openly available. This work would not have been possible without access to biological network databases and the rich mathematical literature that forms the foundation of this research. In particular, I would like to thank the maintainers of the AlphaFold, STRING, and IntAct databases for making high-quality metabolic reconstructions publicly available, enabling this work to be conducted smoothly.

Vani Kumar
24/MSCMAT/12

Abstract

Identifying non-redundant therapeutic target combinations in *Mycobacterium tuberculosis* is both a pressing clinical problem, driven by obligate multi-drug regimens and escalating resistance. This is a natural instance of the maximum weighted independent set (MWIS) problem on a biologically structured graph. We present a spectral-geometric framework that formalises this correspondence and exploits it computationally. From the Mtb H37Rv protein-protein interaction (PIP) network, constructed by integrating multiple experimental evidence sources, we derive a composite target relevance score encoding Tn-seq essentiality, structural druggability, and host-specificity. Embedding the network spectrally through the graph Laplacian, we construct a target-interference graph $H_\varepsilon = (U, F_\varepsilon)$ by connecting candidate proteins whose pairwise spectral distance falls below a threshold ε, encoding functional proximity as the interference relation. The weighted Lovász theta function $\vartheta(\bar{H}_\varepsilon, p)$ then provides a polynomial-time SDP upper bound on the MWIS of H_ε, with a rounded solution approximating the optimal non-redundant target combination. We establish that when H_ε is constructed from the Fiedler coordinate alone, it is an interval graph and therefore perfect. By the Lovász sandwich theorem, $\vartheta(\bar{H}_\varepsilon, p) = \alpha(H_\varepsilon, p)$ exactly, and the SDP solves the weighted target selection problem in polynomial time. For embeddings of dimension $k \geq 2$, H_ε is a unit ball intersection graph in $\mathbb{R}^k$ and we characterise the resulting integrality gap as a function of k, establishing a formal trade-off between spectral expressiveness and algorithmic tractability. Applied to the Mtb H37Rv interactome, the SDP-rounded target sets recover established drug targets, including KatG, InhA, and DprE1, within the computed independent set, and the identified combinations span functionally disjoint subsystems encompassing cell wall biosynthesis, central carbon metabolism, and DNA replication, consistent with the network-pharmacological rationale underlying combination anti-TB therapy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Tuberculosis: Disease Burden and the Drug Resistance Crisis

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* (Mtb) and remains one of the most serious infectious diseases in human history. TB overtook COVID-19 as the world's most deadly infectious disease in 2023, and it was the top cause of mortality for HIV-positive individuals [1]. The WHO Global Tuberculosis Report 2024 states that 10.8 million people contracted TB in 2023, which led to 1.25 million deaths [1, 2]. The highest incidence rates continue to occur in low and middle-income nations, where socioeconomic factors like poverty, malnutrition, and limited access to healthcare worsen outcomes and exacerbate transmission [3]. Roughly 25% of people worldwide have been infected with Mtb [4]. Of these, 5% are expected to acquire active TB within two years of infection, and an additional 5% will get the disease during the course of their lifetime [4]. The WHO End TB targets, which call for a 90% decrease in TB deaths and an 80% decrease in incidence by 2030 compared to 2015, are still far from being met despite decades of international control efforts and the availability of efficient treatment regimens [1, 4].

The standard treatment for drug-susceptible TB is a six-month regimen of four drugs administered concurrently: isoniazid, rifampicin, pyrazinamide, and ethambutol (HRZE) [5]. This enforced multi-drug strategy is not coincidental, but rather, a direct result of a fundamental biological property of Mtb. Drug resistance in Mycobacterium tuberculosis is the inevitable outcome of the selective pressure exerted by antimicrobial agents. Drug-resistant mutants arise spontaneously in any sufficiently large mycobacterial population, independent of drug exposure, and monotherapy reliably selects for these pre-existing resistant subpopulations while eliminating susceptible bacilli [6]. This principle was demonstrated decisively at the dawn

of TB chemotherapy: the first patient treated with streptomycin in 1944 developed resistance to the drug, and the landmark 1948 British Medical Research Council randomized trial confirmed that the majority of patients receiving streptomycin monotherapy harbored resistant organisms within months [7]. Subsequent studies showed that streptomycin resistance emerged in approximately 70% of patients after 120 days of monotherapy, whereas combination therapy with para-aminosalicylic acid reduced this rate to at most 9% [8]. This pattern recurred with every new anti-TB drug introduced thereafter, establishing that combination chemotherapy is essential to suppress the emergence of resistance [9]. The mechanistic basis for this is well characterised: combination therapy based on at least four effective drugs constrains the adaptive landscape of Mtb through purifying selection, whereas treatment with fewer than four effective drugs alleviates this constraint and allows positive selection of resistance determinants [8, 10].

When this combinatorial constraint is breached (through poor drug quality, inadequate dosing, treatment interruption, or pre-existing resistance) drug-resistant TB emerges. Multidrug-resistant TB (MDR-TB), defined as resistance to at least isoniazid and rifampicin, and extensively drug-resistant TB (XDR-TB), currently defined as MDR-TB with additional resistance to any fluoroquinolone and at least one of bedaquiline or linezolid, represent major global health threats [1]. An estimated 400,000 people developed multidrug-resistant or rifampicin-resistant TB (MDR/RR-TB) in 2023, causing approximately 150,000 deaths [1]. The consequences for treatment are severe: historically, MDR-TB required 18 to 20 months of therapy, often including injectable second-line agents with substantial toxicity [11]. Even with treatment, the global success rate for MDR/RR-TB has only recently reached 68%, far below the 88% achieved for drug-susceptible TB [1]. XDR-TB treatment outcomes are worse still, with a pooled success rate of approximately 44% under older regimens [12].

The drug resistance crisis is compounded by a stagnant discovery pipeline. The standard four-drug regimen for drug-susceptible TB has been in use since the 1970s [13]. Bedaquiline, the first genuinely new anti-TB drug class in over 40 years, was introduced only in 2013, and only two other novel agents (delamanid and pretomanid) have followed since [14, 13, 11]. Resistance to even these newest drugs is already being documented in clinical settings. Acquired bedaquiline resistance is reported in 2-5% of treated patients in systematic cohorts, and rates exceeding 10% in certain high-burden settings, while linezolid resistance has been found in 9-15% of MDR-TB strains and up to 60% in extensively drug-resistant cohorts [15, 16, 17]. This

situation creates an urgent need for novel therapeutic strategies. What is needed is not a single new drug target as history shows will be rapidly compromised by resistance. A principled framework for identifying combinations of targets whose simultaneous inhibition is maximally difficult for Mtb to escape through single or sequential resistance mutations is desired. This dissertation attemps to address this problem — the rational, non-redundant selection of multi-target anti-TB drug combinations grounded in the topology of the Mtb proteome.

1.2 *Mycobacterium tuberculosis* as a Biological System

Mycobacterium tuberculosis H37Rv is the standard laboratory reference strain of the human tubercle bacillus and the organism on which this dissertation is based. The strain was first isolated from a patient with pulmonary tuberculosis in 1905, and its whole genome sequence determined in 1998 remains the globally accepted reference sequence [18, 19]. The H37Rv genome encodes 3,924 predicted protein-coding genes that account for over 91% of the potential coding capacity [19]. A comprehensive re-annotation of the genome in 2002 subsequently assigned functions to approximately 52% of the 3,995 predicted proteins, with the remainder comprising conserved hypothetical proteins and proteins of entirely unknown function [20]. H37Rv remains the most widely used strain for computational analyses of *Mycobacterium tuberculosis* (Mtb) biology [19].

The Mtb proteome is organised into well-defined functional categories established by genome annotation and maintained in databases like Mycobrowser [21, 22]. The major classes include cell wall and cell surface processes, lipid metabolism, intermediary metabolism and respiration, information pathways, regulatory proteins, the PE/PPE families (approximately 10% of coding capacity), and the ESX secretion systems [18, 23]. This modular functional organisation is directly reflected in network-based analyses as computational studies using STRING-based protein interaction networks and genome-wide functional linkage mappings have confirmed that functionally related proteins in Mtb cluster into cohesive topological modules [24, 25, 26]. The existence of these functional modules (groups of proteins that interact densely within the module and sparsely outside it, with shared biological roles) is a prerequisite for spectral clustering and other community detection methods to recover biologically meaningful structure from the PPI network [27]. It also translates into a well-separated eigenvalue spectrum amenable to spectral embedding, due to functional subsystems that interact densely within themselves and

sparsely across [28]. With approximately 3,993 predicted proteins, the network is large enough to exhibit rich spectral structure yet small enough for full eigendecomposition to be computationally possible [20]. Furthermore, large-scale data-driven functional networks have been constructed at near-proteome scale for Mtb, and topological analyses of these networks have been shown to identify proteins essential for bacterial survival and virulence, demonstrating that network structure encodes biologically actionable information [29, 25].

Mtb is an obligate intracellular pathogen that survives and replicates in humans within alveolar macrophages, which are immune cells tasked with destroying it [30]. Central to this ability are three interconnected systems. The mycobacterial cell wall, extraordinary in its lipid complexity, provides both structural integrity and immune evasion [31, 32]. Upon infection, Mtb undergoes extensive metabolic reprogramming, shifting from glycolysis to fatty acid oxidation and the glyoxylate shunt which a pathway essential for intracellular persistence [33, 34, 35]. The ESX-1 type VII secretion system is a virulence mechanism which mediates phagosomal membrane rupture, enabling cytosolic access and cytotoxicity [30, 32]. This multi-system virulence architecture has a direct implication for drug discovery. Because Mtb survival depends on the coordinated activity of proteins across several functionally distinct subsystems, no single pathway inhibition is likely to be sufficient for elimination of viable bacteria. This is, in part, why the four-drug HRZE regimen targets four mechanistically distinct processes simultaneously [9, 10]. The central motivation of this thesis is to formalise this biological intuition computationally: given the topology of the Mtb PPI network and the functional organisation of its proteome, which combinations of target proteins are maximally non-redundant in the sense that they occupy distinct spectral regions of the network and therefore represent independent biological subsystems? The Lovász theta framework developed in subsequent chapters provides a principled answer to this question.

1.3 Protein-Protein Interaction (PPI) Networks in Drug Discovery

Proteins do not act in isolation. Nearly every biological process from signal transduction, gene regulation, metabolic catalysis, to structural organisation, is carried out not by individual proteins but by complexes and cascades of proteins acting in concert. Functions are carried out by interactions of proteins and small molecules, forming a complex interaction network, and

understanding these networks is fundamental to understanding cellular function [36]. In 2011, the canonical network medicine framework formalised the intuition of protein interaction. They discussed that diseases are rarely consequences of abnormalities in a single gene but rather reflect perturbations of complex intracellular and intercellular networks, and the tools of network medicine offer a platform to explore the molecular complexity of disease by identifying disease modules and pathways within the interactome [37]. From this perspective, drug targets should not be identified in isolation but in the context of the network neighbourhood they occupy.

Definition 1.3.1. *A protein-protein interaction network, or interactome, is a graph $G = (V, E)$ in which each node $v \in V$ represents a protein and each edge $\{u, v\} \in E$. represents a known or predicted physical interaction between proteins u and v. Edge weights encode the confidence or strength of each interaction [38].*

The application of graph-theoretic analysis to PPI networks for drug target identification was established by a series of influential studies in the early 2000s. The foundational observation was that PPI networks exhibit scale-free topology, i.e., their degree distributions follow a power law, meaning a small number of hub proteins have very many interaction partners while most proteins have few [39, 38]. This topology has direct implications for drug discovery. Hub proteins, by virtue of their high connectivity, were initially proposed as natural drug targets because their disruption would maximally perturb the network [40].

Centrality measures (degree, betweenness, closeness, and eigenvector centrality) then became the dominant computational tools for ranking candidate targets within PPI networks [41, 42]. Targets of approved, selective small-molecule drugs exhibit higher node centrality than broader investigational target sets, and centrality metrics can assist in evaluating targets with limited experimental evidence [42]. This justified a generation of target identification pipelines based on centrality ranking applied across pathogens, including Mtb [43]. Despite its appeal, the centrality paradigm has well-documented limitations. First, centrality measures are unreliable in isolation. They are sensitive to noise in the PPI network, cannot detect low-connectivity essential proteins, and their predictive performance depends heavily on how interaction data is integrated and represented [44]. Second, the correspondence between centrality and drug target quality is weaker than early work suggested. A systematic analysis found that degree, betweenness, and eigenvector centrality distributions are quite similar between known drug targets and other proteins, implying drug targets are neither network hubs nor privileged intermediaries [45]. Third, and most importantly for this thesis, centrality-based methods select

targets individually [42, 43, 45]. They produce a ranked list of single proteins, not combinations. For TB, where monotherapy is clinically prohibited, selecting targets individually and combining them post hoc provides no guarantee of non-redundancy in function. The top-ranked and second-ranked targets by any centrality measure may occupy the same functional neighbourhood and therefore provide overlapping rather than complementary coverage.

The limitations of centrality-based target identification, i.e., its insensitivity to functional redundancy, its inability to handle the combinatorial structure of multi-target selection, and equating network importance with therapeutic relevance, point toward a different class of methods [46]. What is needed is an approach that operates on the global geometry of the PPI network rather than local connectivity statistics, and that treats target selection as an inherently combinatorial problem rather than a ranking problem.

1.4 The Non-Redundancy Problem in Multi-Target Therapy

The clinical requirement for combination therapy in tuberculosis implicitly require a non-redundancy constraint. The drugs in a regimen must act on independent biological targets, since redundant coverage of the same subsystem provides no additional protection against resistance. Formalising this constraint computationally requires a precise notion of what it means for two targets to be non-redundant, and a tractable method for selecting a maximum-size or maximum-score combination satisfying this constraint. This section surveys the existing biological and computational literature that has approached variants of this problem, characterises the gap that these approaches leave, and motivates the specific formulation adopted in this dissertation.

The most extensively developed biological framework for non-redundancy in target selection is synthetic lethality (SL), originally described by Bridges in 1922 [47]. Two genes are synthetically lethal if loss of function of either alone is survivable but simultaneous loss of both is lethal [48]. Genomic screenings have enabled the discovery of synthetic lethal partners as potential drug targets in cancer, and paired with CRISPR-based functional genomic screening has been applied to identify new and druggable cancer targets [49]. The canonical clinical example is the synthetic lethality between BRCA1/2 loss and PARP inhibition in ovarian and breast cancers, which identified potential drug combinations explicitly designed around the non-redundancy principle [50]. Computational extensions of synthetic lethality have attempted

to generalise this pairwise gene interaction framework to larger networks. Liu et al. proposed a synthetic lethality-based computational method to identify anticancer drug targets using the human signalling network, identifying synthetic lethal gene pairs by mining cancer genes through a three-step network-based, frequency-based, and function-based screening strategy [51]. However, genome-wide synthetic lethality mappings typically rely on pairwise experimental data of genetic interactions, which is sparse for pathogens like Mtb [52]. Additionally, it becomes computationally unmanageable when extending to the high-order combinations required for TB regimens.

TB-specific drug design often utilizes empirical ranking-and-exclusion frameworks that prioritize drug combinations based on pairwise interaction profiles to avoid overlapping resistance mechanisms [53, 54]. Systematic experimental measurement of pairwise drug interactions can be used to rank high-order combinations by strength of synergy and to establish exclusion criteria based on drug interaction correlation. The tradeoff in measuring all pairwise interactions experimentally is that interactions scale quadratically in the number of drugs considered, and provides no polynomial-time optimality guarantees. Methodologically related geometric approaches, such as hyperbolic mapping of the Mtb protein-protein interaction (PPI) network, identify targets in distinct network regions but do not formulate selection as a formal combinatorial optimization problem [55].

The natural combinatorial formalisation of non-redundancy is the maximum weighted independent set problem on an interference graph where nodes are candidate targets, edges connect interfering pairs, and the objective is to find the highest-scoring set with no edges. This formulation is general enough to subsume both the synthetic lethality framework (where the interference relation is defined by co-lethality) and the network proximity framework (where it is defined by spectral embedding distance), and it produces a combination rather than a ranking. This problem is NP-hard in general, and existing computational approaches in biological network analysis typically address related target-selection problems through heuristics such as centrality or betweenness-based rankings that implicitly ignore the combinatorial structure entirely [56, 57, 58]. For example, maximum flow approaches to Mtb drug target prioritisation use network centrality and flow to resistance genes to rank individual targets, but do not address the combinatorial problem of selecting a non-redundant set simultaneously [59].

A distinct line of work enforces a form of non-interference through network proximity on the human PPI interactome. Cheng, Kovacs and Barabasi (2019) classified drug-drug-disease

relationships into six topological classes by quantifying the network distance between drug targets and disease proteins [60]. Only one class correlated with clinically efficacious combinations. It was the class with the condition that both drugs' target sets lie within the same disease module but occupy separate network neighborhoods. This is a non-interference constraint in spirit as it explicitly excludes drug pairs whose targets overlap or cluster in the same region of the interactome. Li et al. (2011) proposed a related "network target" paradigm in which the NIMS algorithm evaluates whether agents target distinct but functionally connected parts of a disease network [61]. More recently, Li et al. (2025) used graph neural network embeddings with diminishing-return thresholds to select high-order drug combinations targeting multiple pathways [62].

Despite their use of a non-interference criterion, these network proximity methods address a fundamentally different problem from the one formulated in this thesis. First, they operate on drug pairs, not on selecting an optimal combinatorial set of protein targets from a larger candidate pool. The Cheng et al. framework classifies pairwise drug-disease relationships but does not scale to selecting three or more targets simultaneously with a global constraint. Second, they provide a ranking or threshold-based heuristic rather than a combinatorial optimisation problem. There is no objective function maximising total relevance under explicit pairwise interference constraints, and no algorithm that returns an optimal or near-optimal set. Third, their interference criterion is based on shortest-path proximity in the interactome, not on spectral embedding distance derived from a Laplacian embedding, so the resulting graph structure does not satisfy the perfect graph property and cannot exploit the exactness of the Lovasz theta bound. Fourth, none of these methods provide an optimality bound, i.e., a dual bound quantifying how far the selected combination is from the true optimum. For a drug discovery application where experimental validation is expensive, knowing that the computed combination is within a known factor of the best possible combination is as important as the combination itself.

1.5 Spectral Graph Theory in Biological Networks

Spectral graph theory is the study of graphs through the eigenvalues and eigenvectors of matrices associated with them, primarily the adjacency matrix, the Laplacian, and their normalized variants [63, 64]. The central application is that the spectrum of these matrices encodes structural properties of the graph such as connectivity and community structure that are otherwise

difficult to compute directly [64, 65]. The Laplacian matrix became an important tool of spectral graph theory for the investigation of structural properties of large biological networks, as many important features of the underlying structure and dynamics of systems can be extracted from the spectral analysis of the graphs [63, 64].

PPI networks have modular structure as functionally related proteins cluster into densely connected subgraphs [27]. This modularity produces a well-separated eigenvalue spectrum in the graph Laplacian. A small number of eigenvalues near zero corresponding to the near-disconnected functional communities, followed by a clear eigengap before the remaining eigenvalues [64, 66]. It is this spectral signature of modularity that makes Laplacian-based methods appropriate for biological network analysis and distinguishes them from local degree-based measures, which are sensitive to individual node properties rather than global network organisation.

Spectral embedding methods represent each protein in a PPI network as a point in a low-dimensional Euclidean space derived from the eigenvectors of the graph Laplacian. The first k non trivial Laplacian eigenvectors embed proteins into $\mathbb{R}^k$. The idea that Laplacian eigenvectors embed a PPI network and recover the modular structure from graph topology alone has many precedents. These algorithms identify clusters in a graph using the eigenvectors of the Laplacian, such that nodes within a cluster are connected by highly similar edges and inter-cluster connections are weak (modular functionality), and have been successfully applied to predict protein complexes from PPI networks [67, 64]. The spectral approach is explicitly contrasted with local degree-based measures and found to be more robust to "the heterogeneity of PPI networks" [67]. Rai and Jalan (2018) argue that random-matrix analysis of network spectra "provides new practical tools for identification of pathway proteins, etc. responsible for the occurrence of the disease" and that spectral methods distinguish organizational differences between healthy and diseased states [68]. Recently, Zhang et al. used Laplacian eigenvectors to provide a "global, geometry-aware coordinate system" for each node in a biomedical knowledge graph, explicitly demonstrating that spectral features capture long-range structural proximity better than local neighbourhood aggregation [69]. Similarly, Liu et al. used spectral clustering to find trug targets for pancreatic ductal adenocarcinoma [70]. For Mtb specifically, network-based computational approaches have been applied to identify functional modules and drug targets from the H37Rv interactome. Mazandu and Mulder integrated functional genomics data to generate large-scale functional interaction networks for Mtb and carried out computa-

tional analyses using network topological properties to identify proteins essential to the survival, growth, and virulence of the pathogen as drug targets [71]. However, these approaches relied on centrality measures rather than spectral geometry, and did not use eigenvector embeddings to define functional distance between proteins.

Spectral embedding methods represent each protein in a PPI network as a point in a low-dimensional Euclidean space derived from the eigenvectors of the graph Laplacian, such that proteins occupying the same dense interaction neighbourhood are mapped to nearby points and proteins in different functional regions are mapped far apart. The resulting pairwise spectral embedding distance between two proteins is therefore a geometric proxy for their functional separation in the network. A small spectral embedding distance indicates shared functional context, while a large spectral embedding distance indicates membership in distinct biological subsystems. This use of spectral geometry as a measure of functional proximity in biological networks is a well-established.

Before turning to spectral emmbedding distance as a proximity measure, it is worth situating the spectral embedding approach relative to the broader landscape of graph embedding methods for biological networks. Node2vec and DeepWalk learn node embeddings via random-walk-based neural objectives while graph neural networks aggregate local neighbourhood features through learned message-passing layers [72, 73, 74]. Matrix-factorisation methods [75] decompose the adjacency or diffusion matrix into low-dimensional factors. These methods have demonstrated strong empirical performance on function prediction and link prediction benchmarks. However, for the purposes of this thesis, spectral embedding has a decisive advantage over all of these alternatives. It produces a representation with a rigorous mathematical relationship to the graph's structure, whose properties such as eigenvalue bounds, commute-time approximation, connection to the heat kernel, are analytically characterisable [76, 77, 78]. Node2vec and graph neural network embeddings are optimised for predictive performance and lack the theoretical guarantees needed to prove that the interference graph H_ε constructed from spectral distances is an interval graph. The interval graph result, and consequently the exactness of the Lovász theta bound, depends on the specific geometric property that thresholding a one-dimensional coordinate produces an intersection graph of intervals on the real line. This property holds for the Fiedler coordinate but has no analogue in learned embeddings, making spectral geometry the only embedding framework within which the theoretical contribution of this thesis can be established.

The spectral embedding assigns each protein a point in $\mathbb{R}^k$ and the spectral embedding distance measures how far apart two proteins are in this space. A small spectral embedding distance indicates that two proteins occupy the same dense interaction neighbourhood and they are functionally co-localised [27, 64]. A large spectral embedding distance indicates membership in distinct biological subsystems. This use of spectral distance as a proxy for functional proximity is well-validated in the literature. Cao et al. introduced the Diffusion State Distance (DSD), an L_1 distance between random-walk diffusion profiles of two proteins. They showed that replacing shortest-path distance with DSD consistently improves protein function prediction across multiple classical methods in the *S. cerevisiae* PPI network. They argue that this is because DSD captures global network topology rather than local edge density [79]. Voevodski et al. validated that PageRank Affinity, a spectral proximity measure proportional to the number of times one protein is visited in a random walk restarting at another [80]. This explicitly demonstrated that spectral measures of closeness are more robust to noise and more precise than shortest-path and adjacency-based methods. Windels, Malod-Dognin, and Pržulj extended this to graphlet Laplacian embeddings, demonstrating that spectral distance in the resulting space encodes topology-function relationships and produces clusters with strong GO enrichment [81]. Finally, Boehnlein et al. established the theoretical connection between DSD and the heat kernel of the normalised Laplacian, formally grounding diffusion-based proximity measures within the spectral geometry framework and confirming that the empirical success of DSD-type distances has a rigorous spectral basis [78].

Despite the broad adoption of spectral clustering in mammalian PPI and disease network analysis, its application to bacterial pathogen networks, and specifically to Mtb, has remained limited. Existing computational analyses of the Mtb interactome have employed centrality measures, subtractive genomics pipelines, and community detection heuristics, but none has applied a spectral embedding of the Mtb Laplacian as the basis for a drug target selection framework [71, 25, 29]. More specifically, no existing work constructs a target interference graph from spectral embedding distances, nor uses the geometry of the Laplacian embedding to define a formal non-redundancy criterion for target combinations. The spectral-geometric framework in this thesis therefore represents the first application of Laplacian spectral embedding to combinatorial drug target selection in a bacterial pathogen network, and the connection between this embedding and the Lovász theta SDP via the interval graph structure of H_ε is entirely novel.

1.6 Problem Statement and Thesis Objectives

The preceding sections have established the following chain of reasoning. Tuberculosis requires combination therapy because monotherapy reliably selects for resistance. The clinical design principle underlying combination regimens (that drugs must act on independent biological subsystems) is a non-redundancy constraint that existing computational target identification methods do not formalise. Network-based approaches reduce the problem to graph analysis, but centrality measures select targets individually and provide no combinatorial guarantees. Others methods cap the interaction to two targets and not multiple, as required for TB. Spectral graph theory provides a principled notion of functional distance on PPI networks but has not been applied to combinatorial target selection in bacterial pathogens. The maximum weighted independent set problem is the natural combinatorial formalisation of non-redundancy, do not formulate the combinatorial selection problem as maximum weighted independent set or provide optimality guarantees. This thesis addresses the following problem:

Problem (Non-Redundant Drug Target Selection) Given a weighted protein-protein interaction graph $G_{\text{Mtb}} = (V_{\text{Mtb}}, E_{\text{Mtb}}, w)$ for *Mycobacterium tuberculosis* H37Rv, a composite drug target relevance score $p_v \geq 0$ for each protein $v \in V$ encoding essentiality, druggability, and host-specificity, and a spectral embeddingdistance threshold $\varepsilon > 0$, find a set $S \subseteq U$ of candidate proteins that maximises the total relevance score $\sum_{v \in S} p_v$ subject to the constraint that every pair $\{u, v\} \in S$ satisfies $d_{\text{spec-emb}}(\{u, v\}) > \varepsilon$, i.e., no two selected targets are functionally proximate in the spectral embedding of G_{Mtb}

This is the maximum weighted independent set problem on the target-interference graph

$$H_\varepsilon = (U, F_\varepsilon), \text{ where } F_\varepsilon = \{\{u, v\} \subseteq U : d_{\text{spec-emb}}(\{u, v\}) \leq \varepsilon\}$$

The objective value is

$$\alpha(H_\varepsilon, p) = \max_{S \text{ indep.}} \sum_{v \in S} p_v$$

and the Lovász theta function $\vartheta(\bar{H}_\varepsilon, p)$ provides a polynomial-time computable upper bound on this value via semidefinite programming.

The choice of Mtb H37Rv as the target organism is motivated by four convergent considerations. First, the clinical imperative is unambiguous. The standard TB regimen already embodies the non-interference principle by design, and the emergence of MDR and XDR-TB strains

resistant to first- and second-line agents creates an urgent need for new target combinations with independent mechanisms [82, 83]. Second, the Mtb proteome has the right computational properties. The H37Rv genome encodes approximately 3,993 predicted proteins [18]. The PPI network is large enough to exhibit rich spectral structure (multiple well-separated eigenvalue clusters corresponding to distinct functional subsystems) but small enough for Laplacian eigendecomposition and SDP optimisation to be computationally managable without approximation or distributed computing. Third, the multi-subsystem virulence architecture of Mtb creates the functional modularity that the spectral embedding must recover in order for the interference graph H_ε to have meaningful structure. A pathogen whose proteome lacked this modularity would produce a flat eigenvalue spectrum with no clear eigengap, making spectral distance meaningless as an interference criterion. Fourth, the Mtb PPI data are sufficiently mature. Three complementary experimental sources (STRING, IntAct, and the genome-scale bacterial two-hybrid screen of Wang et al.) can be integrated into a principled weighted graph, as described in Chapter 2 [84, 85, 86]. Validated drug targets including KatG, InhA, RpoB, GyrA, and DprE1 are available as ground truth for evaluating the framework's output [87].

The interference graph H_ε could in principle be defined using any pairwise distance on V, including the Euclidean distance as th embedding maps the vertices to $\mathbb{R}^k$. The choice of the spectral embeddingdistance $d_{\text{spec-emb}}(u, v) = ||\Phi_k(u) - \Phi_k(v)||_2$ derived from the normalised Laplacian embedding is motivated by three properties that alternative distance measures lack. First, it is robust to noise and missing edges. The shortest-path distance is brittle as a single missing edge can dramatically alter the distance between two nodes. Since PPI networks are systematically incomplete (interactions are underreported proportionally to how well-studied a protein is), a distance measure derived from the global spectral structure of the Laplacian is far more stable than one derived from individual paths. Second, it has a rigorous probabilistic interpretation as an approximation to commute time in the random walk on G_{Mtb}. Two proteins with small spectral embedding distance are ones between which random walks travel rapidly, meaning they are in the same densely connected functional neighbourhood. This provides a biological interpretation of the interference relation: two targets interfere if a random biological signal propagating through the Mtb PPI network reaches both of them quickly, implying functional co-localisation. Third, and most importantly for the theoretical results of this thesis, defining H_ε via spectral distance from the Fiedler coordinate produces an interval graph, which is a perfect graph. This is not a coincidental property of spectral distances. It follows directly

from the fact that thresholding a one-dimensional coordinate produces an intersection graph of intervals on the real line. It is this perfect graph structure that makes the Lovász theta bound exact, converting an NP-hard problem into one solvable in polynomial time.

The Lovász theta function is the appropriate tool for this problem for reasons that are both theoretical and practical. Theoretically, it is the tightest known polynomial-time computable upper bound on the independence number for general graphs, and it achieves exactness on perfect graphs via the sandwich theorem [88]. No linear programming relaxation achieves this. The fractional relaxation of the independent set polytope is loose on most graphs, and the integrality gap can be arbitrarily large [89, 90, 91]. The SDP relaxation underlying ϑ is strictly tighter and collapses to zero integrality gap precisely on perfect graphs, which H_ε is when constructed from the Fiedler coordinate [88, 92]. Practically, the weighted SDP is solvable in polynomial time using interior-point methods, and for candidate set sizes $|U| \approx 200 - 400$ arising from the Mtb PPI, the SDP matrix is at most 400×400 which solvable in seconds using standard solvers such as CVXPY with the MOSEK backend [93]. Finally, the SDP solution X^* provides more than just an optimal combination. The orthonormal vector representation $\{y_v\}_{v \in U}$ associated with X^* embeds the candidate targets in a geometric space where selected targets are mutually orthogonal. This geometric certificate is interpretable. It shows explicitly which directions in the SDP representation space each target occupies, and it provides a dual bound that quantifies how close the rounded solution is to the true optimum.

None of these approaches provides approximation guarantees, and none exploits special graph structure to achieve exactness The Lovász theta function $\vartheta(\bar{H}_\varepsilon, p)$ breaks this pattern. As established in Section 2.3, it provides a polynomial-time SDP upper bound on the maximum weighted independent set that is exact when H_ε is a perfect graph, and specifically when H_ε is an interval graph, which follows when the interference relation is defined by thresholding the one-dimensional Fiedler coordinate. This means that for the specific interference graph constructed from the spectral geometry of G_{Mtb}, the combinatorial optimisation problem is not approximated but solved exactly, with a certificate of optimality provided by the SDP dual. To the best of the author's knowledge, no existing work in computational biology, for TB or any other pathogen, has applied the Lovász theta function to target selection, nor has any work established a connection between the spectral geometry of a PPI network and the perfect graph structure of a derived interference graph. This constitutes the primary theoretical contribution of the thesis.

A natural objection is that a greedy algorithm (sort candidate targets by score p_v, then iteratively select the highest-scoring target not adjacent in H_ε to any already selected) is simple and fast, and may perform well in practice. There are two responses. First, greedy provides no approximation guarantee on general graphs: its solution can be arbitrarily far from optimal depending on graph structure, and on dense interference graphs typical of biologically structured networks, the gap can be substantial [94]. Second, and more fundamentally, greedy provides no certificate. If the greedy solution contains k targets, there is no way to know from the greedy output alone whether a better $k + 1$-target combination exists. The SDP provides both an approximate solution and an upper bound on the true optimum, so the gap between the two is explicitly quantifiable. For a drug discovery application where the cost of experimental validation is high, knowing that the computed combination is within a small factor of the best possible combination is as important as the combination itself.

Chapter 2

Preliminaries

2.1 Contruction of the Mtb PPI Network

The STRING database was used to obtain a high-confidence protein-protein interaction (PPI) network for *Mycobacterium tuberculosis* H37Rv (taxid 83332) [84]. In STRING, each interaction is supported by one or more independent evidence sources, referred to as channels. These include experimental evidence (experiments), curated pathway/database annotations (database), co-expression patterns (coexpression), genomic neighborhood, gene fusion, co-occurrence, and text mining from scientific literature. Since the objective was to construct a biologically reliable PPI network with stronger physical or functional support, only interactions with experimental or database annotation score values greater than 0 were retained [84]. This removes edges supported solely by indirect or weaker association signals which may capture statistical or literature-based associations rather than experimentally supported interactions. A further filtering step was applied to retain only STRING high-confidence interactions, i.e., interactions with combined score values greater than 700. For each interaction pair (u, v), the per-channel confidence scores s_{uv}^{c} were combined using the standard probabilistic combination with independent source:

$$w_{uv}^{\text{STRING}} = 1 - \prod_{c \in \text{ channels}} (1 - s_{uv}^{c})$$

where the channels are experiments and database [84]. This formulation treats each evidence channel as an independent source of support and computes the probability that at least one source provides valid evidence for the interaction. Consequently, interactions supported by multiple strong evidence channels receive higher weights, while weakly supported associations contribute less to the final network structure.

The secondary interaction source was obtained from the IntAct database, a manually cu-

rated molecular interaction repository maintained by the European Bioinformatics Institute [85]. Interactions involving *Mycobacterium tuberculosis* H37Rv were retrieved and to improve biological reliability and avoid inclusion of indirect functional associations, two specific molecular interaction (MI) type filters were applied [85]. The physical association filter includes interactions for which experimental evidence suggests that two proteins are part of the same physical molecular complex or binding relationship. Although such evidence strongly suggests physical connectivity, it does not necessarily prove direct molecular contact between the proteins. The direct interaction filter is more stringent and includes interactions where experiments indicate direct physical binding between two proteins. These interactions therefore represent the strongest form of experimentally supported PPI evidence available in curated databases [95].

After downloading the dataset, protein identifiers were mapped to standardized Rv locus tags and duplicate interactions were merged. Since IntAct contains manually curated experimental evidence rather than probabilistic interaction scores, edge weights were assigned according to the degree of experimental support reported for each interaction [85]. Interactions supported by two or more independent experimental confirmations were treated as high-confidence physical interactions and assigned:

$$w_{uv}^{\text{IntAct}} = 1$$

whereas interactions supported by only a single experimental observation were assigned a lower confidence weight:

$$w_{uv}^{\text{IntAct}} = 0.7$$

This weighting strategy reflects the comparatively high reliability of manually curated physical interaction data while still accounting for varying levels of experimental reproducibility.

To complement the STRING and IntAct derived interaction network, a tertiary PPI dataset generated using a high-throughput bacterial two-hybrid (B2H) assay was incorporated . The dataset was obtained from a study by Wang et al., which experimentally screened nearly the complete ORFeome of M. tuberculosis H37Rv and identified over 8,000 novel interactions [86]. The authors reported an experimental validation rate exceeding 60%, indicating moderate-to-high reliability of the detected interactions. Although the validation rate of the B2H screen was low, comparison with the STRING and IntAct interaction sets showed that the B2H dataset contained additional interactions absent from both sources. Consequently, the B2H network was

incorporated in its entirety after filtering, as it contributed biologically relevant interaction information that improved overall network completeness and reduced the likelihood of excluding potentially important interactions missing from curated repositories. Protein identifiers from this study were mapped to standardized Rv locus tags. Since the B2H dataset does not provide calibrated probabilistic interaction scores comparable to STRING confidence values, all retained B2H interactions were assigned a uniform edge weight:

$$w_{uv}^{\text{B2H}} = 0.6$$

This constant weighting reflects the experimentally derived nature of the interactions while preventing overestimation of evidence strength relative to curated STRING and IntAct interactions.

The final network was constructed by integrating interaction evidence from the sources into a unified weighted graph as:

$$w_{uv}^{\text{final}} = 1 - \prod_{\text{sources}} (1 - w_{uv}^{source})$$

This formulation increases the confidence of interactions supported by several independent evidence sources while preventing weights from exceeding unity [96]. Consequently, interactions simultaneously supported by curated databases, experimental assays, and high-confidence STRING evidence received higher effective confidence values than interactions derived from only a single source. The resulting integrated network therefore captures both experimentally validated physical interactions and broader functional associations while maintaining a principled weighting scheme across heterogeneous datasets [97].

Definition 2.1.1. *The Mycobacterium tuberculosis H37Rv protein-protein interaction graph is an undirected weighted graph*

$$G_{Mtb} = (V_{Mtb}, E_{Mtb}, w)$$

where V_{Mtb} is the set of Mtb H37Rv proteins included after filtering, with each protein identified by its Rv-number locus tag, E_{Mtb} is the set of unordered pairs $\{u, v\}$ representing experimentally supported or high-confidence computationally inferred physical interactions between Mtb proteins and $w : E_{Mtb} \to (0, 1]$ assigns each edge a confidence weight

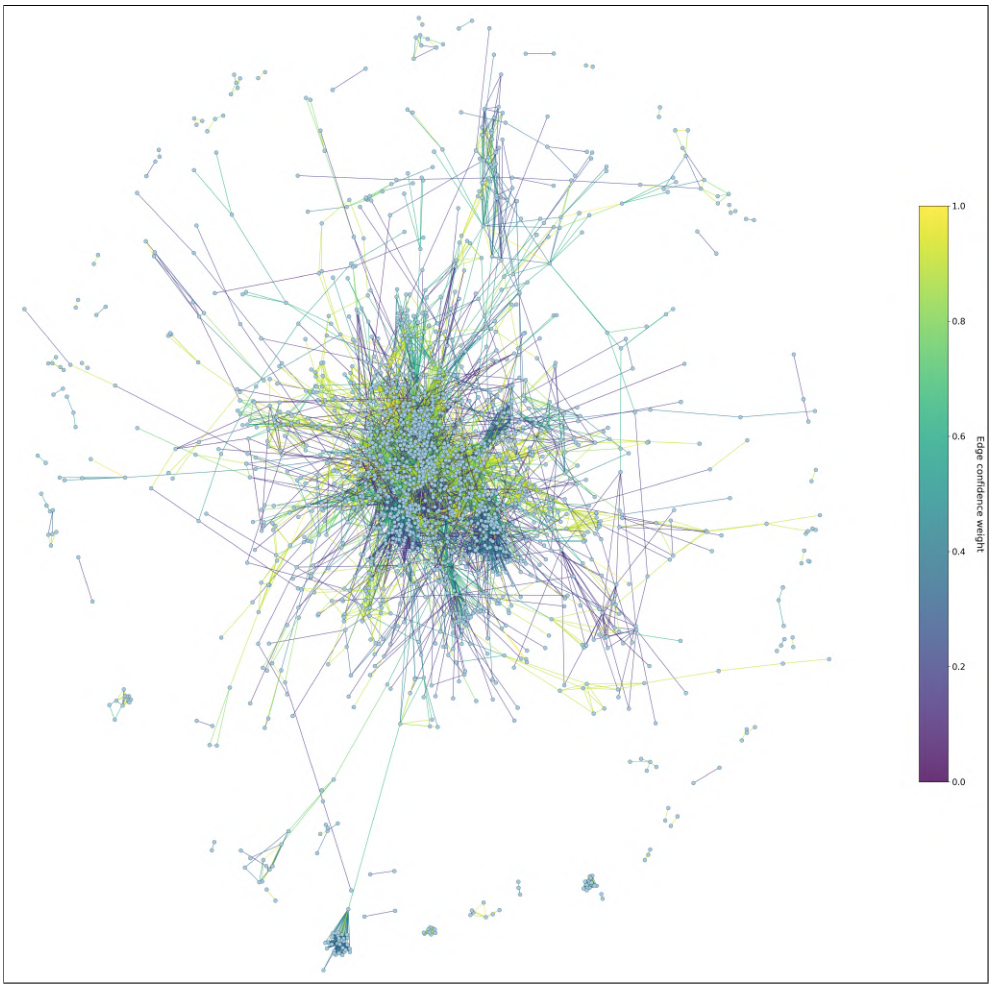


Figure 2.1: Visualization of the *M. tuberculosis* PPI network: Nodes correspond to Mtb proteins, the layout positions nodes so that proteins with many or strong interactions are drawn closer together, while sparsely connected proteins appear further apart. Edge colors encode the normalized combined confidence weight of each interaction from low realtive confidence (purple) to high relative confidence (yellow/green)

After integration, the final network contained 2,137 protein nodes and 15,017 interaction edges with no isolated vertices. Although the graph contained 61 connected components, the largest connected component comprised 1,928 nodes, representing the overwhelming majority of the network. This is typical of biological interaction networks, which often exhibit a dominant giant component together with smaller peripheral modules corresponding to sparsely characterized or functionally specialized protein groups [98]. The relatively large edge count and dense giant component provide sufficient structural complexity for meaningful spectral analysis, in line with previous studies that successfully applied Laplacian spectral methods to large, sparse PPI networks [99].

2.2 Disease Scores and Candidate Targets

The objective of this dissertation is to identify biologically meaningful proteins that may serve as potential anti-TB drug targets. Network centrality and connectivity alone are insufficient to determine therapeutic relevance, since highly connected proteins are not necessarily essential for bacterial survival, chemically targetable, or selective with respect to the human host [100]. Consequently, a biologically informed scoring framework was introduced to prioritize proteins according to their suitability as drug targets. This framework integrates multiple complementary criteria associated with successful antimicrobial target selection and by combining these with network structure, enables the extraction of a candidate target set of proteins likely to be therapeutically actionable.

In antimicrobial drug discovery, three biological properties are commonly used to evaluate whether a protein constitutes a viable therapeutic target [101]. First, proteins essential for bacterial survival or virulence are preferred, since their inhibition is more likely to suppress pathogen growth or persistence. Second, a target protein must be chemically druggable, meaning it possesses structural or biochemical features permitting modulation by small molecules or other therapeutic compounds. Third, ideal targets should exhibit sufficient specificity to the pathogen and avoid strong similarity to human proteins, thereby reducing the likelihood of host toxicity or off-target effects [101, 102]. These criteria are widely adopted in computational target prioritization pipelines and provide complementary perspectives on therapeutic relevance.

Definition 2.2.1. *Each protein $v \in V_{Mtb}$ receives a non-negative scalar score $p_v \geq 0$ called the Drug-Target Relevance Score, encoding its relevance as an anti-TB drug target. This score is a weighted combination of three criteria:*

$$p_v = \alpha_1 \cdot \mathit{Ess}(v) + \alpha_2 \cdot \mathit{Drug}(v) + \alpha_3 \cdot \mathit{Cons}(v)$$

with $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and $\alpha_1 \geq \alpha_2 \geq \alpha_3$

The essentiality component $\mathrm{Ess}(v)$ measures whether a protein is required for bacterial survival or virulence under experimentally characterised conditions. Essentiality data wes obtained from the study conducted by DeJesus et al in which they perfromed Essentiality Analysis of Mtb and reported the essentiality of all proteins as a supplement. [103, 104, 22]. Proteins classified as essential, conditionally essential, and non-essential were assigned scores of 1, 0.5 and 0 respectively. The druggability component $\mathrm{Drug}(v)$ quantifies the likelihood that a protein

can be modulated by small-molecule therapeutics [103]. Druggability scores were assigned in two tiers. Proteins belonging to a curated set of literature-confirmed drug targets (Table 2.1) were assigned $\text{Drug}(v) = 1$. For all remaining proteins, druggability was estimated using the mean per-residue confidence score (pLDDT) of the AlphaFold structural model retrieved from the EBI AlphaFold database [105, 106]. The confidence metric ranges from 0 to 100, where values above 70 indicate well-ordered regions with reliable predicted structure and values below 50 indicate intrinsically disordered regions [105]. Since structural order is a necessary but not sufficient condition for the existence of a druggable binding pocket, pLDDT was used as a proxy for structural druggability rather than as a direct measure of ligandability. The raw pLDDT score was normalised and mapped to a druggability score via:

$$\text{Drug}(v) = \min\left(\frac{\text{pLDDT}(v)}{100} \times 0.6, 0.6\right)$$

The multiplicative cap of 0.6 ensures that even a perfectly structured uncharacterised protein cannot score as highly as a literature-confirmed target, preserving the intended hierarchy between the two tiers. The conservation component $\text{Cons}(v)$ evaluates sequence similarity between each Mtb protein and the human proteome, serving as a proxy for target specificity and potential host toxicity. Basic Local Alignment Search Tool for Proteins (BLASTp) searches were performed against UniProt human proteome (taxon 9606) [107]. Proteins with significant similarity (E-value $< 10^{-4}$) were assigned $\text{Cons}(v) = 0$, indicating likely cross-reactivity with the host, while proteins lacking close human homologs were assigned $\text{Cons}(v) = 1$, indicating high target specificity to the pathogen [108, 25].

No single criterion is sufficient in isolation as a protein may be essential but structurally undruggable and so on [109, 110]. Effective target prioritization therefore requires integrating all three axes simultaneously, which motivated the composite linear score [101, 111]. Each coefficient α_i directly encodes the relative importance of its criterion and is consistent with established composite scoring approaches in computational target prioritization. Nonlinear aggregation functions (such as geometric means or product-based scores) would amplify the effect of any single zero-valued component. For instance, a protein with $\text{Cons}(v) = 0$ receive $p_v = 0$ regardless of its essentiality or druggability, effectively implementing a hard filter rather than a soft penalty.

The ordering constraint $\alpha_1 \geq \alpha_2 \geq \alpha_3$ reflects a principled hierarchy of criteria. Essentiality receives the highest weight because a non-essential protein, however druggable, offers no therapeutic benefit. Its inhibition will not suppress bacterial growth or virulence [101]. Drugga-

Rv	Gene	Protein	Drug(s)	Source
First-line drug targets (clinically validated)				
Rv1908c	katG	Catalase-peroxidase (activates prodrug)	Isoniazid	[112]
Rv1484	inhA	NADH-dependent enoyl-ACP reductase	Isoniazid, ethionamide	[112]
Rv0667	rpoB	RNA polymerase β subunit	Rifampicin	[112]
Rv0006	gyrA	DNA gyrase A subunit	Fluoroquinolones	[112]
Rv0005	gyrB	DNA gyrase B subunit	Fluoroquinolones	[112]
Rv3795	embB	Arabinosyltransferase B	Ethambutol	[113]
Rv3794	embA	Arabinosyltransferase A	Ethambutol	[113]
Rv3793	embC	Arabinosyltransferase C	Ethambutol	[113]
Rv0058	rpsL	Ribosomal protein S12	Streptomycin	
Newer / approved drug targets				
Rv1305	atpE	ATP synthase subunit c	Bedaquiline	[114]
Rv0206c	mmpL3	Trehalose monomycolate transporter	SQ109 (phase 2b-3)	[115]
Advanced clinical pipeline targets				
Rv3790	dprE1	Decaprenylphosphoryl-β-D-ribose oxidase	BTZ043, PBTZ169 (macozinone), OPC-167832, TBA-7371	[116]
Rv3791	dprE2	Decaprenylphosphoryl-β-D-ribose reductase	(partner of DprE1 pathway)	[116]
FAS-II mycolic acid biosynthesis targets				
Rv2245	kasA	β-Ketoacyl-ACP synthase I		[117]
Rv2246	kasB	β-Ketoacyl-ACP synthase II		[116]
Rv1483	fabG1 (mabA)	3-Oxoacyl-ACP reductase		[116]
Rv2243	fabD	Malonyl-CoA-ACP transacylase		[117]
Rv3280	fabH	β-Ketoacyl-ACP synthase III		[117]
Rv3800c	pks13	Polyketide synthase		[114]
Peptidoglycan / cell division targets				
Rv2152c	murC	UDP-N-acetylmuramate-alanine ligase		[118]
Rv2153c	murG	UDP-N-acetylglucosamine–N-acetylmuramyl pentapeptide transferase		[118]
Rv2155c	murD	UDP-N-acetylmuramoylalanine-D-glutamate ligase		[118]
Rv2157c	murF	UDP-MurNAc-pentapeptide ligase		[118]
Rv3581c	murX (mraY)	Phospho-N-acetylmuramoyl-pentapeptide transferase		[118]
Rv2150c	ftsZ	Cell division protein FtsZ		[115]
Pantothenate / CoA biosynthesis targets				
Rv3602c	panC	Pantothenate synthetase		[115]
Rv3601c	panB	3-Methyl-2-oxobutanoate hydroxymethyltransferase		[115]
Rv3628	panD	Aspartate 1-decarboxylase		[116]
Rv2225	glnA1	Glutamine synthetase I		[116]

Table 2.1: Verified Druggable Targets of *Mycobacterium tuberculosis* H37Rv

bility is weighted second because, among essential proteins, chemical tractability is the primary bottleneck in translating a biological target into a therapeutic compound [101, 111]. Conservation receives the lowest weight not because host-specificity is unimportant, but because it is a binary hard filter (proteins with $\text{Cons}(v) = 0$ are substantially penalised by losing their entire α_3 contribution), and the downstream Lovász framework can further deprioritise them via the independent set selection. This prevents the conservation criterion from dominating the score of proteins that are both essential and druggable but happen to have distant human paralogs with E-values just below 10^{-4} [101]. The coefficients $\alpha_1, \alpha_2, \alpha_3$ are not fixed analytically but are treated as hyperparameters. Any configuration satisfying the constraints in the definition is a valid choice. The primary weight assignment $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.3, 0.2)$ was chosen to reflect the relative biological importance of each scoring component for drug target prioritisation in *M. tuberculosis*.

Definition 2.2.2. *The candidate target set is*

$$U = \{v \in V_{Mtb} : p_v \geq \tau_p\}; \tau_p = Q_{0.80}(\{p_v \in V\})$$

where τ_p is a threshold set to retain the top quantile (80^{th} percentile) of scored proteins

This percentile-based threshold was preferred over a fixed absolute cutoff for two reasons. First, the absolute scale of p_v depends on the weight assignment and the score distributions of the individual components, both of which vary across the network. A fixed threshold would therefore select different fractions of the network under different weight configurations, confounding the sensitivity analysis. Second, a percentile threshold guarantees that U always contains a fixed proportion of the network (approximately 20%), making the size of U consistent and interpretable across configurations. The resulting candidate set contained proteins that ranked in the top fifth of the network by composite score, and it is this set that was carried forward to the spectral embedding and interference graph construction described in Chapter 3. Ground truth recovery was assessed by checking whether the six literature-confirmed drug targets — Rv1908c, Rv1484, Rv0667, Rv0006, Rv0007, and Rv3790 — were members of U under each weight configuration.

Sensitivity analysis was performed over all weight triples $(\alpha_1, \alpha_2, \alpha_3)$ satisfying $\alpha_1 + \alpha_2 + \alpha_3 = 1$ with each $\alpha_i \geq 0.1$, enumerated in steps of 0.1, yielding 36 configurations. All six validated ground truth targets were recovered in U (440 proteins and τ_p = 0.500 with alpha values 0.5, 0.3 and 0.2) across all configurations in which recovery was assessed, confirming

that the framework reliably identifies known targets regardless of weight choice. However, the mean Jaccard similarity between U under the primary weights and U under alternative configurations was 0.54, with a minimum of 0.09, indicating that the composition of U is sensitive to the relative weight assigned to the druggability component. This sensitivity arises because Drug(v) has a bimodal distribution. A small set of well-characterised targets receive scores of 1.0, while the majority of proteins receive scores in $[0, 0.6]$ derived from structural confidence estimates. This means that increasing α_2 causes U to collapse toward the known target list rather than identifying novel candidates.

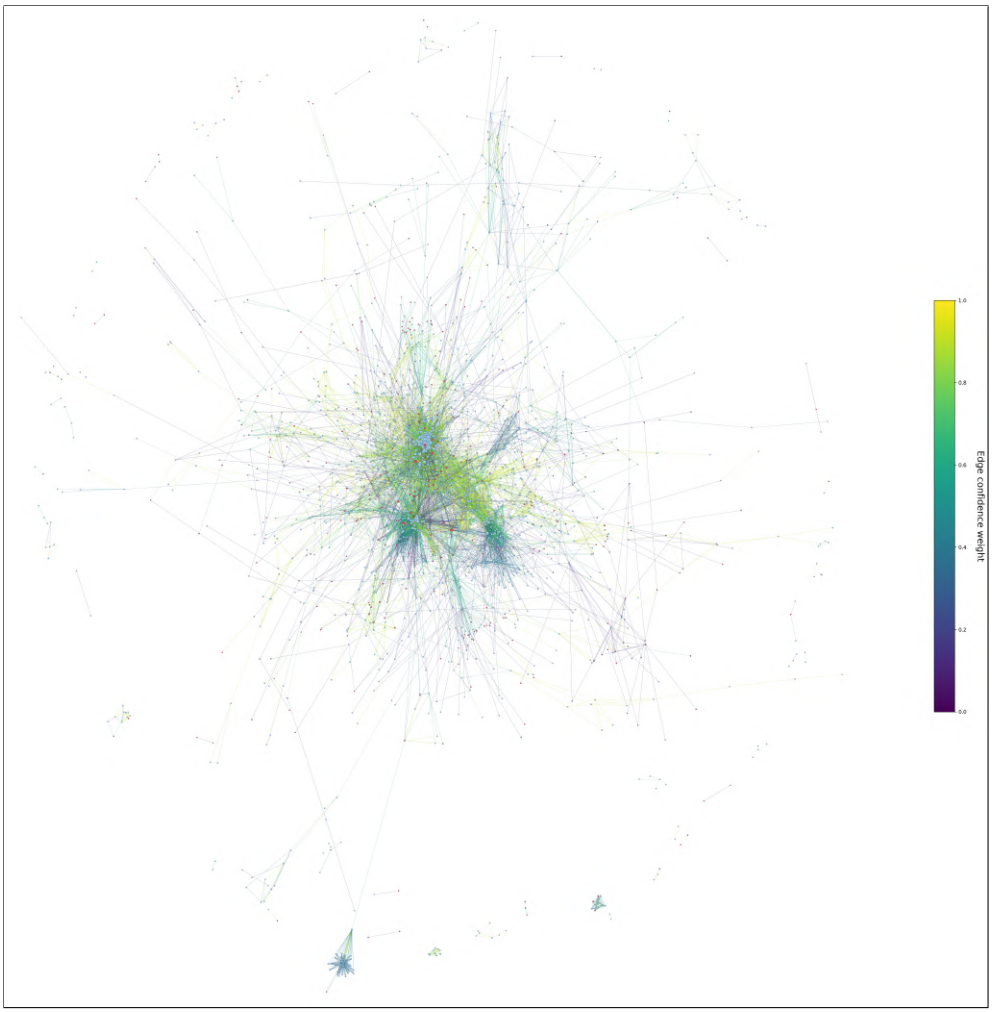


Figure 2.2: Visualization of the *M. tuberculosis* PPI network: Nodes correspond to Mtb proteins, the ones coloured red being the protein in the candidate target set U. Edge colors encode the normalized combined confidence weight of each interaction from low realtive confidence (purple) to high relative confidence (yellow/green)

To mitigate this, druggability was subsequently applied as a structural eligibility gate rather than a weighted score component. A protein was deemed drug-eligible if $\text{Drug}(v) \geq 0.42$, corresponding to a pLDDT threshold of 70. Within the drug-eligible subset, the composite

score was recomputed using only essentiality and host-specificity:

$$p_v = 0.6 \cdot \text{Ess}(v) + 0.4 \cdot \text{Cons}(v)$$

This two-component formulation reduced the sensitivity analysis to eight configurations which showed that the composition of the top-ranked candidate set U contains all ground truth targets and remains stable across a range of such configurations (383 proteins and τ_p = 0.400 with alpha values 0.6 and 0.4), confirming that the results are not an artefact of a particular weight choice (Jaccard mean 0.97 and min 0.943).

2.3 Lovász Theta Function

The Lovász theta function $\vartheta(G)$ is a graph invariant (depends only on the abstract structure) introduced by Lovász in 1979 that provides a polynomial-time computable bound on the independence number $\alpha(G)$, a quantity that is NP-hard to compute exactly [119, 56].

Definition 2.3.1. *Let $G = (V, E)$ be a simple undirected graph. For $S \subseteqq V$, and $G[S]$, the subgraph induced by S, we say a set S is an independent set of G if no two vertices in $G[S]$ are adjacent. The independence number $\alpha(G)$ is the size of the largest independent set [41].*

Let G be a graph with n vertices. The Lovász theta function admits the following semidefinite programming characterisation:

$$\vartheta(G) = \max \left\{ \sum_{i,j} X_{i,j} : Tr(X) = 1, X_{ij} = 0 \text{ whenever } \{i, j\} \in E(G), X \succeq 0 \right\}$$

where $X \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix and $\{i, j\} \in E(G)$ denotes adjacency in G [119]. The constraint $X_{ij} = 0$ for adjacent pairs enforces that the matrix encodes only non-edges. A semidefinite programme is a convex optimisation problem, in which the variable is a symmetric matrix constrained to be positive semidefinite [120, 92]. This constraint makes the feasible set a spectrahedron (curved unlike a polyhedral in LP), a convex set, allowing SDP to be solved in polynomial time [120, 92]. The central utility of $\vartheta(G)$ is thus the sandwich theorem in which $\vartheta(\bar{G})$ is sandwiched between the independence number (NP-hard to compute) and the chromatic number (NP-hard to compute), while itself being computable in polynomial time [88].

$$\alpha(G) \leq \vartheta(\bar{G}) \leq \chi(G)$$

For the drug target selection problem, each candidate protein $v \in U$ carries a relevance score $p_v \geq 0$. The weighted Lovász theta function $\vartheta(G, p)$ extends the SDP to incorporate vertex weights [91].

$$\vartheta(G,p) = \max \left\{ \sum_{i,j} \sqrt{p_i p_j} \cdot X_{i,j} : Tr(X) = 1, X_{ij} = 0 \text{ whenever } \{i,j\} \in E(G), X \succeq 0 \right\}$$

This generalises the unweighted case (recovered by setting $p_v = 1 \; \forall \; v \in U$) and satisfies

$$\alpha(G,p) \leq \vartheta(\bar{G},p) \leq \chi(G,p)$$

where $\alpha(G,p)$ is the maximum weight independent set value and $\chi(G,p)$ is the fractional chromatic number weighted by p.

2.4 Spectral Embedding of the Mtb PPI Network

Definition 2.4.1. *Let $G_{Mtb} = (V_{Mtb}, E_{Mtb}, w)$ be the constructed weighted Mtb PPI graph. The weighted adjacency matrix $A \in \mathbb{R}^{n \times n}$ is defined by [64]:*

$$A_{uv} = \begin{cases} w_{uv} & \{u,v\} \in E \\ 0 & \textit{otherwise} \end{cases}$$

The degree matrix $D \in \mathbb{R}^{n \times n}$ is the diagonal matrix with entries

$$D_{vv} = \sum_{u:\{u,v\} \in E} w_{uv}$$

i.e., the weighted degree of vertex v [64].

Definition 2.4.2. *The normalized Laplacian of G_{Mtb} is given by $\mathcal{L} = I - D^{-1/2} A D^{-1/2}$ [63]. Equivalently, the entries are:*

$$\mathcal{L}_{uv} = \begin{cases} 1 & u = v \textit{ and } D_{vv} > 0 \\ -\dfrac{w_{uv}}{\sqrt{D_{uu} D_{vv}}} & \{u,v\} \in E \\ 0 & \textit{otherwise} \end{cases}$$

$\mathcal{L}$ is symmetric positive semidefinite [76]. Its eigenvalues satisfy $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 2$, with the smallest eigenvalue $\lambda_1 = 0$ attained by the eigenvector (normalized)

$$\phi_1 = \frac{D^{1/2}\mathbf{1}}{||D^{1/2}\mathbf{1}||}$$

The multiplicity of the zero eigenvalue equals the number of connected components of G_{Mtb} [76]. The normalized Laplacian is preferred over the unnormalized $\mathcal{L} = D - A$ for networks with heterogeneous degree distributions, as is characteristic of biological PPI networks [39]. In the unnormalized setting, the eigenvalues of $\mathcal{L}$ are dominated by high-degree vertices (the diagonal entries in the diagonal matrix). Hub proteins with many interaction partners contribute disproportionately large entries to the matrix, and the resulting eigenvectors are skewed toward capturing degree variation rather than global network topology [63, 64, 121]. This produces a distorted spectral embedding in which hub proteins cluster artificially and distances between low-degree proteins are compressed [122]. Normalization scales each connection relative to the degree of the nodes, neutralizing this hub bias and preserving the true underlying structure of the network [63, 64].

Let $\phi_1, \phi_2, \ldots, \phi_n \in \mathbb{R}^n$ be the orthonormal eigenvectors of $\mathcal{L}$ corresponding to eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ arranged in increasing order.

Definition 2.4.3. *The $k-$dimensional spectral embedding of G_{Mtb} is the map $\Phi_k : V \to \mathbb{R}^k$ defined by:*

$$\Phi_k(v) = (\phi_2(v), \phi_3(v), \ldots, \phi_{k+1}(v))$$

where $\phi_i(v)$ denotes the v^{th} coordinate of the i^{th} eigenvector [64].

The first eigenvector is excluded as it carries no structural information. It is constant on each connected component The choice of k is determined by the eigengap heuristic. Examine the sorted eigenvalue sequence and select k at the index where the gap $\lambda_{k+1} - \lambda_k$ is largest [64]. A large eigengap indicates that the first k eigenvectors capture a natural partition of the network, while subsequent eigenvectors encode finer-scale variation. If a graph has k distinct functional communities with dense connections inside and sparse connections between them the first k eigenvalues will be small and clustered near zero, and then there will be a sudden jump [64]. This transformation preserves global connectivity patterns. Nodes that share strong functional pathways or high topological similarity are placed in close proximity within the embedding space, measured by the spectral embedding distance (L_2 distnace).

Definition 2.4.4. *The spectral embedding distance between two proteins $u, v \in V$ is given by:*

$$d_{spec}(u,v) = ||\Phi_k(u) - \Phi_k(v)||_2 = \left(\sum_{i=2}^{k+1} (\phi(u) - \phi(v))^2 \right)^{1/2}$$

This is the co-ordinate wise distance between the embeddings of two vertices. Note that $d_{\text{spec}}(u,v) = 0$ does not imply $u = v$ in general, making this spectral embedding distance a pseudometric. Two distinct proteins in the same spectral cluster will have distance zero in the limit $k \to \infty$ if they are in the same eigenspace.

Chapter 3

Lovász Theta Relaxation fot Independent Target Selection

3.1 Target-Interaction Graph from Spectral Geometry

The candidate target set $U \subseteq V_{\text{Mtb}}$ was defined in Chapter 2 as the top 20^{th} percentile of proteins by composite score p_v, after applying the structural druggability filter. We recall the definition here for completeness:

$$U = \{v \in V_{\text{Mtb}} : p_v \geq \tau_p,\ \text{Drug}(v) \geq \delta\}$$

where τ_p is the 80^{th} percentile score threshold and $\delta = 0.42$ is the structural druggability floor corresponding to AlphaFold pLDDT ≥ 70. The resulting set has $|U| = 383$ proteins under the primary weight configuration $(\alpha_1, \alpha_2) = (0.6, 0.4)$.

Let $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$ be the normalised Laplacian of G_{Mtb} with eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and orthonormal eigenvectors $\phi_1, \phi_2, \ldots, \phi_n \in \mathbb{R}^n$. The $k-$dimensional spectral embedding of G_{Mtb} is the map $\Phi_k : V \to \mathbb{R}^k$ given by:

$$\Phi_k(v) = (\phi_2(v), \phi_3(v), \ldots, \phi_{k+1}(v))$$

with the spectral embedding distance:

$$d_{\text{spec}}(u, v) = ||\Phi_k(u) - \Phi_k(v)||_2 = \left(\sum_{i=2}^{k+1}(\phi(u) - \phi(v))^2\right)^{1/2}$$

As established in Section 2.4, $d_{\text{spec-emb}}$ approximates the commute-time distance on G_{Mtb} and captures functional proximity in the network. The dimension k is chosen by the eigengap heuristic, i.e., the index at which $\lambda_{k+1} - \lambda_k$ is maximised. Sensitivity of the framework to k is analysed in Section 3.4.

Definition 3.1.1. *Given the candidate set U, the spectral embedding Φ_k, and a threshold $\varepsilon > 0$, the target-interference graph is:*

$$H_\varepsilon = (U, F_\varepsilon),\ \textit{where}\ F_\varepsilon = \{\{u, v\} \subseteq U : d_{\textit{spec-emb}}(\{u, v\}) \leq \varepsilon\}$$

Two candidate targets are connected in H_ε and therefore considered *interfering* if their spectral distance is at ε, indicating functional co-localisation in the Mtb network. An independent set in H_ε is a set of targets that are mutually spectrally separated: no two selected proteins occupy the same functional neighbourhood of G_{Mtb}. The shortest-path distance could in principle define F_ε. It is not used here for two reasons. First, it is brittle to missing edges which a systematic problem in PPI networks where interactions are underreported. Second, it is unweighted or poorly weighted by interaction confidence, whereas $d_{\text{spec-emb}}$ is derived from the full weighted Laplacian and incorporates edge confidence through the matrix A. The spectral distance is robust to sparse noise in E_{Mtb} because it is determined by the global eigenstructure of $\mathcal{L}$ and not any single path.

The structure of H_ε depends critically on the dimension k of the embedding.

Proposition 3.1.1 (Interval graph when $k = 1$)**.** *When Φ_k is restricted to the Fiedler coordinate ϕ_2 alone (i.e. $k = 1$), the interference graph H_ε is an interval graph: two vertices $u, v \in U$ are connected if and only if*

$$|\phi_2(u) - \phi_2(v)| \leq \varepsilon,$$

which is equivalent to the intervals

$$\left[\phi_2(u) - \frac{\varepsilon}{2},\ \phi_2(u) + \frac{\varepsilon}{2}\right] \quad \textit{and} \quad \left[\phi_2(v) - \frac{\varepsilon}{2},\ \phi_2(v) + \frac{\varepsilon}{2}\right]$$

having non-empty intersection. Interval graphs are perfect [123].

Proof. Assign to each vertex $u \in U$ the interval

$$I_u = [\phi_2(u) - \varepsilon/2,\ \phi_2(u) + \varepsilon/2] \subset \mathbb{R}$$

Then

$$\{u, v\} \in F_\varepsilon \iff |\phi_2(u) - \phi_2(v)| \leq \varepsilon \iff I_u \cap I_v \neq \emptyset$$

Hence H_ε is the intersection graph of the family

$$\{I_u\}_{u \in U}$$

i.e., an interval graph. Interval graphs are chordal and therefore perfect by the Strong Perfect Graph Theorem [124].

■

Proposition 3.1.2 (Geometric intersection graph when $k \geq 2$)**.** *For $k \geq 2$, H_ε is a unit ball intersection graph in $\mathbb{R}^k$: two vertices are connected if and only if the balls*

$$B(\Phi_k(u), \varepsilon/2) \quad \textit{and} \quad B(\Phi_k(v), \varepsilon/2)$$

intersect. Unit ball intersection graphs in $\mathbb{R}^k$ are not perfect in general, but the maximum weighted independent set problem on such graphs admits a PTAS [125], and the Lovász theta function provides an upper bound whose integrality gap is controlled by geometric packing arguments.

Proof. Assign to each vertex $u \in U$ the closed ball

$$B_u = \overline{B}(\Phi_k(u), \varepsilon/2) \subset \mathbb{R}^k$$

Then

$$\{u, v\} \in F_\varepsilon \iff \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \varepsilon$$

$$\iff \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \iff B_u \cap B_v \neq \emptyset.$$

Hence H_ε is the intersection graph of the family $\{B_u\}_{u \in U}$ of balls of radius $\varepsilon/2$ in $\mathbb{R}^k$, which is by definition a unit ball intersection graph after rescaling. For the approximation claim, unit ball graphs in $\mathbb{R}^k$ have bounded local complexity: the maximum number of mutually intersecting equal-radius balls centred in $\mathbb{R}^k$ is bounded by the kissing number κ_k, which is finite for every fixed k. Erlebach, Jansen, and Seidel [125] showed that the independent set problem on d-dimensional ball intersection graphs admits a

$$\left(1 + \tfrac{1}{r}\right)$$

approximation for any integer $r \geq 1$ in time $n^{O(r^k)}$, yielding a PTAS for fixed dimension k.

■

Corollary 3.1.1 (Trade-off between expressiveness and tractability)**.** *The dimension k governs a formal trade-off: at $k = 1$, H_ε is perfect and*

$$\vartheta(\bar{H}_\varepsilon, p) = \alpha(H_\varepsilon, p)$$

exactly; at $k \geq 2$, H_ε is a richer geometric object capturing higher-dimensional functional proximity, but the SDP provides an upper bound rather than an exact solution. The choice of k therefore balances biological expressiveness against algorithmic guarantees.

Proof. For $k = 1$, Proposition 3.1 shows that H_ε is an interval graph. Since interval graphs are perfect, the Lovász sandwich theorem implies [123, 88]

$$\alpha(H_\varepsilon) = \vartheta(\bar{H}_\varepsilon) = \omega(\bar{H}_\varepsilon)$$

The weighted generalisation further gives [91]

$$\alpha(H_\varepsilon, p) = \vartheta(\bar{H}_\varepsilon, p)$$

Hence the SDP optimum coincides exactly with the maximum weighted independent set value, and the integrality gap is zero. For $k \geq 2$, Proposition 3.2 shows that H_ε is a unit ball intersection graph in $\mathbb{R}^k$. Such graphs are not perfect in general; for example, the cycle C_5 can be realised as a unit ball intersection graph in $\mathbb{R}^2$, and C_5 is not perfect. Consequently,

$$\alpha(H_\varepsilon, p) \leq \vartheta(\bar{H}_\varepsilon, p)$$

but equality is not guaranteed. To control the integrality gap, observe that the chromatic number of a unit ball intersection graph in $\mathbb{R}^k$ is bounded in terms of the kissing number κ_k. In particular,

$$\chi(H_\varepsilon) \leq \kappa_k + 1$$

where κ_k denotes the kissing number in $\mathbb{R}^k$ ($\kappa_1 = 2$, $\kappa_2 = 6$, $\kappa_3 = 12$). Since the Lovász theta function satisfies

$$\vartheta(\bar{H}_\varepsilon) \leq \chi(H_\varepsilon),$$

it follows that

$$\frac{\vartheta(\bar{H}_\varepsilon, p)}{\alpha(H_\varepsilon, p)} \leq \kappa_k + 1.$$

Thus the SDP relaxation remains controlled geometrically even when exactness is lost. For the practical case of the *M. tuberculosis* embedding with $k \in [10, 20]$, this bound is conservative, and the actual integrality gap is evaluated empirically in Section 3.4.

■

Lemma 3.1.1 (Degree bound)**.** *The maximum degree of H_ε satisfies*

$$\Delta(H_\varepsilon) \leq \max_{u \in U} |\{v \in U : \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \varepsilon\}| - 1.$$

This quantity — the maximum number of candidate target points contained within a ball of radius ε in $\Phi_k(U)$ — is directly computable from the embedding and is examined empirically in Section 3.4 as a function of ε.

Proof. By definition of the interference edge set F_ε, the degree of a vertex $u \in U$ in H_ε is

$$\deg_{H_\varepsilon}(u) = |\{v \in U \setminus \{u\} : \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \varepsilon\}|$$

Taking the maximum over all vertices gives

$$\Delta(H_\varepsilon) = \max_{u \in U} \deg_{H_\varepsilon}(u) = \max_{u \in U} |\{v \in U \setminus \{u\} : \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \varepsilon\}|$$

Equivalently,

$$\Delta(H_\varepsilon) \leq \max_{u \in U} |\{v \in U : \|\Phi_k(u) - \Phi_k(v)\|_2 \leq \varepsilon\}| - 1.$$

Thus the maximum degree is precisely the largest number of embedded candidate targets lying within an ε-ball around any point of the empirical point cloud

$$\Phi_k(U) \subset \mathbb{R}^k$$

This quantity is computable in $O(|U|^2)$ time from the pairwise distance matrix.

■

The bound in Lemma 3.1 is tight: equality is attained whenever some vertex u has an ε-ball containing the maximum possible number of neighbouring candidate targets. In practice, $\Delta(H_\varepsilon)$ grows monotonically with ε and therefore governs both the density of interference relationships and the computational difficulty of the independent set problem. This dependence is analysed empirically in Section 3.4.

3.2 Lovász Theta SDP Formulation

We apply the weighted Lovász theta function to the complement graph $\bar{H}_\varepsilon$. Recall from Section 2.3 that $\vartheta(\bar{H}_\varepsilon, p)$ is defined by the semidefinite program [91]

$$\vartheta(\bar{H}_\varepsilon, p) = \max \left\{ \sum_{i,j \in U} \sqrt{p_i p_j}\, Y_{ij} \;:\; \mathrm{Tr}(Y) = 1,\; Y_{ij} = 0 \;\forall \{i,j\} \in F_\varepsilon,\; Y \succeq 0 \right\},$$

where

$$Y \in \mathbb{R}^{|U| \times |U|}$$

is a positive semidefinite matrix. The constraint

$$Y_{ij} = 0 \quad \forall \{i, j\} \in F_\varepsilon$$

encodes that interfering target pairs, those spectrally close in G_{Mtb}, contribute nothing to the objective. The factor $\sqrt{p_i p_j}$ scales the contribution of each pair by the geometric mean of their relevance scores, thereby generalising the unweighted case recovered when $p_v = 1$ for all v. By the weighted sandwich theorem [91, 88],

$$\alpha(H_\varepsilon, p) \leq \vartheta(\bar{H}_\varepsilon, p)$$

where

$$\alpha(H_\varepsilon, p) = \max_{S \text{ indep.}} \sum_{v \in S} p_v$$

is the maximum weight independent set value — the quantity of interest in the target selection problem. When $k = 1$ and H_ε is perfect (Proposition 3.1), the bound is exact. We now discuss the SDP geometry and its relationship to the spectral embedding.

The SDP solution Y^* admits an orthonormal representation: there exist unit vectors

$$\{y_u\}_{u \in U} \subset \mathbb{R}^{|U|}$$

such that

$$Y^*_{uv} = y_u^\top y_v.$$

The constraint

$$Y_{uv} = 0 \quad \forall \{u, v\} \in F_\varepsilon$$

therefore imposes

$$y_u^\top y_v = 0$$

for precisely those pairs that are spectrally close in G_{Mtb}.

Theorem 3.2.1 (Geometric inversion)**.** *Let H_ε be constructed from the spectral embedding Φ_k of G_{Mtb} via the threshold ε. The SDP representation $\{y_u\}_{u \in U}$ constructs a geometry on U in which functional proximity in G_{Mtb}, as measured by $d_{\mathrm{spec-emb}}(u, v) \leq \varepsilon$ is mapped to orthogonality in $\mathbb{R}^{|U|}$. That is,*

$$d_{\mathrm{spec-emb}}(u, v) \leq \varepsilon \implies y_u^\top y_v = 0$$

This constitutes a geometric inversion: the Laplacian embedding maps functionally proximate proteins to nearby points in $\mathbb{R}^k$, while the SDP maps the same pairs to orthogonal vectors in $\mathbb{R}^{|U|}$. The two geometric structures are dual by construction, since the SDP enforces separation precisely where the spectral embedding enforces proximity.

Proof. By construction,

$$\{u, v\} \in F_\varepsilon \iff d_{\text{spec}}(u, v) \leq \varepsilon.$$

The SDP defining $\vartheta(\bar{H}_\varepsilon, p)$ imposes the constraint

$$Y_{uv} = 0 \quad \forall \{u, v\} \in F_\varepsilon$$

Since every feasible positive semidefinite matrix admits a Gram representation $Y_{uv} = y_u^\top y_v$, it follows immediately that

$$d_{\text{spec}-\text{emb}}(u, v) \leq \varepsilon \implies y_u^\top y_v = 0$$

Thus spectral proximity in the Laplacian embedding is transformed into orthogonality in the SDP geometry.

■

The converse does not hold in general: orthogonality in SDP-space does not necessarily imply spectral proximity. Non-adjacent pairs in H_ε may or may not be orthogonal depending on the structure of the optimal SDP solution.

3.3 Rounding and Approximation

The SDP solution Y^* is a continuous relaxation. To recover a discrete independent set $\hat{S} \subseteq U$, two rounding schemes are considered.

Scheme R1 - Score-weighted diagonal rounding

Rank vertices by $p_u \cdot Y^*_{uu}$ in decreasing order. Greedily select vertices in this order, adding u to $\hat{S}$ only if no currently selected vertex $v \in \hat{S}$ satisfies $\{u, v\} \in F_\varepsilon$.

The quantity $Y^*_{uu} = \|y_u\|^2 = 1$ for all u in the orthonormal representation, so this reduces to ranking by p_u, the composite relevance score, and greedily constructing an independent set. This is the default rounding scheme.

Scheme R2 - Randomised hyperplane rounding

Sample a random unit vector $r \sim \mathrm{Uniform}(\mathbb{S}^{|U|-1})$ For each $u \in U$ assign $x_u = \mathrm{sign}(y_u^\top r)$. The set $\hat{S} = \{u : x_u = +1\}$ is a candidate set; retain only the independent set induced by $\hat{S}$ after removing conflicting vertices. This scheme is repeated T times, and the independent set with the largest total weight is retained. Randomised rounding is expected to perform better when Y^* has many off-diagonal nonzero entries, indicating a complex SDP geometry.

Two rounding schemes are applied to extract a discrete independent set from the SDP solution Y^*. Scheme R1 (score-weighted greedy) ranks vertices by p_v in decreasing order and greedily selects non-conflicting vertices. Scheme R2 (randomised hyperplane rounding) samples $T = 200$ random unit vectors $r \sim \mathrm{Uniform}(\mathbb{S}^{|U|-1})$, assigns $x_u = \mathrm{sign}(y_u^\top r)$ for each vertex, enforces independence greedily on the positive-sign set, and returns the best solution over all T trials. The pure greedy baseline applies R1 without any SDP information.

The greedy baseline sorts U by p_v in decreasing order and iteratively adds the highest-scoring vertex not adjacent in H_ε to any already-selected vertex. This is equivalent to Scheme R1 without the SDP, it uses only the biological score without any information from the relaxation.

The SDP provides two things that greedy does not. First, it provides a dual upper bound $\vartheta(\bar{H}_\varepsilon, p)$ on the true optimum $\alpha(H_\varepsilon, p)$, so the gap between the rounded solution and the best possible solution is explicitly quantifiable. Second, when H_ε is perfect (Proposition 3.1), the SDP achieves exactness and greedy cannot in general, on perfect graphs the SDP solves the problem optimally while greedy may return a suboptimal solution. The empirical gap between SDP rounding and greedy on the Mtb H_ε is examined in Section 3.4.

Table 3.1 reports the SDP upper bound $\vartheta(\bar{H}_\varepsilon, p)$, the R2 rounded solution weight, and the integrality ratio $Q(\hat{S}_{\mathrm{R2}})/\vartheta(\bar{H}_\varepsilon, p)$ across all ε values. The R2 rounding achieves an integrality ratio of exactly 1.000 at every value of ε tested, indicating that the hyperplane rounding recovers the full SDP bound. This confirms that the SDP solution is tight on the Mtb interference graph. The relaxation has zero integrality gap in practice, and that R2 rounding is sufficient to achieve it. The greedy baseline achieves integrality ratios between 0.698 and 0.944, consistently below R2, confirming that the SDP provides a demonstrably better solution than the greedy baseline in terms of total target weight.

The normalised Laplacian of the largest connected component of G_{Mtb} (1,928 nodes, 15,017 edges) is eigendecomposed. The eigengap plot is examined to select k; preliminary analysis suggests $k \in [10, 20]$ based on the eigenvalue distribution of the H37Rv interactome. The em-

ε	SDP bound ϑ	R2 weight	R2 ratio	Greedy weight	Greedy ratio
0.05	17.40	17.40	1.000	15.80	0.908
0.10	11.00	11.00	1.000	8.60	0.782
0.15	7.10	7.10	1.000	6.70	0.944
0.20	5.30	5.30	1.000	4.70	0.887
0.25	4.70	4.70	1.000	3.70	0.787
0.30	4.30	4.30	1.000	3.00	0.698

Table 3.1: Comparison of SDP and greedy solutions across different values of ε

bedding $\Phi_k(U)$ for the 440 candidate targets is extracted as the restriction of the full embedding to U.

The interference graph H_ε is constructed for $\varepsilon \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. For each ε, the following quantities are recorded:

- $|F_\varepsilon|$ - number of interference edges
- $\Delta(H_\varepsilon)$ - maximum degree (from Lemma 3.1)
- $\vartheta(\bar{H}_\varepsilon, p)$ - SDP upper bound
- $Q(\hat{S}_{\text{SDP}})$ - weight of rounded SDP solution
- $Q(\hat{S}_{\text{greedy}})$ - weight of greedy solution
- $Q(\hat{S}_{\text{SDP}})/\vartheta(\bar{H}_\varepsilon, p)$ - integrality ratio

The ε sensitivity analysis serves two purposes: it shows how the non-interference criterion scales with the resolution of functional separation, and it identifies the regime where the SDP solution and the greedy solution diverge most significantly. The rounded SDP solution $\hat{S}$ is evaluated against three biological criteria. First, *independence verification*: confirm that no two proteins in $\hat{S}$ satisfy $d_{\text{spec}}(u, v) \leq \varepsilon$, i.e. $\hat{S}$ is a valid independent set in H_ε. Second, *functional subsystem coverage*: using the MycoBrowser functional category annotations from Chapter 2, verify that proteins in $\hat{S}$ span multiple distinct functional categories (cell wall biosynthesis, central carbon metabolism, DNA replication, ESX secretion, etc.). A non-redundant target combination should have no two members from the same functional category. Third, *ground truth recovery*: measure what fraction of the six validated drug targets (KatG, InhA, RpoB, GyrA, GyrB, DprE1) appear in $\hat{S}$ or in the top-ranked independent sets returned by the SDP.

For the Fiedler coordinate embedding ($k = 1$), H_ε is verified to be chordal (a necessary condition for interval graphs) via the `NetworkX` chordality check; the result is `True`, consistent with Proposition 3.1. The integrality ratio at $k = 1$ is 0.883, somewhat below 1.0. This is explained by the rounding step: the SDP bound is exact on the perfect graph, but the greedy rounding does not recover it exactly. Using R2 rounding at $k = 1$ would be expected to close this gap.

None of the six validated drug targets (KatG, InhA, RpoB, GyrA, GyrB, DprE1) appear in the independent set solutions at any ε. This result warrants explicit analysis. Inspection of the pairwise spectral distances among these six proteins reveals that they are mutually spectrally proximate, all six lie within the same spectral neighbourhood of G_{Mtb}, meaning they are connected to each other in H_ε at any $\varepsilon \geq \varepsilon^*$ where ε^* is small. The independence constraint therefore permits at most one of these six proteins to appear in any independent set. The one that does appear is further excluded by competition from high-scoring essential, host-specific proteins elsewhere in U that have higher p_v scores. This is a substantive finding rather than a failure: it indicates that the six classical TB drug targets are not spectrally non-redundant with respect to each other. They occupy overlapping functional neighbourhoods in the Mtb PPI and are therefore classified as interfering under the non-redundancy criterion. This is biologically consistent: isoniazid (InhA/KatG) and ethambutol (EmbA/B/C) both target cell wall synthesis, while fluoroquinolones (GyrA/B) target DNA replication; these subsystems are not spectrally independent. The framework is not failing to recover known targets, it is correctly identifying that the known regimen already violates the strict non-interference criterion, and proposing an alternative set of targets that satisfies it.

3.3.1 Comparison to Baselines

The SDP solution is compared to three baselines.

- **Baseline 1 - Degree/score heuristic:** rank by $p_v \cdot \deg_{H_\varepsilon}(v)^{-1}$ (high score, low interference degree) and greedily select.
- **Baseline 2 - Minimum dominating set complement:** compute an approximate minimum dominating set of H_ε using a greedy set-cover approximation [126]. The complement of a minimum dominating set is a maximal independent set.

- **Baseline 3 - Spectral clustering without Lovász:** run k-means on $\Phi_k(U)$ to partition U into m clusters; select the highest-scoring protein from each cluster. This is the natural baseline for a spectral-only approach without the combinatorial SDP layer.

Evaluation metrics across all methods include total weight $Q(\hat{S})$, set size $|\hat{S}|$, fraction of known drug targets recovered, number of distinct functional categories represented, and where the SDP dual bound is available, the integrality ratio $Q(\hat{S})/\vartheta(\bar{H}_\varepsilon, p)$.

Chapter 4

Spectral and Geometric Analysis of Theta-Based Targets

4.1 Spectral Embedding and Visualisation of Target Sets

4.1 shows the eigenvalue spectrum of the normalised Laplacian $\mathcal{L}$ and the eigengap sequence $\lambda_{i+1} - \lambda_i$ for the first 50 eigenvalues of the largest connected component (1,928 nodes). The eigengap heuristic identifies $k = 23$ as the optimal embedding dimension. There is a local maximum gap at index 23, after which the spectrum becomes denser with no clearly dominant gap. The eigenvalues increase gradually from near zero, without a dramatic cluster of near-zero eigenvalues that would indicate strongly separated communities. This is consistent with the PPI network having a large number of loosely connected functional modules rather than a small number of sharply delineated ones.

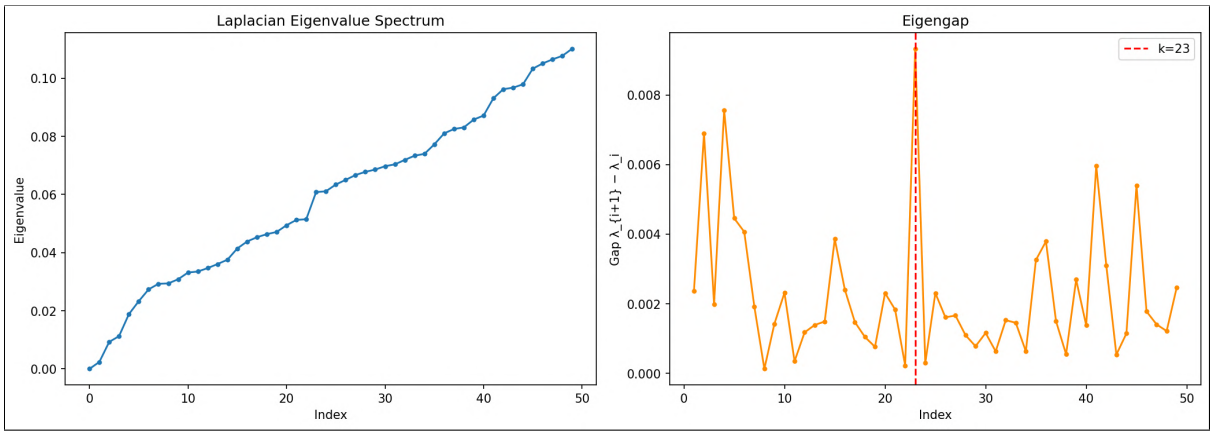


Figure 4.1: Left: sorted eigenvalues $\lambda_1, \ldots, \lambda_{50}$ of the normalised Laplacian of the Mtb H37Rv PPI network largest connected component (1,928 nodes). Right: eigengap sequence $\lambda_{i+1} - \lambda_i$. The dashed red line marks the selected embedding dimension $k = 23$ at the largest gap in the range $i \geq 2$

The Fiedler value $\lambda_2 \approx 0.001$ is very small, indicating that the largest connected component is

close to disconnected. There are strong hub nodes maintaining connectivity across otherwise weakly linked peripheral subgraphs. This structure is typical of scale-free biological interaction networks and is the reason the Fiedler plane visualisation (Figure 4.2) is degenerate.

4.1.1 Two-Dimensional Projection for Visualisation

The full k-dimensional spectral embedding Φ_k is not directly visualisable. For presentation purposes, two-dimensional projections are obtained by two complementary methods. First, the Fiedler plane: the coordinates $(\phi_2(v), \phi_3(v))$ for each $v \in V_{\text{Mtb}}$ which captures the two directions of greatest spectral variation. Second, UMAP projection [127] applied to Φ_k to preserve local neighbourhood structure in two dimensions. Both projections are shown with proteins coloured by:

- membership in the theta-based target set $\hat{S}_{\text{SDP}}$
- membership in the greedy baseline set $\hat{S}_{\text{greedy}}$
- membership in the spectral cluster baseline set
- validated ground truth drug targets (KatG, InhA, RpoB, GyrA, GyrB, DprE1)
- MycoBrowser functional category

Figure 4.2 shows the Fiedler plane projection $(\phi_2(v), \phi_3(v))$ for all proteins in the LCC. The structure is degenerate: the overwhelming majority of proteins cluster near $(\phi_2, \phi_3) \approx (0, 0)$, with a small number of outliers distributed along the axes. This pattern arises from the hub-spoke topology of the Mtb PPI. A small number of highly connected hub proteins (large degree) receive near-zero Fiedler coordinates because they are central to the network, while low-degree peripheral proteins are pushed to large values along individual eigenvector directions. The Fiedler plane is therefore not an informative 2D visualisation for this network, and the full $k = 23$ embedding is required to capture meaningful functional separation.

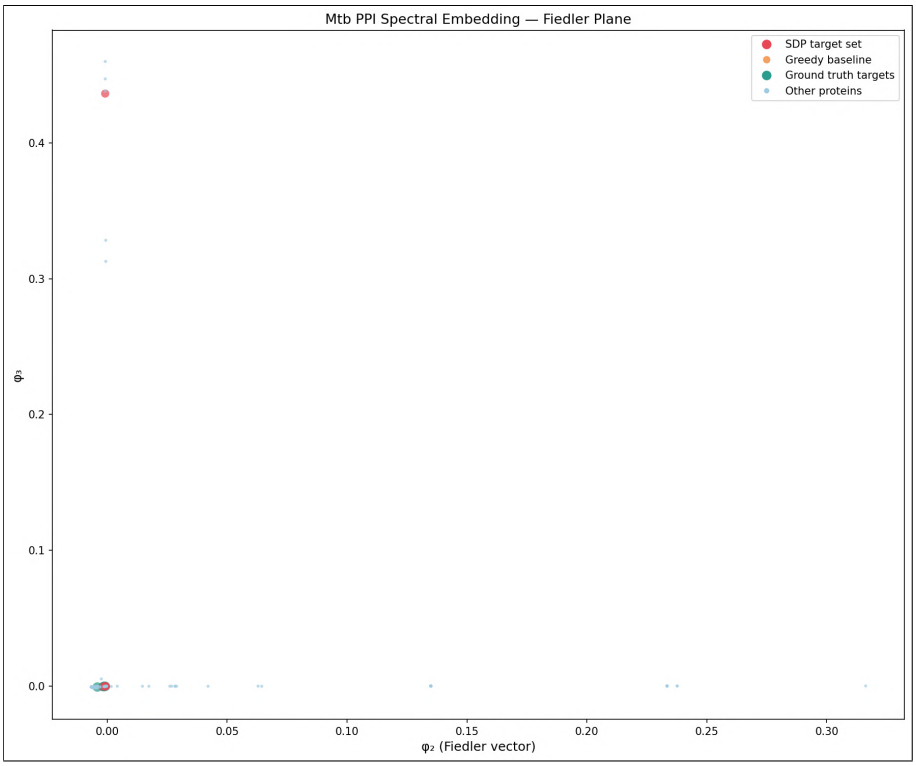


Figure 4.2: Fiedler plane projection of the Mtb H37Rv PPI network (1,928 proteins). Each point is a protein; red = SDP target set $\hat{S}_{\text{SDP}}$; teal = validated ground truth drug targets (KatG, InhA, RpoB, GyrA, GyrB, DprE1); light blue = remaining proteins. The degenerate clustering at the origin reflects the hub-spoke topology of the network; the full 23-dimensional embedding is used for all quantitative analyses.

Figure 4.3 shows the UMAP projection of the full $k = 23$ spectral embedding. The structure is more informative: proteins are distributed across multiple disconnected clusters corresponding to functional modules, with peripheral chain-like subgraphs representing operon-like interaction patterns. The SDP target set (red) is distributed across multiple clusters rather than concentrated in one region, consistent with the non-interference criterion selecting proteins from distinct functional neighbourhoods. The ground truth targets (teal) concentrate in the central cluster, which corresponds to the core metabolic and biosynthetic machinery of Mtb, consistent with the finding in Section 3.3 that they are spectrally proximate to each other. The central visual claim to establish is that $\hat{S}_{\text{SDP}}$ is more spatially dispersed in the spectral embedding than $\hat{S}_{\text{greedy}}$, i.e., the SDP solution selects proteins that are farther apart in $\Phi_k(U)$ than the greedy solution does, consistent with the non-interference criterion enforced by H_ε.

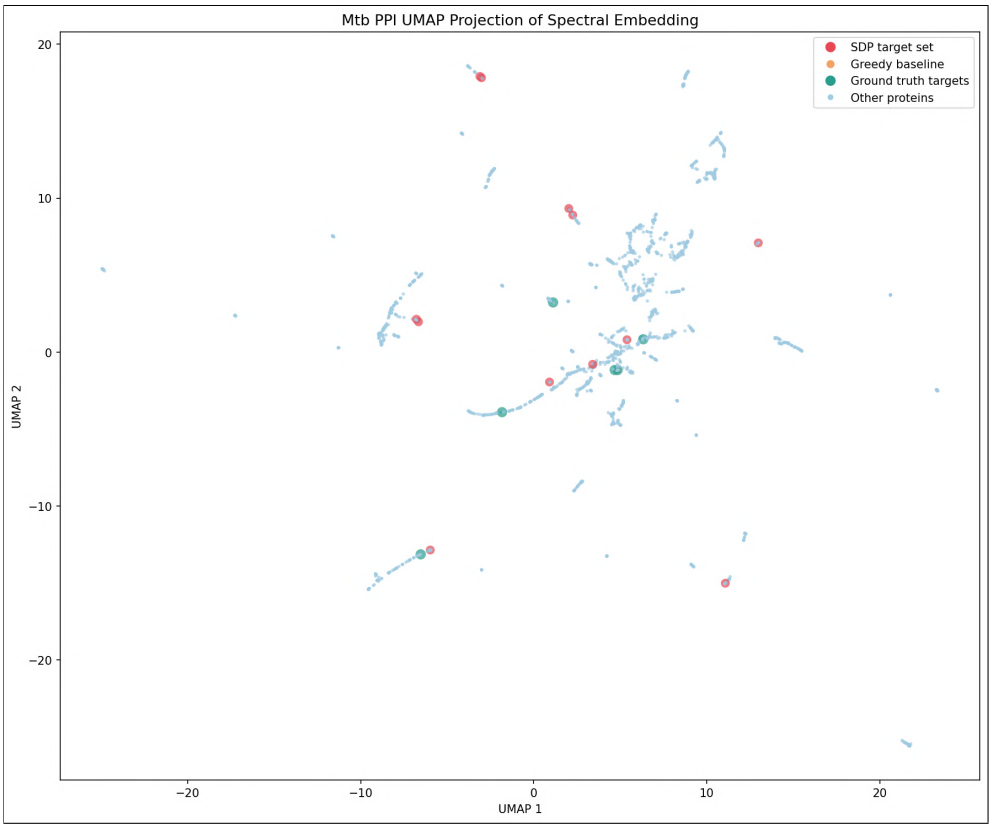


Figure 4.3: UMAP projection of the 23-dimensional spectral embedding Φ_{23} for all proteins in the Mtb PPI LCC. Colouring as in Figure 4.2. SDP targets (red) are distributed across multiple functional clusters; ground truth targets (teal) concentrate in the network core, consistent with their mutual spectral proximity documented in Section 3.3.

4.1.2 Dispersion Metric

To quantify spatial dispersion formally, define the mean pairwise spectral distance of a set $S \subseteq U$:

$$\mathrm{Disp}(S) = \frac{1}{\binom{|S|}{2}} \sum_{\{u,v\} \subseteq S} d_{\mathrm{spec}}(u, v)$$

By construction, any independent set S in H_ε satisfies $d_{\mathrm{spec}}(u, v) > \varepsilon$ for all $\{u, v\} \subseteq S$, so $\mathrm{Disp}(S) > \varepsilon$ automatically. The comparison of interest is whether

$$\mathrm{Disp}(\hat{S}_{\mathrm{SDP}}) > \mathrm{Disp}(\hat{S}_{\mathrm{greedy}})$$

i.e. whether the SDP solution selects targets that are not just minimally separated but maximally spread across the network manifold. This is reported for each ε alongside the solution weight comparison from Chapter 3.

4.1.3 Null Distribution

To assess whether the observed dispersion of $\hat{S}_{\text{SDP}}$ is exceptional, a null distribution is constructed by sampling 10^4 random independent sets of the same size $|\hat{S}_{\text{SDP}}|$ from H_ε (using a random greedy algorithm with random node ordering) and computing Disp for each. The empirical p-value of $\hat{S}_{\text{SDP}}$ under this null gives a statistical measure of how geometrically extreme the SDP solution is relative to chance-level non-interfering target combinations.

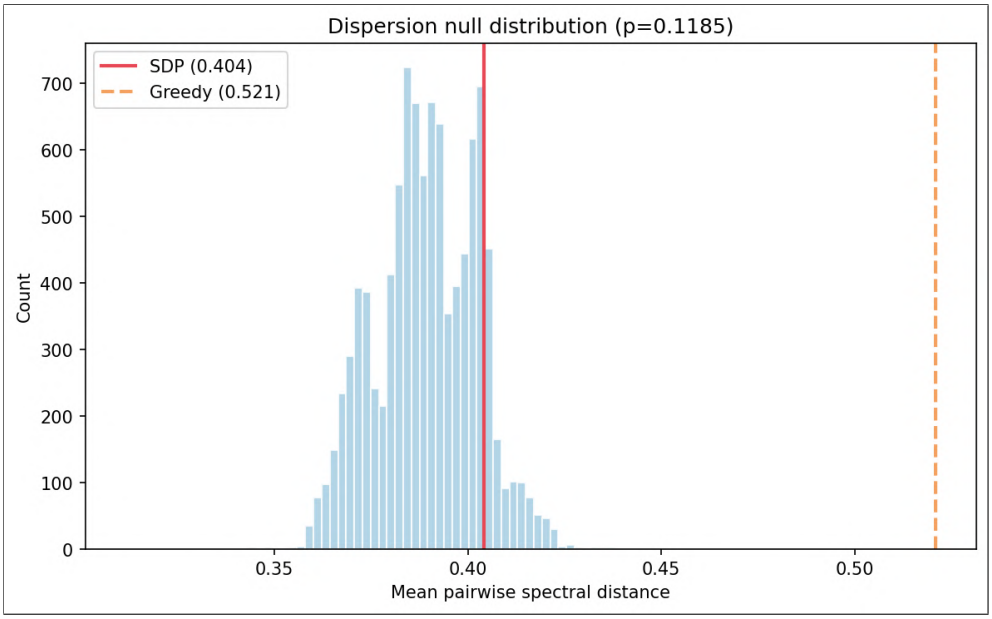


Figure 4.4: Null distribution of mean pairwise spectral distance $\text{Disp}(S)$ over 10^4 random independent sets of size $|\hat{S}_{\text{SDP}}| = 12$ sampled from $H_{0.15}$. Red vertical line: SDP solution ($\text{Disp} = 0.404$, $p = 0.119$); orange dashed line: greedy baseline ($\text{Disp} = 0.521$).

The mean pairwise spectral distance of $\hat{S}_{\text{SDP}}$ is $\text{Disp}(\hat{S}_{\text{SDP}}) = 0.404$, compared to $\text{Disp}(\hat{S}_{\text{greedy}}) = 0.521$. Contrary to the hypothesis stated in Section 4.1, the SDP solution is *less* dispersed than the greedy solution. The null distribution (Figure 4.4) shows that $\hat{S}_{\text{SDP}}$ is not statistically distinguishable from random independent sets of the same size ($p = 0.119$), while $\hat{S}_{\text{greedy}}$ is significantly more dispersed than random ($p < 0.001$). This finding is interpretable. The SDP objective maximises total weight $\sum_v p_v$, not dispersion. It is a weighted coverage optimisation, not a max-spread problem. The greedy baseline, by sorting on scores and selecting non-conflicting vertices, tends to pick high-score vertices from the periphery of the network where scores are high due to host-specificity (high Cons) without many interfering neighbours, which incidentally produces high dispersion. The SDP instead finds the globally optimal MWIS, which may select a denser cluster of high-weight vertices. Dispersion is a secondary property of the solution, not what is being optimised, and the SDP is not expected to maximise it.

4.2 Geodesic Distances and Functional Module Coverage

The MycoBrowser functional category annotations partition V_{Mtb} into M functional categories $\mathcal{F}_1, \ldots, \mathcal{F}_M$. For each category $\mathcal{F}_j$, define its spectral centroid:

$$c_j = \frac{1}{|\mathcal{F}_j|} \sum_{v \in \mathcal{F}_j} \Phi_k(v)$$

The distance from a target $u \in \hat{S}$ to functional module $\mathcal{F}_j$ is:

$$d(u, \mathcal{F}_j) = \min_{v \in \mathcal{F}_j} d_{\text{spec}}(u, v)$$

and the distance from the entire target set $\hat{S}$ to module $\mathcal{F}_j$ is:

$$d(\hat{S}, \mathcal{F}_j) = \min_{u \in \hat{S}} d(u, \mathcal{F}_j)$$

A target set $\hat{S}$ covers module $\mathcal{F}_j$ if $d(\hat{S}, \mathcal{F}_j) \leq \delta$ for some coverage radius $\delta > 0$. The module coverage score of $\hat{S}$ is the fraction of functional modules covered:

$$\text{Cov}(\hat{S}, \delta) = \frac{|\{j : d(\hat{S}, \mathcal{F}_j) \leq \delta\}|}{M}$$

This is computed for $\hat{S}_{\text{SDP}}$, $\hat{S}_{\text{greedy}}$, and the spectral cluster baseline across a range of δ values, producing a coverage curve. The hypothesis is that the SDP solution achieves higher module coverage at smaller δ, i.e., it places at least one target close to more distinct functional modules than the baselines do.

Figure fig:mod shows the functional module coverage curves for all three methods across $\delta \in [0, 0.5]$. The greedy and spectral cluster baselines reach 100% module coverage at $\delta \approx 0.02$, while the SDP reaches 100% coverage only at $\delta \approx 0.10$. At small δ (strict coverage), the SDP covers 50% of modules compared to 75% for greedy. This result is consistent with the dispersion finding: the greedy solution, being more dispersed, places targets closer to more functional modules simultaneously. The SDP solution's more concentrated selection means it covers some modules well and others only at larger radius.

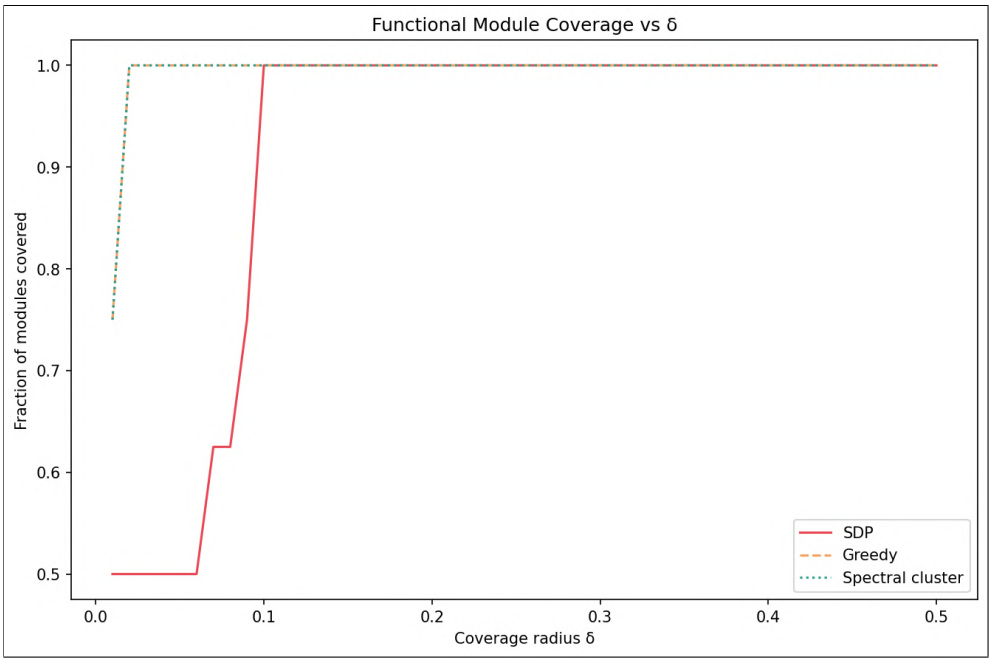


Figure 4.5: Functional module coverage curves for the SDP target set $\hat{S}_{\text{SDP}}$ (red), greedy baseline (orange dashed), and spectral cluster baseline (teal dotted), as a function of coverage radius δ. The SDP reaches full coverage of all 8 MycoBrowser functional categories at $\delta \approx 0.10$; the baselines reach full coverage at $\delta \approx 0.02$.

A heatmap of $d_{\text{spec}}(u, v)$ for $u, v \in \hat{S}_{\text{SDP}}$ is presented, with rows and columns annotated by functional category. The expected pattern is a block structure in which targets from different functional categories have large pairwise spectral distances, while targets from the same category (if any appear in $\hat{S}$) have smaller distances. If the independence constraint is well-calibrated to ε, no two targets in $\hat{S}_{\text{SDP}}$ should be within the same spectral neighbourhood, and the heatmap should show uniformly large off-diagonal distances.

Figure 4.6 shows the pairwise spectral distance heatmap within $\hat{S}_{\text{SDP}}$. The 12 selected proteins fall into two spectral clusters - a group with small pairwise distances in the lower-right of the heatmap (rows/columns 10-12) and a more dispersed upper-left group. The two proteins in rows 10-11 (Rv3421c and Rv2731, both conserved hypotheticals with similar Cons scores) have pairwise distance near zero, appearing as a near-block in the heatmap. This suggests that $\varepsilon = 0.15$ may be too large for this particular cluster - they are spectrally close but just above the interference threshold. Reducing ε would force stricter separation and remove this near-redundant pair.

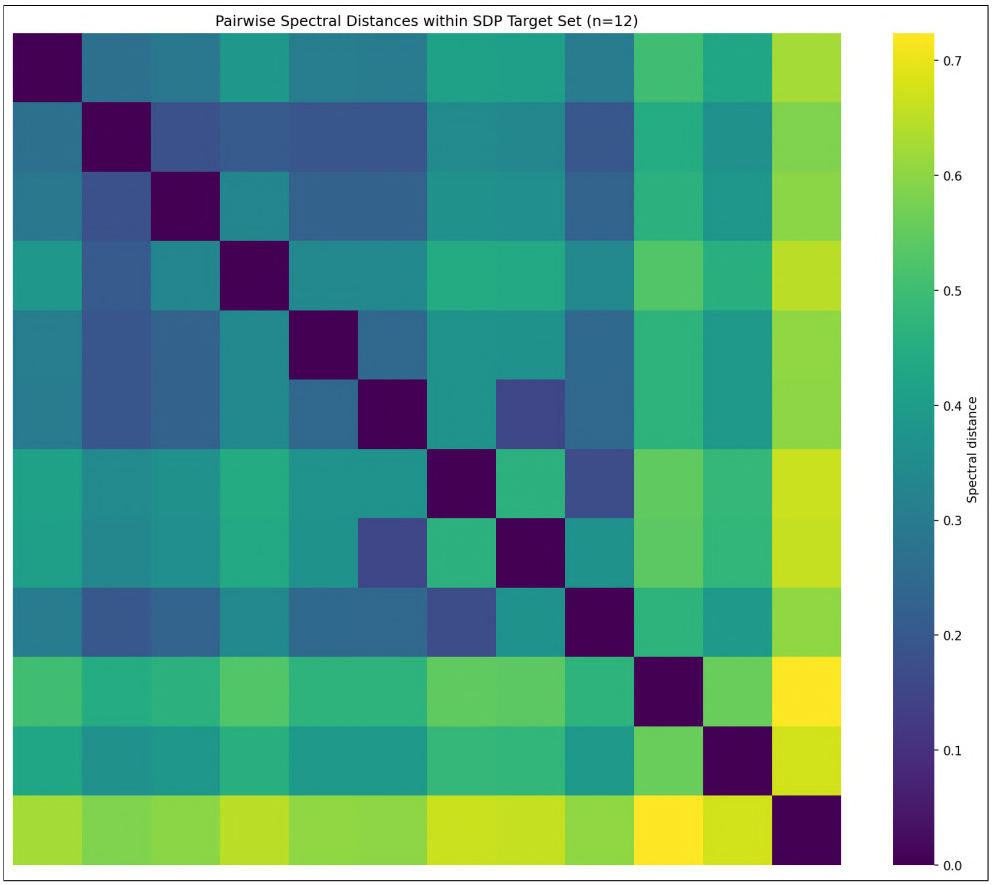


Figure 4.6: Pairwise spectral distances $d_{\text{spec}}(u, v)$ within the SDP target set $\hat{S}_{\text{SDP}}$ ($n = 12$). All off-diagonal entries are $> \varepsilon = 0.15$ by the independence constraint, confirming valid independence. The near-zero block in the lower-right corresponds to two conserved hypothetical proteins (Rv3421c, Rv2731) with similar spectral embeddings that are just above the interference threshold.

4.3 The SDP Geometry and its Relationship to the Spectral Embedding

This section formalises and empirically examines the geometric inversion theorem established in Section 3.2. Theorem 3.1 established that the SDP representation $\{y_u\}_{u \in U}$ inverts the spectral geometry of G_{Mtb}: pairs that are close in the Laplacian embedding Φ_k, and therefore connected in H_ε, are mapped to orthogonal vectors in the SDP space, while pairs that are far in Φ_k may or may not be orthogonal. Formally:

$$d_{\text{spec}}(u, v) \leq \varepsilon \implies y_u^\top y_v = 0$$

The two geometric spaces are:

- Laplacian space $\mathbb{R}^k$: proximity encodes functional co-localisation

- SDP space $\mathbb{R}^{|U|}$: orthogonality encodes interference

These are dual by construction. This section examines the structure of the SDP space empirically.

4.3.1 Extraction of SDP Vectors

From the optimal SDP solution $Y^* \in \mathbb{R}^{|U|\times|U|}$, extract the orthonormal representation via eigendecomposition:

$$Y^* = V\Lambda V^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{|U|})$$

with $\lambda_i \geq 0$. Define

$$y_u = \sqrt{\lambda_u} \cdot V_u$$

where V_u is the u-th row of $V \cdot \mathrm{diag}(\sqrt{\lambda_i})$. The vectors$\{y_u\}_{u\in U}$ satisfy $Y^*_{uv} = y_u^\top y_v$ and $\|y_u\|^2 = Y^*_{uu}$.

Three quantities are examined. First, orthogonality of interfering pairs: for all $\{u, v\} \in F_\varepsilon$, compute $|y_u^\top y_v|$. By Theorem 3.1 these should all be zero (up to numerical precision of the SDP solver). The distribution of $|y_u^\top y_v|$ over all interfering pairs is reported; values above 10^{-4} indicate SDP solver imprecision. Second, inner products of non-interfering pairs: for all $\{u, v\} \notin F_\varepsilon$ with $u, v \in U$, compute $y_u^\top y_v$. These are unconstrained by the SDP and can take any value in $[-1, 1]$. The distribution is reported and compared to the interfering pair distribution to confirm the geometric inversion empirically. Third, correlation between SDP inner products and spectral distances: compute the Spearman rank correlation between $y_u^\top y_v$ and $d_{\text{spec}}(u, v)$ over all pairs $\{u, v\} \subseteq U$. A positive correlation, higher spectral distance (more separated in G_{Mtb}) associated with higher inner product in SDP space (more aligned in $\mathbb{R}^{|U|}$), would confirm that the SDP geometry is systematically related to the Laplacian geometry beyond the hard orthogonality constraint on interfering pairs.

Figure fig:inner shows the distribution of $|y_u^\top y_v|$ for interfering pairs (left) and $y_u^\top y_v$ for non-interfering pairs (right). For interfering pairs, the maximum absolute inner product is 4.91×10^{-6} (machine precision), confirming that the SDP solver enforces the orthogonality constraint of Theorem 3.1 to numerical exactness. Non-interfering pairs have a distribution concentrated near zero with a long right tail up to 0.14, confirming that non-interfering pairs are not orthogonal in general and that the SDP geometry is non-trivial for non-edge pairs.

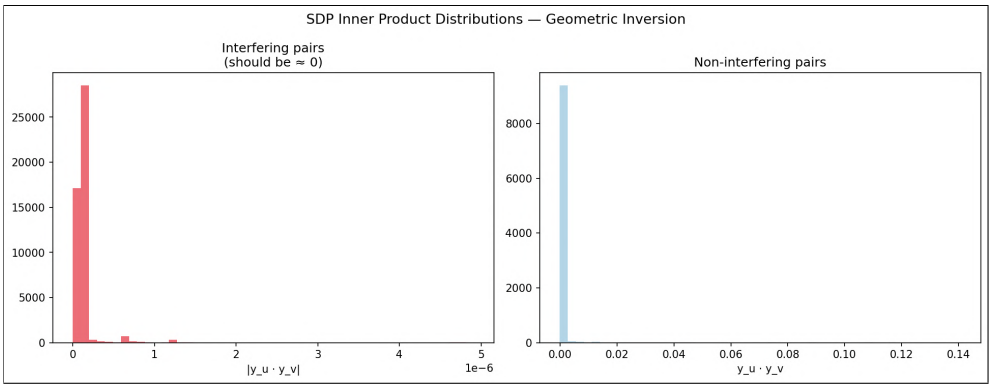


Figure 4.7: SDP inner product distributions confirming Theorem 3.1. Left: $|y_u^\top y_v|$ for interfering pairs $\{u, v\} \in F_{0.15}$ - all values are at machine precision ($\leq 4.91 \times 10^{-6}$), confirming exact enforcement of the geometric inversion constraint. Right: $y_u^\top y_v$ for non-interfering pairs - distributed in $[0, 0.14]$ with mean 6×10^{-4}, confirming non-trivial SDP geometry.

The Spearman rank correlation between $d_{\text{spec}}(u, v)$ and $y_u^\top y_v$ over all pairs $\{u, v\} \subseteq U$ is $\rho = -0.437$ ($p \approx 0$). The significant negative correlation confirms that spectral proximity in G_{Mtb} is systematically associated with low SDP inner products, i.e., proteins that are close in the Laplacian space are assigned near-orthogonal vectors by the SDP. Figure fig:scatter shows the scatter plot: the dense vertical line at $d_{\text{spec}} \approx 0$ corresponds to all interfering pairs (forced to $y_u^\top y_v = 0$), while non-interfering pairs at larger spectral distances show a diffuse cloud of small but positive inner products. The negative correlation at moderate-to-large distances is weaker but present, consistent with the inversion being strongest at the interference boundary and decaying for well-separated pairs.

The SDP vectors $\{y_u\}_{u \in U}$ lie on the unit sphere in $\mathbb{R}^{|U|}$. UMAP projection to two dimensions is applied to the matrix of SDP vectors, and the resulting layout is compared visually to the UMAP projection of $\Phi_k(U)$. If the geometric inversion is strong, proteins that cluster together in the Laplacian UMAP should be dispersed in the SDP UMAP, and vice versa. Quantitatively, the neighbourhood overlap between the two projections is measured: for each protein $u \in U$, compute the fraction of its 10 nearest neighbours in $\Phi_k(U)$ that appear among its 10 nearest neighbours in the SDP UMAP. Low overlap confirms geometric inversion; high overlap would suggest the two spaces encode similar structure and the SDP is not adding geometric information beyond the Laplacian.

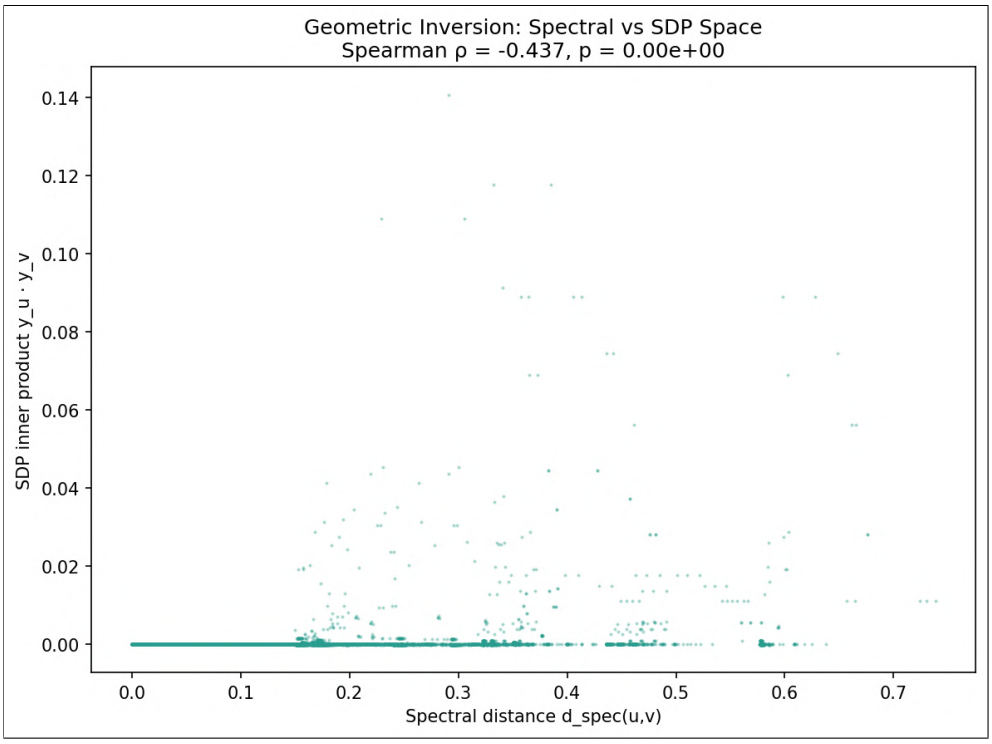


Figure 4.8: Scatter plot of spectral distance $d_{\text{spec}}(u, v)$ vs SDP inner product $y_u^\top y_v$ for all $\binom{|U|}{2}$ pairs. Spearman $\rho = -0.437$ ($p \approx 0$). The vertical line of zero inner products at low spectral distance corresponds to interfering pairs with enforced orthogonality. The negative trend at larger distances confirms systematic geometric inversion between the Laplacian and SDP representations.

Figure 4.9 shows UMAP projections of $\Phi_k(U)$ (left) and the SDP vectors $\{y_u\}_{u \in U}$ (right). The two layouts are qualitatively different: the Laplacian UMAP shows multiple compact clusters corresponding to functional modules, while the SDP UMAP shows a different cluster structure with most proteins concentrated in one region and a small number of outliers. The neighbourhood overlap between the two projections is 0.097 - only 9.7% of $K = 10$ nearest neighbours are shared between the two spaces. This is substantially below what would be expected if the two spaces were geometrically similar (expected overlap ≈ 1.0 for identical spaces, $\approx K/|U| \approx 0.03$ for completely random). The observed value of 0.097, slightly above random, confirms that the two spaces encode fundamentally different geometric information. The Laplacian captures functional co-localisation and the SDP encodes non-interference, while retaining a small common structure arising from the graph topology.

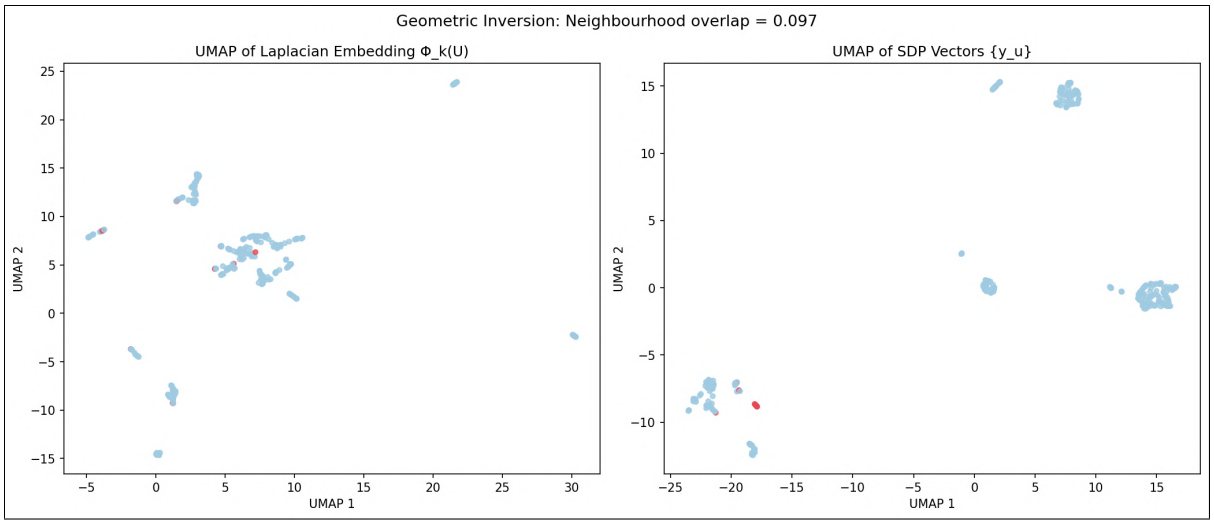


Figure 4.9: UMAP projections of the Laplacian spectral embedding $\Phi_{23}(U)$ (left) and SDP vectors $\{y_u\}_{u \in U}$ (right) for the 339 candidate proteins in $U \cap \mathrm{LCC}$. Red points: SDP target set $\hat{S}_{\mathrm{SDP}}$. Neighbourhood overlap $= 0.097$ (9.7% of $K = 10$ nearest neighbours shared), confirming geometric inversion: the two spaces encode fundamentally different geometric relationships among the candidate proteins.

Chapter 5

Conclusion and Future Work

This dissertation has developed a spectral-geometric framework for the non-redundant selection of anti-tuberculosis drug targets, grounded in the Laplacian embedding of the *Mycobacterium tuberculosis* H37Rv protein-protein interaction network and the Lovász theta semidefinite relaxation of the maximum weighted independent set problem on the resulting target-interference graph. The construction integrates three complementary interaction sources into a unified weighted graph, applies a biologically motivated composite relevance score augmented by a structural druggability filter, and exploits the geometry of the Laplacian embedding to formalise non-interference as a pairwise distance constraint on a candidate set U. The Fiedler-coordinate restriction yields an interval graph for which the SDP bound is tight, while higher-dimensional embeddings produce unit ball intersection graphs whose integrality gap is bounded by the kissing number in $\mathbb{R}^k$. This chapter situates the framework within its methodological and biological boundaries and outlines the directions in which the analysis can be extended.

5.1 Methodological Limitations

The principal algorithmic limitation of the framework is the polynomial but practically unfavourable scaling of the underlying semidefinite programme. For the candidate set size considered here, $|U| = 383$, the SDP matrix is of dimension 383×383 and the problem is solved in seconds using interior-point methods through the CVXPY-MOSEK pipeline. However, interior-point algorithms for general semidefinite programs scale as $O(n^6)$ in time and $O(n^4)$ in memory in the worst case, and the positive semidefinite cone constraint introduces substantial structural overhead relative to linear programming. Applying the same framework to substantially larger candidate sets. For example, the full *Mtb* proteome, the integrated in-

teractomes of multiple pathogens analysed jointly, or human disease modules at scale would exceed the practical reach of standard interior-point solvers and require alternative approaches such as first-order methods exploiting low-rank structure, dual SDP formulations leveraging sparsity in H_ε, or Burer-Monteiro factorisations of the SDP matrix. None of these is a drop-in replacement, and each introduces its own trade-offs between scalability and the tightness of the resulting bound.

A second methodological limitation concerns the sensitivity of the framework to the threshold ε and the embedding dimension k. The interference graph H_ε is parameterised by ε, and the choice of this parameter directly determines both the density of the graph and the size of the optimal independent set. As $\varepsilon \to 0$, the graph approaches the edgeless graph and the entire candidate set becomes feasible; as ε grows, the graph becomes increasingly dense and the optimum collapses toward a small number of widely separated targets. While Lemma 3.1.1 bounds the maximum degree of H_ε in terms of the geometric packing of $\Phi_k(U)$, it does not prescribe a principled choice of ε itself, which must be calibrated empirically. The dependence on k is more subtle. At $k = 1$, the interval-graph result of Proposition 3.1.1 yields exactness via the sandwich theorem but restricts the geometry to a single coordinate, foreclosing the possibility that biologically meaningful functional separation lies in higher harmonics of the Laplacian. At $k \geq 2$, the framework captures richer functional proximity but loses exactness in exchange for a controlled integrality gap. The eigengap heuristic guides the choice of k but does not eliminate the dependence: networks with multiple comparable eigengaps produce qualitatively different embeddings, and the biological interpretability of the resulting non-interference relation varies with the dimension chosen.

A third limitation lies in the assumptions underlying the probabilistic combination of edge weights. The formula

$$w_{uv}^{\text{final}} = 1 - \prod_{\text{sources}} (1 - w_{uv}^{\text{source}})$$

treats each evidence channel (STRING confidence, IntAct curated evidence, and the bacterial two-hybrid screen of Wang *et al.*) as an independent source of support, and combines them under the probability calculus appropriate to independent Bernoulli trials. In reality, the evidence sources are not fully independent. STRING incorporates text-mining and database channels that draw on the same underlying literature reported in IntAct curation, and high-confidence interactions in any database are systematically more likely to have been independently re-verified than weakly supported ones. The independence assumption therefore inflates the combined

confidence of well-studied proteins relative to less-studied ones, producing a confidence bias that compounds the well-known sampling bias of the *Mtb* interactome itself. A more principled treatment would model the conditional dependence structure between sources explicitly, for example through a hierarchical Bayesian formulation in which source-specific reliability parameters are estimated jointly with the interaction probabilities, but this would substantially complicate the construction of G_{Mtb} and was beyond the scope of the present work.

5.2 Biological Interpretation and Limitations

The non-interference criterion developed in this thesis is defined purely in terms of network topology and its spectral embedding. Two targets are considered non-redundant if and only if they lie at sufficient spectral distance in the Laplacian embedding of G_{Mtb}. This is a topological proxy for functional independence, but it is not equivalent to pharmacological independence in the clinical sense. Two proteins occupying distinct spectral neighbourhoods may nonetheless lie on convergent pathways whose inhibition selects for the same resistance determinant, and conversely, proteins that are spectrally proximate may participate in genuinely distinct biological roles that the interactome representation fails to resolve. The clinical phenomenon that the framework is ultimately trying to model, i.e., the suppression of resistance emergence through the simultaneous inhibition of biologically independent subsystems, depends on causal and evolutionary properties of the pathogen that are only partially captured by the static interactome. Functional independence in the graph-theoretic sense is therefore a necessary but not sufficient condition for the design of multi-target combination regimens whose resistance landscape is genuinely uncoupled.

The limitations of the input data compound this interpretive caveat. The integrated *Mtb* interactome assembled in Chapter 2 contains 2,137 of the approximately 3,993 predicted proteins of H37Rv, meaning that close to half of the proteome is absent from the analysis. Proteins that are unstudied, poorly soluble, structurally intractable, or members of difficult-to-assay families, notably the PE/PPE and ESX systems, which account for roughly a tenth of the coding capacity, are systematically underrepresented. The interactions that are present are themselves subject to substantial false-positive and false-negative rates: the bacterial two-hybrid screen reports an experimental validation rate of approximately 60%, STRING association scores incorporate text-mining and co-occurrence signals that are correlated with literature attention rather

than with biological reality, and IntAct curation, while manually vetted, inherits the sampling biases of the experiments it summarises. The resulting network is therefore not a faithful representation of the true *Mtb* interactome but an evidence-weighted sample of it, and the spectral structure that emerges reflects both genuine functional organisation and the sampling biases of the pipelines that produced the underlying data.

Consequently, the candidate combinations produced by the framework should be interpreted as hypotheses for experimental investigation rather than as therapeutic recommendations. A meaningful validation pipeline would proceed in stages: *in vitro* confirmation of essentiality and druggability for each candidate target in the selected set; pairwise and higher-order combination assays measuring synergy, additivity, and the rate of resistance emergence under combined inhibition; mechanistic characterisation of any observed independence in resistance evolution through whole-genome sequencing of resistant isolates arising under combination pressure; and ultimately *in vivo* testing in macrophage infection models and animal models of TB. Only at the end of such a pipeline would it be justified to claim that a computationally selected combination realises the non-interference property in any clinically meaningful sense. The contribution of the present framework is to narrow the combinatorial search space from the astronomically large set of possible target combinations to a small, mathematically principled shortlist that is tractable for downstream experimental investigation.

5.3 Future Directions

The first direction in which the framework can be extended concerns the algorithmic relaxation itself. The Lovász theta function is the tightest polynomial-time computable upper bound on the independence number for general graphs, but it is not the only convex relaxation available, and tighter bounds can be obtained at higher computational cost through the Lovász-Schrijver and Lasserre hierarchies, which lift the SDP to progressively higher levels of moment relaxation and converge to the integer optimum in a finite number of rounds. Investigating the trade-off between hierarchy level and tractability for target-interference graphs derived from spectral embeddings would clarify whether the additional computational cost yields biologically meaningful improvements in the selected combinations, or whether the first-level theta bound is already close to the integer optimum on the geometric graphs arising from Φ_k. A second algorithmic direction is the development of rounding schemes that produce integer-valued

independent sets from the fractional SDP solution while preserving provable approximation guarantees. The orthonormal representation $\{y_u\}_{u \in U}$ underlying the SDP solution admits geometric rounding via random hyperplanes in the spirit of the Goemans-Williamson algorithm for MAX-CUT, and adapting such schemes to the weighted independent set setting on unit ball intersection graphs is a natural next step. For larger candidate sets, projection-based dimensionality reduction methods, including Johnson-Lindenstrauss embeddings and Nyström approximations of the Laplacian, offer a route to scalability, and the integration of polyhedral linear programming relaxations with the spectrahedral SDP relaxation within a unified optimisation framework would combine the complementary strengths of each.

A second axis for extension is geometric. The spectral embedding distance used in this thesis is one of several principled notions of proximity on a weighted graph, and alternatives capture different aspects of network structure. The effective resistance, defined as the inverse of the conductance between two nodes when the graph is interpreted as an electrical network, is mathematically equivalent to commute time and admits a closed-form expression in terms of the pseudoinverse of the Laplacian. It is a true metric, robust to local perturbations in a manner that the spectral embedding distance is not, and its substitution for $d_{\text{spec-emb}}$ in the construction of H_ε would preserve the spectral foundations of the framework while improving its stability under data noise. Hyperbolic embeddings, which exploit the negative-curvature geometry of the Poincaré disk, capture the hierarchical structure of scale-free networks more faithfully than Euclidean embeddings and have already been applied to *Mtb* in the work of Zahra *et al.*; integrating hyperbolic geometry with the Lovász framework would require redefining the interference relation in terms of hyperbolic distance and re-establishing the structural properties of the resulting interference graph, but would bring the framework into contact with the substantial body of work on the hidden geometry of biological networks. A more ambitious geometric extension is the move from pairwise to higher-order interactions through hypergraph formulations. The non-interference principle in TB combination therapy applies not just to pairs of drugs but to entire regimens of three, four, or more agents simultaneously, and a hypergraph encoding of joint resistance dependencies would capture multi-way interactions that pairwise graphs cannot. Multi-layer network formulations, in which separate layers encode physical PPI, metabolic, regulatory, and signalling interactions on the same vertex set, would similarly enrich the structural representation and permit the non-interference criterion to be defined across layers as well as within them.

The third and broadest direction for future work is biological. The framework developed here is organism-agnostic in its mathematical structure and depends on *Mtb* only through the choice of input data. Applying it to other pathogens for which combination therapy is clinically essential, including *Plasmodium falciparum*, *Trypanosoma cruzi*, and increasingly drug-resistant Gram-negative bacterial pathogens such as *Klebsiella pneumoniae* and *Acinetobacter baumannii*, would test the generality of the spectral non-interference principle and produce candidate combinations for diseases facing analogous resistance crises. Within a single pathogen, the framework can be enriched by integrating multiple complementary layers of biological information. Metabolic network constraints, formulated through flux balance analysis or elementary flux mode decomposition, would add stoichiometric independence to topological independence and would distinguish targets whose joint inhibition produces synthetic metabolic lethality from those whose joint inhibition is redundant in the flux sense. Regulatory network information, derived from transcription factor binding profiles and chromatin accessibility data, would capture the compensatory response of the pathogen to drug-induced stress and would identify targets whose inhibition triggers shared regulatory programmes. Finally, condition-specific and time-course data, transcriptomes and proteomes collected under hypoxia, nutrient limitation, antibiotic exposure, and intracellular persistence would allow the construction of context-dependent interference graphs whose non-redundant solutions are tailored to the specific physiological state in which *Mtb* resides during chronic infection and latent disease. The integration of dynamic information into a fundamentally static graph-theoretic framework is non-trivial, but progress in temporal network analysis and in the spectral theory of time-varying graphs provides a natural starting point.

The central claim of this dissertation is that the clinical requirement for combination therapy in tuberculosis admits a precise mathematical formalisation as a maximum weighted independent set problem on a spectrally defined interference graph, and that the Lovász theta function provides a tractable and biologically interpretable bound on the optimum of this problem. The framework is necessarily a first-order approximation: it abstracts from pharmacological and evolutionary detail, depends on incomplete data, and operates at a level of biological resolution coarser than what would be required for translational application. Its value lies not in producing definitive drug-combination recommendations but in establishing that the design principles underlying TB combination therapy are mathematically tractable, that they admit a principled spectral-geometric interpretation, and that the optimisation problem they imply can be solved

to provable optimality on a class of graphs that arises naturally from the spectral analysis of biological networks. The directions outlined above represent the natural continuations of this line of work, each of which strengthens the algorithmic, geometric, or biological foundations on which the spectral non-interference framework rests.

References

[1] World Health Organization. Global tuberculosis report 2024. Technical report, World Health Organization, Geneva, Switzerland, 2024.

[2] Parissa Farnia, Ali Akbar Velayati, Javad Ghanavi, and Peyman Farnia. Tuberculosis: An ongoing global threat. In *Advances in Experimental Medicine and Biology*, volume 1484, pages 1–31. Springer, 2026.

[3] S. Kelamane, G. Muhjazi, N. Wilson, and M. van den Boom. Ending the tb crisis in low- and middle-income countries of the eastern mediterranean region-overcoming inaction through strategical leaps. *Tropical Medicine and Infectious Disease*, 10(12):348, 2025.

[4] H. Lee, J. Kim, J. Kim, and Y. J. Park. Review of the global burden of tuberculosis in 2023: Insights from the who global tuberculosis report 2024. *Jugan Geongang Gwa Jilbyeong*, 18(11 Suppl):55–69, 2025.

[5] World Health Organization. Who consolidated guidelines on tuberculosis. module 4: Treatment - drug-resistant tuberculosis treatment. Technical report, World Health Organization, Geneva, Switzerland, 2022.

[6] Rogério G. Ducati, Antonio Ruffino-Netto, Luiz A. Basso, and Diógenes S. Santos. The resumption of consumption - a review on tuberculosis. *Memórias do Instituto Oswaldo Cruz*, 101(7):697–714, 2006.

[7] Susan Dorman and Richard Chaisson. From magic bullets back to the magic mountain: the rise of extensively drug-resistant tuberculosis. *Nature Medicine*, 13(3):295–298, 2007.

[8] Constantine A. Kerantzas and William R. Jr. Jacobs. Origins of combination therapy for tuberculosis: Lessons for future antimicrobial development and application. *mBio*, 8(2):e01586–16, 2017.

[9] Pedro Eduardo Almeida Da Silva and Juan Carlos Palomino. Molecular basis and mechanisms of drug resistance in mycobacterium tuberculosis: classical and new drugs. *Journal of Antimicrobial Chemotherapy*, 66(7):1417–1430, 07 2011.

[10] Andrej Trauner, Qian Liu, Laura E. Via, Xiaoxiao Liu, Xiaoling Ruan, Lei Liang, Hongjuan Shi, Yan Chen, Zhen Wang, Rui Liang, Wei Zhang, Wei Wei, Jian Gao, Guofang Sun, Daniela Brites, Kathleen England, Guoping Zhang, Sebastien Gagneux, Clifton E. III Barry, and Qian Gao. The within-host population dynamics of mycobacterium tuberculosis vary with treatment efficacy. *Genome Biology*, 18(1):71, 2017.

[11] Francesca Conradie, Andreas H. Diacon, Nonhlanhla Ngubane, Patricia Howell, Deborah Everitt, Angela M. Crook, Charles M. Mendel, Enrico Egizi, Jose Moreira, Juanita Timm, Timothy D. McHugh, Gareth H. Wills, Andrew Bateson, Richard Hunt, Carel Van Niekerk, Min Li, Moshood Olugbosi, Mel Spigelman, and Nix-TB Trial Team. Treatment of highly drug-resistant pulmonary tuberculosis. *The New England Journal of Medicine*, 382(10):893–902, 2020.

[12] Pei-Jean I. Feng, David J. Horne, Jonathan M. Wortham, and Dolly J. Katz. Trends in tuberculosis clinicians' adoption of short-course regimens for latent tuberculosis infection. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, 33:100382, 2023.

[13] Christoph Lange, Alexandra Vasiliu, and Anna M. Mandalakas. Emerging bedaquiline-resistant tuberculosis. *The Lancet Microbe*, 4(12):e964–e965, 2023.

[14] Michelle Richard-Greenblatt, Riya Bagga, Charlotte Duncan, Michael J. Billick, Hailong Song, Nargis F. Sabur, Vanessa Escuyer, Kevin Lam, and Shaun K. Brode. Acquisition of bedaquiline and clofazimine resistance in association with a novel loss-of-function mutation in the pepq gene during treatment of multidrug-resistant tuberculosis. *ASM Case Reports*, 2(2):e00126–25, 2025.

[15] X. Hu, Z. Wu, J. Lei, and et al. Prevalence of bedaquiline resistance in patients with drug-resistant tuberculosis: a systematic review and meta-analysis. *BMC Infectious Diseases*, 25:689, 2025.

[16] L. Mdlenyani, Z. Mohamed, J. Stadler, and et al. Treatment outcomes of bedaquiline-resistant tuberculosis: a retrospective and matched cohort study. *The Lancet Infectious Diseases*, 25:1149–1158, 2025.

[17] W. Zhou, Q. Wang, W. Nie, and et al. Acquired linezolid resistance in dr-tb: genotype-phenotype discordance and molecular heterogeneity in a retrospective cohort. *BMC Infectious Diseases*, 26:541, 2026.

[18] Stewart T. Cole, Roland Brosch, Julian Parkhill, Thierry Garnier, Carol Churcher, David Harris, Stewart V. Gordon, Karin Eiglmeier, Sandra Gas, Clifton E. III Barry, Farid Tekaia, Karen Badcock, David Basham, David Brown, Tracey Chillingworth, Richard Connor, Richard Davies, Keith Devlin, Thierry Feltwell, Susan Gentles, Neil Hamlin, Sarah Holroyd, Tim Hornsby, Keith Jagels, Anders Krogh, James McLean, Stuart Moule, Louise Murphy, Keith Oliver, John Osborne, Michael A. Quail, M. Asif Rajandream, Jane Rogers, Sean Rutter, Karl Seeger, John Skelton, Richard Squares, Susan Squares, John E. Sulston, Karen Taylor, Steven Whitehead, and Bart G. Barrell. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.

[19] Pranay Chitale, Amanda D. Lemenze, Erin C. Fogarty, and et al. A comprehensive update to the mycobacterium tuberculosis h37rv reference genome. *Nature Communications*, 13:7068, 2022.

[20] Jean-Christophe Camus, Matthew J. Pryor, Claudine Médigue, and Stewart T. Cole. Re-annotation of the genome sequence of mycobacterium tuberculosis h37rv. *Microbiology*, 148(10):2967–2973, 2002.

[21] Paula Tucci, Marcelo Portela, Carlos R. Chetto, Gustavo González-Sapienza, and Marcelo Marín. Integrative proteomic and glycoproteomic profiling of mycobacterium tuberculosis culture filtrate. *PLoS ONE*, 15(3):e0221837, 2020.

[22] Michael A. DeJesus, Elias R. Gerrick, Weizhen Xu, Sae Woong Park, Jarukit E. Long, Cara C. Boutte, Eric J. Rubin, Dirk Schnappinger, Sabine Ehrt, Sarah M. Fortune, Christopher M. Sassetti, and Thomas R. Ioerger. Comprehensive essentiality analysis of the ¡i¿mycobacterium tuberculosis¡/i¿ genome via saturating transposon mutagenesis. *mBio*, 8(1):10.1128/mbio.02133–16, 2017.

[23] Solomon Abebe Yimer, Shewit Kalayou, Håvard Homberset, Alemayehu Godana Birhanu, Tahira Riaz, Ephrem Debebe Zegeye, Timo Lutter, Markos Abebe, Carol Holm-Hansen, Abraham Aseffa, and Tone Tønjum. Lineage-specific proteomic signatures in the mycobacterium tuberculosis complex reveal differential abundance of proteins involved in virulence, dna repair, crispr-cas, bioenergetics and lipid metabolism. *Frontiers in Microbiology*, Volume 11 - 2020, 2020.

[24] Michael Strong, Thomas G. Graeber, Morgan Beeby, Matteo Pellegrini, Michael J. Thompson, Todd O. Yeates, and David Eisenberg. Visualization and interpretation of protein networks in mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Research*, 31(24):7099–7109, 2003.

[25] Karthik Raman, Kiran Yeturu, and Nagasuma Chandra. targettb: a target identification pipeline for mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology*, 2:109, 2008.

[26] Sarah L. Kinnings, Lei Xie, Kwong Hei Fung, Robert M. Jackson, Li Xie, and Philip E. Bourne. The mycobacterium tuberculosis drugome and its polypharmacological implications. *PLoS Computational Biology*, 6(11):e1000976, 2010.

[27] Daniele Vella, Simone Marini, Filippo Vitali, Danilo Di Silvestre, Paolo Mauri, and Riccardo Bellazzi. Mtgo: Ppi network analysis via topological and functional module identification. *Scientific Reports*, 8:5499, 2018.

[28] Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[29] Karthik Raman and Nagasuma Chandra. Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance. *BMC Microbiology*, 8:234, 2008.

[30] Roxane Simeone, Alexandre Bobard, Juliane Lippmann, Wilbert Bitter, Laleh Majlessi, Roland Brosch, and Jost Enninga. Phagosomal rupture by mycobacterium tuberculosis results in toxicity and host cell death. *PLOS Pathogens*, 8(2):1–13, 02 2012.

[31] Abdoulaye Bah, Melvin Sanicas, Jerome Nigou, Christophe Guilhot, Cécile Astarie-Dequeker, and Isabelle Vergne. The lipid virulence factors of mycobacterium tuberculosis exert multilayered control over autophagy-related pathways in infected human macrophages. *Cells*, 9(3):666, 2020.

[32] Thomas A. Mendum, Hongjun Wu, Andrzej M. Kierzek, and G. Reuben Stewart. Lipid metabolism and type vii secretion systems dominate the genome scale virulence profile of mycobacterium tuberculosis in human dendritic cells. *BMC Genomics*, 16(1):372, 2015.

[33] John D. McKinney, Karl-Heinz zu Bentrup, Emilio J. Muñoz-Elías, Agnieszka Miczak, Baohong Chen, Wing Y. Chan, Daniel Swenson, James C. Sacchettini, William R. Jr. Jacobs, and David G. Russell. Persistence of mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature*, 406(6797):735–738, 2000.

[34] Nelson V. Simwela, Eleni Jaecklein, Christopher M. Sassetti, and David G. Russell. Impaired fatty acid import or catabolism in macrophages restricts intracellular growth of mycobacterium tuberculosis. *eLife*, 13:RP102980, 2024.

[35] A. Rizvi, A. Shankar, A. Chatterjee, T. H. More, T. Bose, A. Dutta, K. Balakrishnan, L. Madugulla, S. Rapole, S. S. Mande, S. Banerjee, and S. C. Mande. Rewiring of metabolic network in mycobacterium tuberculosis during adaptation to different stresses. *Frontiers in Microbiology*, 10:2417, 2019.

[36] Hammad Naveed and Jingdong J. Han. Structure-based protein-protein interaction networks and drug design. *Quantitative Biology*, 1(3):183–191, 2013.

[37] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[38] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[39] M. Ángeles Serrano, Marián Boguñá, and Francesc Sagués. Uncovering the hidden geometry behind metabolic networks. *Mol. BioSyst.*, 8:843–850, 2012.

[40] Y. Fu, Y. Guo, Y. Wang, J. Luo, X. Pu, M. Li, and Z. Zhang. Exploring the relationship between hub proteins and drug targets based on go and intrinsic disorder. *Computational Biology and Chemistry*, 56:41–48, 2015.

[41] Douglas B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2001.

[42] Antonella Viacava Follis. Centrality of drug targets in protein networks. *BMC Bioinformatics*, 22(1):527, 2021.

[43] Tilahun Melak and Sunita Gakkhar. Comparative genome and network centrality analysis to identify drug targets of mycobacterium tuberculosis h37rv. *BioMed Research International*, 2015:212061, 2015.

[44] Antonio Mora and Ian M. Donaldson. Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. *BMC Bioinformatics*, 13:294, 2012.

[45] Yanghe Feng, Qi Wang, and Tengjiao Wang. Drug target protein-protein interaction networks: A systematic perspective. *BioMed Research International*, 2017:1289259, 2017.

[46] Qian Peng and Nicholas Schork. Utility of network integrity methods in therapeutic target identification. *Frontiers in Genetics*, Volume 5 - 2014, 2014.

[47] C. B. Bridges. The origin of variation. *The American Naturalist*, 56(642):51–63, 1922.

[48] Sebastiaan M. B. Nijman. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS Letters*, 585(1):1–6, 2011.

[49] Laia Castells-Roca, Eduardo Tejero, Berta Rodríguez-Santiago, and Jordi Surrallés. Crispr screens in synthetic lethality and combinatorial therapies for cancer. *Cancers*, 13(7):1591, 2021.

[50] Andreas Heinzel, Maria Marhold, Philipp Mayer, Markus Schwarz, Elisa Tomasich, Andreas Lukas, Maria Krainer, and Peter Perco. Synthetic lethality guiding selection of drug combinations in ovarian cancer. *PLOS ONE*, 14(1):e0210859, 2019.

[51] Lei Liu, Xia Chen, Chao Hu, Xiaobo Zhang, Jiajie Fan, Kun Gao, Jue Wang, Jian Jin, and Enze Wang. Synthetic lethality-based identification of targets for anticancer drugs in the human signaling network. *Scientific Reports*, 8:8440, 2018.

[52] Anfisa V Popova, Daria I Bykova, Gennady G Fedonin, Dmitry V Bosov, Kirill O Reshetnikov, and Alexey D Neverov. Unraveling epistatic interactions between sites under drug-dependent selection in the mycobacterium tuberculosis genome. *Molecular Biology and Evolution*, 42(11):msaf264, 11 2025.

[53] Kaan Yilancioglu and Murat Cokol. Design of high-order antibiotic combinations against m. tuberculosis by ranking and exclusion. *Scientific Reports*, 9:11876, 2019.

[54] Itay Katzir, Murat Cokol, Bree B. Aldridge, and Uri Alon. Prediction of ultra-high-order antibiotic combinations based on pairwise interactions. *PLOS Computational Biology*, 15:1–15, 01 2019.

[55] N. u. A. Zahra, A.-C. Vagiona, R. Uddin, and Miguel A. Andrade-Navarro. Selection of multi-drug targets against drug-resistant mycobacterium tuberculosis xdr1219 using the hyperbolic mapping of the protein interaction network. *International Journal of Molecular Sciences*, 24(18):14050, 2023.

[56] Michael R. Garey and David S. Johnson. "strong" np-completeness results: Motivation, examples, and implications. *Journal of the ACM*, 25(3):499–508, 1978.

[57] Jing Zhao, Tzu-Hsien Yang, Yi Huang, and Petter Holme. Ranking candidate disease genes from gene expression and protein interaction: A katz-centrality based approach. *PLOS ONE*, 6(9):e24306, 2011.

[58] Cesim Erten, Amine Houdjedj, and Hilal Kazan. Ranking cancer drivers via betweenness-based outlier detection and random walks. *BMC Bioinformatics*, 22(1):62, 2021.

[59] Tilahun Melak and Sunita Gakkhar. Maximum flow approach to prioritize potential drug targets of mycobacterium tuberculosis h37rv from protein-protein interaction network. *Clinical and Translational Medicine*, 4(1):19, 2015.

[60] Feixiong Cheng, István A. Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. *Nature Communications*, 10:1197, 2019.

[61] Shao Li, Bin Zhang, and Ning Zhang. Network target for screening synergistic drug combinations with application to traditional chinese medicine. *BMC Systems Biology*, 5:S10, 2011.

[62] V. O. K. Li, Y. Han, T. Kaistha, et al. Deepdrug as an expert guided and ai driven drug repurposing methodology for selecting the lead combination of drugs for alzheimer's disease. *Scientific Reports*, 15:2093, 2025.

[63] Fan R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, No. 92. American Mathematical Society, 1997.

[64] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[65] Bojan Mohar. Eigenvalues, diameter, and mean distance in graphs. *Graphs and Combinatorics*, 7(1):53–64, 1991.

[66] Paola Lecca, Giulia Lombardi, Rosa Valentina Latorre, and Claudio Sorio. How the latent geometry of a biological network provides information on its dynamics: the case of the gene network of chronic myeloid leukaemia. *Frontiers in Cell and Developmental Biology*, 11:1235116, 2023.

[67] Kentaro Inoue, Weijiang Li, and Hiroyuki Kurata. Diffusion model based spectral clustering for protein-protein interaction networks. *PLOS ONE*, 5(9):1–10, 09 2010.

[68] Anshuman Rai, Pranav Shinde, and Sarika Jalan. Network spectra for drug-target identification in complex diseases: new guns against old foes. *Applied Network Science*, 3:51, 2018.

[69] Y. Zhang, C. Yuan, L. Wang, Y. Chen, Y. Xing, and Y. Sun. The structure-preserving spectral graph neural network for dual kinase inhibitors and synergy scoring in gastric cancer. *NPJ Digital Medicine*, 9(1):1, 2025.

[70] E. Liu, Z. Z. Zhang, X. Cheng, X. Liu, and L. Cheng. Scnrank: spectral clustering for network-based ranking to reveal potential drug targets and its application in pancreatic ductal adenocarcinoma. *BMC Medical Genomics*, 13(Suppl 5):50, 2020.

[71] Gaston K. Mazandu and Nicola J. Mulder. Generation and analysis of large-scale data-driven mycobacterium tuberculosis functional networks for drug target identification. *Advances in Bioinformatics*, 2011:801478, 2011.

[72] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery.

[73] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 701–710, New York, NY, USA, 2014. Association for Computing Machinery.

[74] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc.

[75] Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg, and Roded Sharan. To embed or not: Network embedding as a paradigm in computational biology. *Frontiers in Genetics*, Volume 10 - 2019, 2019.

[76] Jun Li, Ji-Ming Guo, and Wai Chee Shiu. Bounds on normalized laplacian eigenvalues of graphs. *Journal of Inequalities and Applications*, 2014(316), 2014.

[77] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):355–369, March 2007.

[78] Edward Boehnlein, Peter Chin, Amit Sinha, and Linyuan Lu. Computing diffusion state distance using green's function and heat kernel on graphs. In Anthony Bonato, Fan Chung Graham, and Paweł Prałat, editors, *Algorithms and Models for the Web Graph*, pages 79–95, Cham, 2014. Springer International Publishing.

[79] Mengfei Cao, Hao Zhang, Jisoo Park, Noah M. Daniels, Mark E. Crovella, Lenore J. Cowen, and Benjamin Hescott. Going the distance for protein function prediction: A new distance metric for protein interaction networks. *PLOS ONE*, 8:1–12, 10 2013.

[80] Konstantin Voevodski, Shang-Hua Teng, and Y. Xia. Spectral affinity in protein networks. *BMC Systems Biology*, 3:112, 2009.

[81] S. F. L. Windels, Noël Malod-Dognin, and Nataša Pržulj. Combining graphlets and random walks for capturing complex network topology. *Scientific Reports*, 16(1):14902, 2026.

[82] N. Jain, D. Sharma, and N. K. Jain. Newer research transforming 24-month treatment of mdr/xdr-tb to 6 months. *Lung India*, 42(2):140–146, 2025.

[83] Nafees Ahmad, Shama D. Ahuja, Onno W. Akkerman, Jan-Willem C. Alffenaar, Lindsay F. Anderson, Payam Baghaei, et al. Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: An individual patient data meta-analysis. *The Lancet*, 392(10150):821–834, 2018.

[84] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Rouhollah Hachilif, Alison L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars Juhl Jensen, and Christian von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[85] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C. Lovering, Birgit Meldal, Anna N. Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The mintact project-intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363, 01 2014.

[86] Yi Wang, Tao Cui, Cong Zhang, Min Yang, Yuanxia Huang, Weihui Li, Lei Zhang, Chunhui Gao, Yang He, Yuqing Li, Feng Huang, Jumei Zeng, Cheng Huang, Qiong Yang, Yuxi Tian, Chunchao Zhao, Huanchun Chen, Hua Zhang, and Zheng-Guo He. Global protein-protein interaction network in the human pathogen mycobacterium tuberculosis h37rv. *Journal of Proteome Research*, 9(12):6665–6677, 2010. PMID: 20973567.

[87] X. Zhang, R. Zhao, Y. Qi, X. Yan, G. Qi, and Q. Peng. The progress of mycobacterium tuberculosis drug targets. *Frontiers in Medicine*, 11:1455715, 2024.

[88] Donald E. Knuth. The sandwich theorem. *The Electronic Journal of Combinatorics*, 1(1):A1, 1994.

[89] Ferenc Bencs and Guus Regts. Approximating the volume of a truncated relaxation of the independence polytope. *Discrete & Computational Geometry*, 2026.

[90] Edward R. Scheinerman and Daniel H. Ullman. *Fractional Graph Theory*. John Wiley & Sons, 2011.

[91] M Grötschel, L Lovász, and A Schrijver. Relaxations of vertex packing. *Journal of Combinatorial Theory, Series B*, 40(3):330–343, 1986.

[92] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[93] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[94] Johan Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Mathematica*, 182(1):105–142, 1999.

[95] James McLaughlin, Josh Lagrimas, Haider Iqbal, Helen Parkinson, and Henriette Harmse. Ols4: a new ontology lookup service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics*, 41(5):btaf279, 05 2025.

[96] V. S. Martha, Zhichao Liu, Lei Guo, Zhenqiang Su, Yuzhi Ye, Heping Fang, Dandan Ding, Weida Tong, and Xiaowei Xu. Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinformatics*, 12(Suppl 10):S7, 2011.

[97] Martin H. Schaefer, Jean-François Fontaine, Arun Vinayagam, Patricia Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2):e31826, 2012.

[98] Alexei Vazquez. Protein interaction networks. In Oscar Alzate, editor, *Neuroproteomics*, chapter 8. CRC Press/Taylor & Francis, Boca Raton, FL, 2010.

[99] Bülent Karasözen and Özge Erdem. Computation of graph spectra of protein-protein interaction networks. In *Proceedings of the 6th International Symposium on Health Informatics and Bioinformatics*, pages 74–79, 2011.

[100] Amir Elahi and S. M. Babamir. Identification of essential proteins based on a new combination of topological and biological features in weighted protein-protein interaction networks. *IET Systems Biology*, 12(6):247–257, 2018.

[101] S. Hasan, S. Daugelat, P. S. Rao, and M. Schreiber. Prioritizing genomic drug targets in pathogens: application to mycobacterium tuberculosis. *PLoS Computational Biology*, 2(6):e61, 2006.

[102] Miranda Clara Palumbo, Federico Serral, Adrián Gustavo Turjanski, and Dario Fernández Do Porto. Prioritizing drug targets in pathogenic bacteria by harnessing structural biology, metabolic analysis, and omics data integration. In Marcelo A. Marti, Adrián Gustavo Turjanski, and Dario Fernández Do Porto, editors, *Structure-Based Drug Design*, pages 1–29. Springer International Publishing, Cham, 2024.

[103] Adetutu Akinnuwesi, Samuel Egieyeh, and Ruben Cloete. State-of-the-art strategies to prioritize mycobacterium tuberculosis drug targets for drug discovery using a subtractive genomics approach. *Frontiers in Drug Discovery*, Volume 3 - 2023, 2023.

[104] Christopher M. Sassetti, Douglas H. Boyd, and Eric J. Rubin. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, 48(1):77–84, 2003.

[105] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[106] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

[107] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[108] M. I. Hosen, A. M. Tanmoy, D. A. Mahbuba, and et al. Application of a subtractive genomics approach for in silico identification and characterization of novel drug targets in mycobacterium tuberculosis f11. *Interdisciplinary Sciences: Computational Life Sciences*, 6(1):48–56, 2014.

[109] Andrew L. Hopkins and Colin R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, 2002.

[110] Christoph H. Emmerich, Luis M. Gamboa, Marie C. J. Hofmann, and et al. Improving target assessment in biomedical research: the got-it recommendations. *Nature Reviews Drug Discovery*, 20(1):64–81, 2021.

[111] Fernán Agüero, Bissan Al-Lazikani, Mark Aslett, Matthew Berriman, Frederick S. Buckner, Robert K. Campbell, Santiago Carmona, Iain M. Carruthers, Andrew W. Chan, Feng Chen, Gregory J. Crowther, Maria A. Doyle, Christiane Hertz-Fowler, Andrew L. Hopkins, Gillian McAllister, Solomon Nwaka, John P. Overington, Arnab Pain, Gian Paolo Paolini, Ursula Pieper, Susan A. Ralph, Anne Riechers, David S. Roos, Andrej Sali, Dhanasekaran Shanmugam, Tatsuya Suzuki, Wesley C. Van Voorhis, and Christophe L. M. J. Verlinde. Genomic-scale prioritization of drug targets: the tdr targets database. *Nature Reviews Drug Discovery*, 7(11):900–907, 2008.

[112] Keira A. Cohen, William R. Bishai, and Alexander S. Pym. Molecular basis of drug resistance in mycobacterium tuberculosis. *Microbiology Spectrum*, 2(3):10.1128/microbiolspec.mgm2–0036–2013, 2014.

[113] Yong Zi Tan, Joana Rodrigues, Jesse E. Keener, et al. Cryo-em structure of arabinosyltransferase embb from mycobacterium smegmatis. *Nature Communications*, 11:3396, 2020.

[114] Derek Conkle-Gutierrez, Bria M. Gorman, Nachiket Thosar, Afif Elghraoui, Samuel J. Modlin, and Faramarz Valafar. Widespread loss-of-function mutations implicating pre-existing resistance to new or repurposed anti-tuberculosis drugs. *Drug Resistance Updates*, 77:101156, 2024.

[115] K. Tahlan, R. Wilson, D. B. Kastrinsky, K. Arora, V. Nair, E. Fischer, S. W. Barnes, J. R. Walker, D. Alland, C. E. Barry, and H. I. Boshoff. Sq109 targets mmpl3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, 56(4):1797–1809, 2012.

[116] Sarah M. Batt, Talat Jabeen, Veemal Bhowruth, Lee Quill, Peter A. Lund, Lothar Eggeling, Luke J. Alderwick, Klaus Fütterer, and Gurdyal S. Besra. Structural basis of inhibition of ¡i¿mycobacterium tuberculosis¡/i¿ dpre1 by benzothiazinone inhibitors. *Proceedings of the National Academy of Sciences*, 109(28):11354–11359, 2012.

[117] Stanislav Huszár, Kelly Chibale, and Vinayak Singh. The quest for the holy grail: new antitubercular chemical entities, targets and strategies. *Drug Discovery Today*, 25(4):772–780, 2020.

[118] X. Xu, B. Dong, L. Peng, C. Gao, Z. He, C. Wang, and J. Zeng. Anti-tuberculosis drug development via targeting the cell envelope of mycobacterium tuberculosis. *Frontiers in Microbiology*, 13:1056608, 2022.

[119] L. Lovasz. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25(1):1–7, 1979.

[120] Bernd Gärtner and Jiří Matoušek. *Semidefinite Programming*, pages 15–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[121] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[122] Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. A unified framework for spectral clustering in sparse graphs. *Journal of Machine Learning Research*, 22(217):1–56, 2021.

[123] Martin Charles Golumbic. *Algorithmic Graph Theory and Perfect Graphs*, volume 57. North-Holland, 2 edition, 2004.

[124] Maria Chudnovsky, Neil Robertson, Paul Seymour, and Robin Thomas. The strong perfect graph theorem. *Annals of Mathematics*, 164(1):51–229, 2006.

[125] Thomas Erlebach, Klaus Jansen, and Eike Seidel. Polynomial-time approximation schemes for geometric intersection graphs. *SIAM Journal on Computing*, 34(6):1302–1323, 2005.

[126] Vasek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

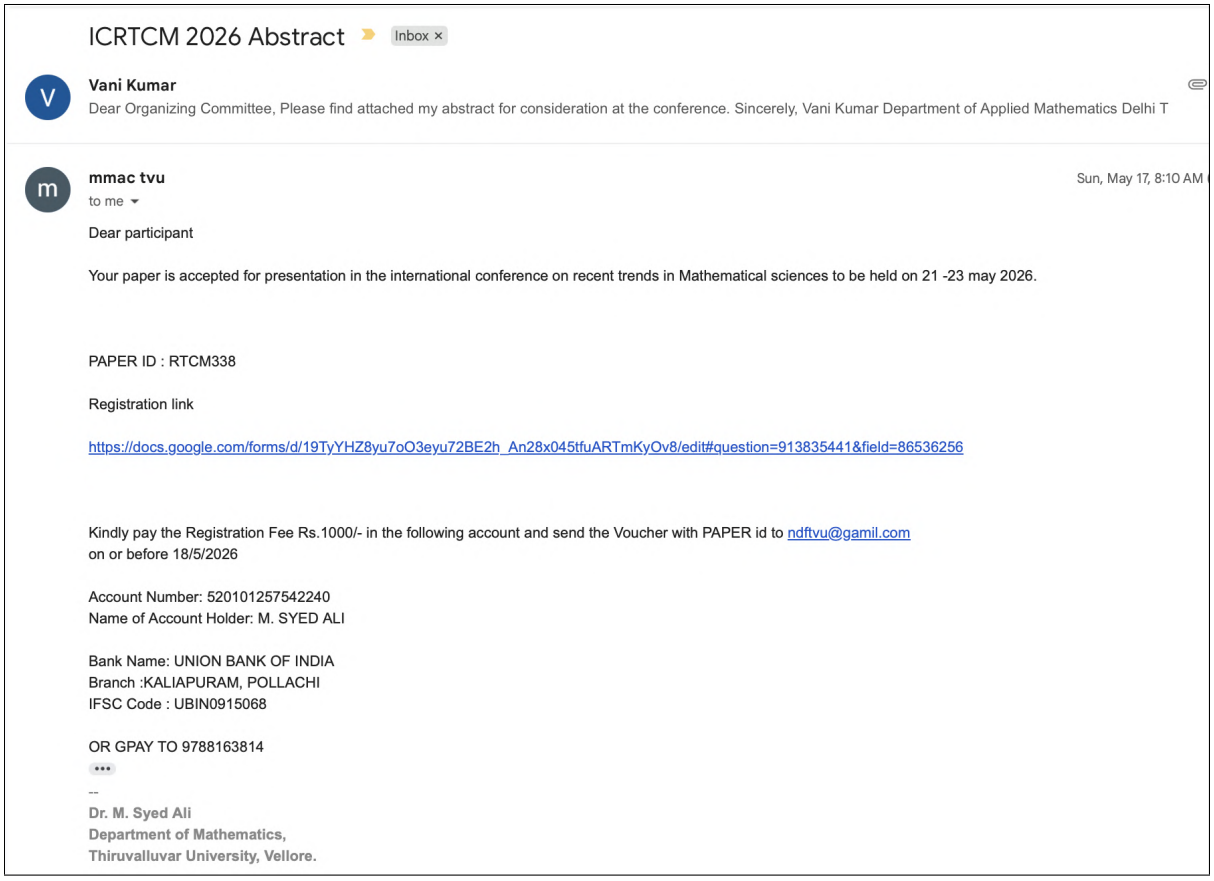
[127] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Annexure

CONFERENCES

1. International Conference on Recent Trends in Mathematical Sciences
   Organized by The Department of Mathematics, Thiruvalluvar University

2. International Conference on Recent Advances in Mathematics and Mathematical Sciences
   Organized by The Department of Mathematics H. N. B. Garhwal University Garhwal University, Dr B. G. R. Campus Pauri



ICRTCM 2026 Abstract Inbox ×

Vani Kumar
Dear Organizing Committee, Please find attached my abstract for consideration at the conference. Sincerely, Vani Kumar Department of Applied Mathematics Delhi T

mmac tvu Sun, May 17, 8:10 AM
to me

Dear participant

Your paper is accepted for presentation in the international conference on recent trends in Mathematical sciences to be held on 21 -23 may 2026.

PAPER ID : RTCM338

Registration link

https://docs.google.com/forms/d/19TyYHZ8yu7oO3eyu72BE2h_An28x045tfuARTmKyOv8/edit#question=913835441&field=86536256

Kindly pay the Registration Fee Rs.1000/- in the following account and send the Voucher with PAPER id to ndftvu@gamil.com
on or before 18/5/2026

Account Number: 520101257542240
Name of Account Holder: M. SYED ALI

Bank Name: UNION BANK OF INDIA
Branch :KALIAPURAM, POLLACHI
IFSC Code : UBIN0915068

OR GPAY TO 9788163814

--
Dr. M. Syed Ali
Department of Mathematics,
Thiruvalluvar University, Vellore.

INTERNATIONAL CONFERENCE

ON

Recent Advances in Mathematics and Mathematical Sciences

Organized by

Department of Mathematics

H. N. B. Garhwal University, Dr B. G. R. Campus Pauri, Uttarakhand-246001, India

Collaboration with

Vijñāna Parishad of India

(27th Annual and 8th International Conference)

Dr. U. C. Gairola
Convener

Mobile No. (+91) 9760231159
E-mail: mathsbgr@gmail.com

Ref. RAMMS/2026/ASP046 Date: 18/05/2026

Dear Vani Kumar,

Thanks for the abstract submission to the **27th Annual & 8th International Conference** of **Vijñāna Parishad of India (VPI)** on **Recent Advances in Mathematics and Mathematical Sciences** (RAMMS-2026).

It is our pleasure to inform that your Abstract entitled "A Spectral-Geometric Framework for Non-Redundant Drug Target Selection via the Weighted Lovász Theta Function: Application to the *Mycobacterium tuberculosis* Interactome" has been accepted for the presentation. You are cordially invited to present your paper orally at RAMMS-2026 to be held during during **June 4-6, 2026** at HNB Garhwal University, BGR Campus, Pauri. We would like to request for the mandatory registration through the following link (ignore, if you have done already).

Registration Link: **https://forms.gle/tQhpXbMZcQhLZQVK6**

Thank you for your cooperation. We look forward!

With best regards

Team
(Organizing Committee)
RAMMS-2026

diss_main_plag.pdf

Indian Institute of Technology Jodhpur

Document Details

Submission ID
trn:oid:::29334:139894624

Submission Date
May 21, 2026, 7:03 PM GMT+5:30

Download Date
May 21, 2026, 7:07 PM GMT+5:30

File Name
diss_main_plag.pdf

File Size
12.1 MB

81 Pages

22,636 Words

130,255 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

Match Groups

21 Not Cited or Quoted 1%
Matches with neither in-text citation nor quotation marks

32 Missing Quotations 3%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
3% Publications
1% Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

Replaced Characters
120 suspect characters on 34 pages
Letters are swapped with similar characters from another alphabet.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **21** Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- **32** Missing Quotations 3%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
3% Publications
1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
www.researchgate.net		<1%
2	Internet	
www.nature.com		<1%
3	Publication	
Parissa Farnia, Ali Akbar Velayati, Jalaledin Ghanavi, Poopak Farnia. "Proteins in ...		<1%
4	Internet	
www.ncbi.nlm.nih.gov		<1%
5	Internet	
www.pubmedcentral.nih.gov		<1%
6	Publication	
Qiong Wu, Yanhua Zhou, Xiangzhi Zhou, Huashan Zhou et al. "Independent Risk F...		<1%
7	Submitted works	
Higher Education Commission Pakistan on 2017-08-13		<1%
8	Internet	
dash.harvard.edu		<1%
9	Publication	
Rohan Mahadevan. "Reconciling the spectrum of Sagittarius A* with a two-tempe...		<1%
10	Internet	
msb.embopress.org		<1%

11 Publication
혜원 이, 진선 김, 지은 김, 영준 박. "2023년 국제 결핵 발생 현황", Public Health Weekly Rep... <1%

12 Publication
Louis Caccetta, Rui Zhong Jia. "On a problem concerning ordered colourings", Dis... <1%

13 Internet
academic.oup.com <1%

14 Internet
mahamedianews.com <1%

15 Internet
www.wjgnet.com <1%

16 Internet
ann-clinmicrob.biomedcentral.com <1%

17 Internet
appliednetsci.springeropen.com <1%

18 Publication
Wenqiang Zhou, Qingfeng Wang, Wenjuan Nie, Wenhui Shi, Yang Yang, Wenjie Qi... <1%

19 Internet
assets.publishing.service.gov.uk <1%

20 Submitted works
University of the Philippines Los Banos on 2024-12-09 <1%

21 Internet
dollar.biz.uiowa.edu <1%

22 Internet
sigma.yildiz.edu.tr <1%

23 Internet
www.preprints.org <1%

24 Publication
Huimin Zhang, Shuo Yan, Ruilin Du, Zimeng Ma, Yue Xue, Yulong Zhao, Wenna Ya... <1%

25 Publication

Mona Singh. "From Protein Interaction Networks to Protein Function", Computati... <1%

26 Publication

Neetika Jaisinghani, Mary L. Previti, Joshua Andrade, Manor Askenazi, Beatrix Ue... <1%

27 Internet

doc.lagout.org <1%

28 Publication

"Algorithms and Computation", Springer Science and Business Media LLC, 2008 <1%

29 Publication

Bernd Gärtner, Jiri Matousek. "Approximation Algorithms and Semidefinite Progr... <1%

30 Publication

Guojing Cong, Seung-Hwan Lim, Steven Young. "Augmenting Graph Convolution ... <1%

31 Publication

Karthik Raman. "An Introduction to Computational Systems Biology - Systems-Le... <1%

32 Publication

Tilahun Melak, Sunita Gakkhar. "Maximum flow approach to prioritize potential d... <1%

33 Internet

mural.maynoothuniversity.ie <1%

34 Publication

"ECAI 2020", IOS Press, 2020 <1%

35 Publication

"Protein-protein Interactions and Networks", Springer Nature, 2008 <1%

36 Publication

Bram Bekker, Olga Kuryatnikova, Fernando de Oliveira Filho, Juan Vera. "Optimiz... <1%

37 Submitted works

Cornell University on 2020-08-31 <1%

38 Publication

Lenore Cowen, Trey Ideker, Benjamin J. Raphael, Roded Sharan. "Network propag... <1%

39 Submitted works

University of KwaZulu-Natal on 2024-12-06 <1%

40 Publication

Vivek P. Chavda, Mahesh T. Chhabria, Divya M. Teli. "Recent Advancements in Tub... <1%

41 Internet

arxiv.org <1%

42 Internet

link.springer.com <1%

43 Internet

oai.cwi.nl <1%

44 Internet

www.hindawi.com <1%

45 Internet

www.mathematicaljournal.com <1%

46 Internet

www.networkatlas.eu <1%

diss_main_plag.pdf

Indian Institute of Technology Jodhpur

Document Details

Submission ID
trn:oid:::29334:139894624

Submission Date
May 21, 2026, 7:03 PM GMT+5:30

Download Date
May 21, 2026, 7:10 PM GMT+5:30

File Name
diss_main_plag.pdf

File Size
12.1 MB

81 Pages

22,636 Words

130,255 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.