

Integrated Subtractive Proteomics and Molecular Docking Reveal a Novel Therapeutic Target and Tuberosin as a Lead Phytochemical Against *Streptococcus pneumoniae*

Ahaana Jain
ahaana200609@gmail.com

ABSTRACT

Streptococcus pneumoniae remains a leading cause of community-acquired pneumonia, meningitis and invasive pneumococcal disease, and rising antimicrobial resistance has undermined existing therapeutics, creating an urgent need for novel, microbiome-sparing drug targets. We applied an integrated, six-stage computational pipeline—combining subtractive proteomics, homology modelling and structure-based virtual screening—to the hypothetical proteome of *S. pneumoniae* strain OTU45 (GenBank CP113114.1). Of 282 hypothetical proteins, sequential filtering on physicochemical druggability, VaxiJen-predicted non-antigenicity and zero BLASTp identity to ten representative gut-commensal proteomes shortlisted 70 candidates (24.8%). TMHMM 2.0 and InterProScan/PHOBIUS topology prediction, combined in a seven-criterion composite scoring matrix (max 35), identified OTU45_03800, a six-pass integral membrane protein of unknown function, as the top-ranked target (score 30/35; GRAVY +0.985; VaxiJen 0.293; 28.19 kDa). A homology model of OTU45_03800 built in SWISS-MODEL was docked in AutoDock Vina (via PyRx) against 211 ADMET-compliant phytochemicals drawn from an initial library of 523 natural products screened using Lipinski, Ghose, Veber, Egan and Muegge drug-likeness filters and SwissADME-based pharmacokinetic profiling. Nineteen compounds achieved binding energies ≤ -8.1 kcal/mol; although nordihydrotoxiferine, dracorubin and morellin bound most strongly (-9.6 to -9.4 kcal/mol), each carried disqualifying ADMET liabilities. Integrated binding-affinity and drug-likeness ranking identified Tuberosin, a pterocarpan-class phytochemical (-8.5 kcal/mol), as the overall lead: it satisfied all five drug-likeness filters with zero violations, showed high predicted gastrointestinal absorption and blood–brain-barrier permeability, and engaged the OTU45_03800 binding cavity with no unfavourable steric contacts. γ -Galbacin, Hispaglabridin-B, Berbamine and Sesamin constituted a ranked series of secondary candidates. This study nominates OTU45_03800 as a previously unexplored, computationally validated antibacterial target and Tuberosin as a developable natural-product lead, providing a rational framework for next-generation anti-pneumococcal drug discovery pending experimental validation.

Aug 2026
Vol 10, No 4.

Keywords: Streptococcus pneumoniae; subtractive proteomics; hypothetical membrane protein; molecular docking; ADMET; drug-likeness; natural products; Tuberodin; antibacterial drug discovery

ABBREVIATIONS

ADMET, Absorption, Distribution, Metabolism, Excretion, and Toxicity; AMR, antimicrobial resistance; BBB, blood–brain barrier; BE, binding energy; BLASTp, Basic Local Alignment Search Tool for proteins; cLogP, consensus calculated logarithm of the octanol–water partition coefficient; GI, gastrointestinal; GRAVY, grand average of hydropathicity; HP, hypothetical protein; II, instability index; kDa, kilodalton; MW, molecular weight; NCBI, National Center for Biotechnology Information; PAINS, pan-assay interference compounds; PCV, pneumococcal conjugate vaccine; pI, isoelectric point; PDB, Protein Data Bank; PDBQT, Protein Data Bank, Partial Charge (Q), and Atom Type (T); QMEANDisCo, Qualitative Model Energy Analysis–Distance Constraints (SWISS-MODEL quality estimation score); SA score, synthetic accessibility score; TM, transmembrane; TMH, transmembrane helix; TPSA, topological polar surface area;

INTRODUCTION

Streptococcus pneumoniae remains one of the leading bacterial pathogens responsible for community-acquired pneumonia, meningitis, otitis media, septicemia, and other invasive pneumococcal diseases, contributing substantially to global morbidity and mortality, particularly among children under five years of age, older adults, and immunocompromised individuals. Despite the widespread implementation of pneumococcal conjugate vaccines (PCVs), the emergence of non-vaccine serotypes and the rapid dissemination of antimicrobial resistance (AMR) have significantly undermined current therapeutic strategies. Increasing resistance to β -lactams, macrolides, fluoroquinolones, and tetracyclines has reduced treatment efficacy, highlighting the urgent need for novel antibacterial agents targeting previously unexplored bacterial proteins (Khan et al., 2022; Syaad et al., 2023).

Traditional antibiotic discovery has largely focused on well-characterized metabolic enzymes and essential cellular processes. However, decades of intensive antibiotic use have accelerated the evolution of resistance mechanisms against these conventional targets. Consequently, modern antimicrobial research has shifted toward the identification of novel essential proteins with limited homology to human and commensal microbial proteins, thereby reducing the likelihood of host toxicity and microbiome disruption. Advances in bacterial genomics, comparative proteomics, and structural bioinformatics have transformed antimicrobial target discovery by enabling genome-wide identification of essential, pathogen-specific proteins suitable for therapeutic intervention. Among these computational strategies, subtractive proteomics has emerged as one of the most effective approaches for prioritizing drug targets by sequentially eliminating redundant proteins, homologs of host proteins, and proteins shared with

beneficial microbiota while retaining essential pathogen-specific proteins involved in bacterial survival (Khan et al., 2022; Alanzi et al., 2025).

Subtractive proteomics integrates multiple bioinformatics analyses, including essentiality prediction, metabolic pathway analysis, subcellular localization, transmembrane topology, antigenicity assessment, and druggability evaluation, to identify proteins with high therapeutic potential. This strategy has been successfully applied to several clinically important pathogens. For example, Khan et al. (2022) identified novel therapeutic targets in *S. pneumoniae* serotype 14 through systematic subtractive genomics, whereas Syaad et al. (2023) employed a similar workflow in the *S. pneumoniae* D39 strain to prioritize essential proteins suitable for molecular docking and virtual screening. Likewise, Alanzi et al. (2025) demonstrated the applicability of subtractive proteomics in identifying novel membrane-associated drug targets in multidrug-resistant *Klebsiella michiganensis*. Collectively, these studies demonstrate that computational target prioritization substantially accelerates early-stage antibacterial drug discovery while minimizing experimental cost and time.

An emerging area of antimicrobial research focuses on hypothetical proteins, which constitute a considerable proportion of bacterial genomes yet remain functionally uncharacterized. These proteins frequently represent lineage-specific or species-specific molecules that may perform indispensable biological functions without exhibiting significant similarity to known proteins. Because hypothetical proteins are generally absent from existing antibiotic resistance pathways, they represent an attractive source of completely novel therapeutic targets. Furthermore, membrane-associated hypothetical proteins are particularly valuable because their structural accessibility facilitates small-molecule binding and increases the likelihood of successful pharmacological modulation. Nevertheless, relatively few studies have systematically investigated hypothetical membrane proteins in *S. pneumoniae*, leaving an important gap in pneumococcal drug discovery.

Following target identification, structure-based virtual screening has become an indispensable component of modern drug discovery. Molecular docking predicts ligand–protein interactions by estimating binding affinity and identifying key intermolecular interactions within the active site, thereby enabling rapid prioritization of compounds before experimental validation. However, binding affinity alone is insufficient to predict successful drug candidates because many compounds with excellent docking scores possess poor pharmacokinetic characteristics. Consequently, contemporary computational drug discovery integrates molecular docking with comprehensive absorption, distribution, metabolism, excretion, and toxicity (ADMET) evaluation to identify compounds with favorable physicochemical and pharmacological properties (Lipinski et al., 1997).

The Lipinski Rule of Five established a fundamental framework for predicting oral drug-likeness based on molecular weight, lipophilicity, and hydrogen-bonding characteristics. Since its introduction, this framework has been expanded by complementary filters, including the Veber, Ghose, Egan, and Muegge criteria, together with computational platforms such as SwissADME that evaluate gastrointestinal absorption, blood–brain barrier permeability, synthetic accessibility, structural alerts, and bioavailability.

These computational approaches substantially improve the identification of compounds with greater clinical development potential and have become standard practice in contemporary computer-aided drug discovery (Lipinski et al., 1997; Syaad et al., 2023).

Natural products continue to represent one of the richest sources of structurally diverse bioactive molecules for antibacterial drug development. Approximately half of currently approved pharmaceuticals originate directly or indirectly from natural products, reflecting their exceptional chemical diversity and evolutionary optimization for biological activity. Medicinal plants synthesize numerous secondary metabolites, including alkaloids, flavonoids, terpenoids, lignans, coumarins, phenolics, and saponins, that exhibit antimicrobial, antioxidant, anti-inflammatory, anticancer, and immunomodulatory activities. Among medicinal plants, *Glycyrrhiza glabra* (licorice) has attracted considerable attention because of its abundance of bioactive phytochemicals, including glycyrrhizin, glycyrrhetic acid, glabridin, liquiritin, isoliquiritigenin, hispaglabridins, and other isoflavonoids that display broad-spectrum pharmacological activities against bacterial and viral pathogens (Sharma et al., 2018; Wang et al., 2022).

Several phytochemical classes, including pterocarpan, lignans, flavonoids, and benzyloquinoline alkaloids, have demonstrated significant antibacterial activity against Gram-positive bacteria through mechanisms involving membrane disruption, inhibition of cell-wall biosynthesis, oxidative stress induction, and interference with essential bacterial enzymes. These findings suggest that plant-derived compounds constitute valuable chemical scaffolds for discovering novel antibacterial agents capable of overcoming existing antimicrobial resistance mechanisms. Importantly, computational screening enables rapid evaluation of hundreds of phytochemicals against newly identified bacterial targets, greatly accelerating lead compound identification before experimental testing.

In the present study, we employed an integrated computational drug discovery pipeline to identify novel therapeutic candidates against *S. pneumoniae* OTU45. A comprehensive subtractive proteomics strategy was implemented to identify essential, non-host homologous, non-antigenic, gut microbiota-specific hypothetical proteins with favorable physicochemical characteristics. The prioritized target, OTU45_03800, was subsequently modeled structurally and subjected to molecular docking against a curated library of phytochemicals. The top-ranked compounds were further evaluated using comprehensive SwissADME-based drug-likeness and pharmacokinetic analyses to identify compounds possessing both strong binding affinity and favorable pharmaceutical characteristics. This integrated workflow identified Tuberosin, a pterocarpan-class phytochemical, as the most promising lead compound based on its favorable interaction profile, absence of unfavorable steric interactions, compliance with multiple drug-likeness criteria, and excellent predicted oral bioavailability. Collectively, this study provides a rational computational framework for identifying previously unexplored membrane protein targets and naturally derived lead compounds for the development of next-generation therapeutics against drug-resistant *S. pneumoniae*.

METHODOLOGY

1. Overview of the study design

This study employed an integrated, six-stage computational pipeline combining subtractive genomics/proteomics, homology-based structure prediction and structure-based virtual screening to (i) identify a novel, microbiome-sparing drug target in the hypothetical proteome of *Streptococcus pneumoniae* strain OTU45 (GenBank accession CP113114.1), and (ii) screen a phytochemical library for candidate inhibitors of the prioritised target. The pipeline proceeded through sequential stages of proteome retrieval, physicochemical profiling, antigenicity screening, gut-commensal non-homology filtering and transmembrane topology prediction to shortlist candidate drug targets; homology modelling and structural preparation of the top-ranked target; acquisition and drug-likeness filtering of a natural-product ligand library; molecular docking-based virtual screening; and post-docking interaction analysis and composite ranking of lead compounds. Each stage is described in turn below.

Box 1 | Schematic summary of the six-stage computational pipeline, added in response to reviewer comments on methodological transparency.

Stage	Key step(s)	Databases / tools & key parameters	Principal output
1. Proteome retrieval	Retrieval of the complete <i>S. pneumoniae</i> OTU45 proteome; extraction of proteins annotated “hypothetical protein”	NCBI genome database, accession CP113114.1 (FASTA)	282 hypothetical proteins
2. Subtractive filtering	Physicochemical profiling; antigenicity screening; gut-commensal non-homology filtering; transmembrane topology prediction; composite scoring	ProtParam/ExPASy; VaxiJen 2.0 (threshold 0.5); BLASTp vs. 10 gut-commensal proteomes (0% identity retained); TMHMM 2.0 + InterProScan 5/PHOBIUS; 7-criterion matrix (max 35)	70 candidates shortlisted (24.8%); OTU45_03800 top-ranked (30/35)
3. Structural preparation	Homology modelling of OTU45_03800; receptor structural preparation	SWISS-MODEL (ProMod3 engine; QMEANDisCo scoring); UCSF Chimera; BIOVIA Discovery Studio Visualizer v21.1.0.20298	Energy-minimised receptor, OTU45_03800_.pd bqt
4. Ligand library & ADMET	Phytochemical library acquisition; structure preparation; drug-likeness/ADMET pre-screening	Dr. Duke’s Phytochemical & Ethnobotanical Databases; PubChem; Open Babel/PyRx v0.8; SwissADME (Lipinski, Ghose, Veber, Egan, Muegge; PAINS; Brenk)	523 candidate phytochemicals filtered to 211 ADMET-compliant compounds (40.3%)

5. Molecular docking	Structure-based virtual screening of ADMET-compliant ligands against OTU45_03800	AutoDock Vina (Lamarckian genetic algorithm) via PyRx Vina Wizard; grid centre (0.7591, 0.2666, 0.7681); size 59.49 × 39.92 × 55.34 Å; exhaustiveness 8	19 compounds with binding energy ≤ −8.1 kcal/mol
6. Interaction analysis & ranking	Post-docking 2D/3D interaction analysis; composite ranking integrating binding energy, ADMET profile and interaction quality	BIOVIA Discovery Studio Visualizer; UCSF Chimera	Tuberosin selected as lead compound; γ -Galbacin, Hispaglabridin-B, Berbamine and Sesamin as secondary candidates

2. Retrieval of the *S. pneumoniae* OTU45 proteome and identification of hypothetical proteins

The complete annotated proteome of *S. pneumoniae* strain OTU45 (accession CP113114.1) was retrieved from the National Center for Biotechnology Information (NCBI) genome database in FASTA format. Proteins annotated as “hypothetical protein” were extracted computationally from the whole-proteome dataset to form the working dataset for subtractive analysis, on the premise that uncharacterised proteins are disproportionately likely to yield genuinely novel drug targets with no pre-existing resistance mechanisms, in contrast to well-characterised metabolic enzymes that have already been extensively explored as antibacterial targets (Khan et al., 2022; Syaad et al., 2023). This yielded a dataset of 282 hypothetical proteins that was carried forward into the sequential subtractive filtering stages described in Section 3.

3. Sequential subtractive proteomics pipeline

3.1 Physicochemical profiling

Each of the 282 hypothetical proteins was subjected to physicochemical characterisation using the ProtParam tool on the ExPASy server (Gasteiger et al., 2005). Molecular weight, theoretical isoelectric point (pI), instability index, aliphatic index and grand average of hydropathicity (GRAVY) were computed for every sequence. These parameters were used as an initial druggability filter, prioritising proteins of compact molecular weight, favourable thermostability (high aliphatic index, low instability index) and, for membrane-target candidates, high predicted hydrophobicity (positive GRAVY score), consistent with the physicochemical criteria applied in comparable subtractive-proteomics studies of *S. pneumoniae* (Syaad et al., 2023) and other Gram-negative and Gram-positive pathogens (Alanzi et al., 2025).

3.2 Antigenicity screening

The antigenic potential of each candidate was predicted using VaxiJen 2.0 (Doytchinova & Flower, 2007), an alignment-independent server based on auto cross-covariance transformation of physicochemical amino acid properties. A default prediction threshold of 0.5 was applied, with sequences scoring below this cut-off classified as non-antigenic. Because the present study pursued a small-molecule inhibitor strategy rather than a vaccine-antigen strategy, non-antigenic proteins were preferentially retained: a non-antigenic target minimises competitive host-immune engagement that could otherwise reduce effective drug concentration at the target site, a selection logic analogous to that applied by Syaad et al. (2023) in reserving non-antigenic proteins for downstream small-molecule inhibitor development.

3.3 Gut-microbiota non-homology filtering

To bias target selection toward microbiome-sparing, narrow-spectrum candidates, each hypothetical protein was screened by BLASTp (Altschul et al., 1990) against the reference proteomes of ten representative human gut-commensal species: *Bacteroides* sp., *Faecalibacterium prausnitzii*, *Lactiplantibacillus plantarum*, *Bifidobacterium longum*, *Akkermansia muciniphila*, *Lacticaseibacillus rhamnosus*, *Ruminococcus bromii*, *Roseburia hominis*, *Segatella (Prevotella) copri* and *Clostridium scindens*. Proteins returning 0% sequence identity across all ten commensal proteomes were retained as candidates with no predicted commensal cross-reactivity. This gut-microbiota non-homology filter was implemented as an explicit, dedicated pipeline stage a feature not incorporated by comparable prior subtractive-proteomics studies of *S. pneumoniae* (Khan et al., 2022; Syaad et al., 2023) reflecting the clinical imperative of preserving the gut microbiome during narrow-spectrum antibacterial drug design.

3.4 Transmembrane topology prediction

Transmembrane helical topology was predicted for all candidates using two independent algorithms: TMHMM 2.0, a hidden Markov model-based predictor of transmembrane helices (Krogh et al., 2001), and PHOBIUS, accessed via the InterProScan 5 platform (Jones et al., 2014), which implements a combined transmembrane-topology and signal-peptide prediction model (Käll et al., 2004). For each candidate, the number and boundaries of predicted transmembrane helices, the probability of N-terminal intracellular localisation, and the extent of cross-tool concordance between TMHMM and PHOBIUS annotation tracks (cytoplasmic regions, non-cytoplasmic regions, and TM helices predicted by each tool) were recorded. Candidates exhibiting multiple well-resolved transmembrane helices with strong agreement between the two independent prediction tools were prioritised, on the basis that integral membrane transporter/channel-forming scaffolds present hydrophobic cavities well suited to small-molecule engagement and are established target classes in antibacterial drug discovery.

3.5 Composite druggability scoring and final target selection

Candidates surviving the preceding filters were evaluated against a seven-criterion composite scoring matrix integrating: (i) transmembrane topology and cross-tool concordance, (ii) non-antigenicity, (iii) gut-microbiota non-homology, (iv) physicochemical druggability (molecular weight, pI, GRAVY, aliphatic index, instability index), (v) absence of close host or commensal homologues, (vi) predicted

structural/functional accessibility, and (vii) hypothetical-protein novelty status. Each candidate was scored against this matrix (maximum composite score 35), and the highest-scoring protein was carried forward as the primary drug target for structure-based virtual screening. This sequential filtering process reduced the initial dataset of 282 hypothetical proteins to a shortlist of 70 candidate proteins (24.8%), from which the six-pass integral membrane protein OTU45_03800 was selected as the top-ranked candidate (composite score 30/35), on the strength of its six-transmembrane topology with the strongest TMHMM–PHOBIUS concordance of any candidate, non-antigenic character (VaxiJen 0.293), zero sequence identity to all ten gut-commensal proteomes screened, and a strongly hydrophobic, thermostable physicochemical profile (28.19 kDa; pI 9.33; GRAVY +0.985; aliphatic index 132.05; instability index 38.96).

4. Homology modelling and structural preparation of OTU45_03800

As no experimentally determined structure was available for OTU45_03800, a three-dimensional homology model was generated using the SWISS-MODEL automated protein structure homology-modelling server (Waterhouse et al., 2018). The OTU45_03800 amino acid sequence was submitted to SWISS-MODEL, which identified the most suitable structural template(s) by sequence and profile-based template search, generated the model using the ProMod3 modelling engine, and assessed model quality using the QMEANDisCo global and local quality-estimation scores. The resulting homology model was downloaded in PDB format and prepared for docking using UCSF Chimera (Pettersen et al., 2004) and BIOVIA Discovery Studio Visualizer (Dassault Systèmes BIOVIA, 2021). Structural preparation comprised removal of crystallographic/model water molecules and heteroatoms, addition of polar hydrogens, assignment of Kollman/Gasteiger partial charges, and energy minimisation using UCSF Chimera, followed by visual inspection of the six-transmembrane fold and identification of the putative ligand-binding cavity using Discovery Studio. The prepared receptor structure was subsequently converted to PDBQT format (receptor file OTU45_03800_.pdbqt) for use as the docking target in AutoDock Vina.

5. Phytochemical ligand library acquisition and structure preparation

A natural-product ligand library was compiled from Dr. Duke's Phytochemical and Ethnobotanical Databases, an online repository developed by the United States Department of Agriculture, Agricultural Research Service, cataloguing phytochemical constituents, documented biological activities and ethnobotanical uses of higher plants (U.S. Department of Agriculture, Agricultural Research Service, 1992–2016). Compounds were screened for reported antibacterial, antimicrobial or related pharmacological activity, yielding an initial candidate library of 523 phytochemicals. The corresponding two-dimensional and three-dimensional structures (SDF format) of all 523 compounds were retrieved from the PubChem Compound database (Kim et al., 2023) using their PubChem Compound Identifier (CID) or compound name. Ligand structures were subsequently energy-minimised and converted to PDBQT format using the Open Babel utility integrated within PyRx (Dallakyan & Olson, 2015) in preparation for ADMET screening and molecular docking.

6. ADMET profiling and drug-likeness screening

Aug 2026
Vol 10, No 4.

All 523 phytochemicals were pre-screened for oral drug-likeness prior to docking using Lipinski's rule of five (Lipinski et al., 1997), which flags poor predicted oral absorption/permeation when two or more of the following criteria are violated: molecular weight > 500 Da, calculated LogP > 5, hydrogen-bond donors > 5, and hydrogen-bond acceptors > 10. Compounds were further profiled using the SwissADME web platform (Daina et al., 2017), which extended the Lipinski assessment to the Ghose, Veber, Egan and Muegge drug-likeness filters, together with BOILED-Egg prediction of gastrointestinal absorption and blood-brain barrier permeability, pan-assay interference compound (PAINS) alerts, Brenk structural-alert screening, bioavailability scoring, and synthetic accessibility scoring. Compounds violating two or more Lipinski criteria were excluded at this stage. Application of this multi-parametric ADMET filter reduced the candidate library from 523 to 211 compounds (40.3% retention), which were carried forward into molecular docking-based virtual screening.

7. Molecular docking-based virtual screening

7.1 Docking software and workflow

Structure-based virtual screening was performed using PyRx version 0.8 (Dallakyan & Olson, 2015), an open-source virtual screening suite that integrates AutoDock Vina (Trott & Olson, 2010) via its built-in Vina Wizard docking interface. The prepared OTU45_03800 homology model (receptor file: OTU45_03800_.pdbqt) was imported into PyRx as the rigid macromolecular target, and the 211 ADMET-compliant phytochemical ligands were imported as flexible ligand structures, with rotatable bonds assigned automatically by the AutoDock Vina Wizard.

7.2 Grid box configuration

A docking grid box centred on the predicted transmembrane ligand-binding cavity of OTU45_03800 was defined within the AutoDock Vina Wizard using the following coordinates and dimensions:

Receptor: OTU45_03800_.pdbqt

Center (Å): center_x = 0.7591, center_y = 0.2666, center_z = 0.7681

Size (Å): size_x = 59.4940506935, size_y = 39.9164006996, size_z = 55.3438607025

Exhaustiveness: 8

This grid box was sized to encompass the entire six-transmembrane helical bundle and the associated putative binding cavity of OTU45_03800, allowing an unbiased, blind-docking-style search of the transmembrane region for each ligand while remaining computationally tractable at the specified exhaustiveness setting. An exhaustiveness value of 8 (the AutoDock Vina default) was applied uniformly across all 211 docking runs to ensure a consistent and comparable sampling of the conformational search space for every ligand, following the docking protocol convention established by Trott and Olson (2010) and applied in comparable *S. pneumoniae* and Gram-negative subtractive-proteomics docking studies (Syaad et al., 2023; Alanzi et al., 2025).

7.3 Docking execution and scoring

For each of the 211 ligands, AutoDock Vina was executed via the PyRx Vina Wizard using the empirical scoring function and iterated local search (Lamarckian genetic-algorithm-derived) global optimisation procedure described by Trott and Olson (2010), under the grid box and exhaustiveness parameters specified above. For each ligand, PyRx returned up to nine ranked binding poses together with their predicted binding free energies (kcal/mol); the pose with the most negative (most favourable) binding energy was retained as the representative docking result for that compound. Compounds were ranked in ascending order of binding energy, and a threshold of ≤ -8.1 kcal/mol was applied to define the set of highest-affinity hits for downstream interaction and composite-ranking analysis.

8. Post-docking interaction analysis and composite ranking of lead candidates

The docked protein–ligand complexes for the top-ranked compounds were visualised and analysed in BIOVIA Discovery Studio Visualizer (Dassault Systèmes BIOVIA, 2021) to characterise the two-dimensional and three-dimensional molecular interaction profile of each pose, including conventional and carbon hydrogen bonds, π – π stacked and T-shaped interactions, π -alkyl and alkyl hydrophobic contacts, π -sulfur and π -sigma interactions, and any unfavourable steric clashes or acceptor–acceptor contacts. UCSF Chimera (Pettersen et al., 2004) was used in parallel for structural superposition, cavity visualisation and figure preparation. Final lead-compound selection integrated AutoDock Vina binding energy with the SwissADME-derived drug-likeness profile (number of violations across the Lipinski, Ghose, Veber, Egan and Muegge filters), predicted gastrointestinal absorption and blood–brain barrier permeability, bioavailability score, synthetic accessibility score, and the qualitative absence of unfavourable docking contacts, following the composite selection logic balancing binding affinity against ADMET compliance and interaction quality rather than binding energy alone applied in comparable *S. pneumoniae* and *Klebsiella michiganensis* subtractive-proteomics and docking studies (Syaad et al., 2023; Alanzi et al., 2025).

RESULTS

PART I Drug-target identification

1. Subtractive profiling defines a 282-member hypothetical proteome for target discovery

To identify novel, druggable targets in *Streptococcus pneumoniae* OTU45, we subjected the full set of computationally predicted hypothetical proteins to an integrated in silico pipeline combining physicochemical characterization, structural prediction, antigenicity screening and comparative homology filtering against representative human gut commensals (*Bacteroides* sp., *Faecalibacterium prausnitzii*, *Lactiplantibacillus plantarum*, *Bifidobacterium longum*, *Akkermansia muciniphila*, *Lacticaseibacillus rhamnosus*, *Ruminococcus bromii*, *Roseburia hominis*, *Segatella* (Prevotella) *copri* and *Clostridium scindens*). A total of 282 hypothetical proteins with no assigned function were retrieved and fully

characterised, ranging from 20 to 1,197 residues (mean $\pm$ s.d., 126.6 ± 109.8 aa) and 2.2–130.9 kDa in predicted molecular weight.

2. Physicochemical profiling and antigenicity screening

ProtParam-based profiling classified 194 of 282 proteins (68.8%) as thermodynamically stable (instability index < 40) and the remaining 88 (31.2%) as unstable. Predicted isoelectric points spanned a broad basic-to-acidic range, and grand average of hydropathicity (GRAVY) values indicated a predominantly hydrophilic hypothetical proteome, consistent with a mixture of cytoplasmic and surface-exposed proteins. Secondary-structure prediction showed random coil as the dominant conformation (mean $33.2 \pm 13.9\%$). Protective antigen scoring (VaxiJen 2.0) classified 168 of 282 proteins (59.6%) as predicted antigens and 114 (40.4%) as non-antigens. Non-antigenic candidates (VaxiJen score ≤ 0.5) were retained as priorities for small-molecule inhibitor development, as their low immunodominance minimises competitive immune clearance that could confound drug efficacy in vivo.

3. Gut-microbiota homology filtering identifies microbiome-sparing candidates

BLASTp comparison of each hypothetical protein against ten representative gut-commensal proteomes revealed that 197 of 282 proteins (69.9%) returned 0% identity to all ten commensal species; a further 9 (3.2%) showed $\leq 30\%$ maximum identity; 42 (14.9%) showed 30–60%; and only 34 (12.1%) exceeded 60% identity to at least one commensal. This distribution confirms that the majority of *S. pneumoniae* hypothetical proteins are structurally distinct from gut-microbiota proteins, supporting their suitability as selective, microbiome-sparing drug targets.

4. Integrated prioritization identifies top-ranked candidates

Antigenicity, stability and microbiome non-homology were combined into a composite priority score (weighted 0.5 : 0.25 : 0.25, respectively). Intersecting all three criteria yielded a shortlist of 70 hypothetical proteins (24.8%) as high-confidence, host- and microbiome-sparing target candidates (Table 1). The highest-priority candidate, HP-77, achieved a composite score of 0.90 a small (54-residue, 6.56 kDa), highly basic (pI 9.87) protein with a protective antigen score of 0.99, aliphatic index 131.85, and zero gut-commensal identity.

Table 1 | Top ten prioritized hypothetical-protein drug-target candidates ranked by composite priority score.

Rank	ID	Length (aa)	pI	MW (kDa)	Instab. index	Aliphatic index	GRAV Y	Antigen score	Max. % id.	Priority score
1	HP-77	54	9.87	6.56	35.25	131.85	0.69	0.99	0	0.90
2	HP-82	199	6.09	21.58	25.22	77.44	−0.59	0.83	0	0.86
3	HP-265	68	9.46	8.29	14.27	116.03	0.43	0.83	0	0.86

4	HP-35	70	9.46	8.21	28.07	175.43	1.19	0.82	0	0.85
5	HP-92	250	9.32	29.79	26.87	117.72	0.81	0.81	0	0.85
6	HP-195	95	10.13	11.02	22.27	165.16	1.50	0.81	0	0.85
7	HP-274	61	9.75	6.94	25.60	76.89	-0.74	0.80	0	0.85
8	HP-161	67	4.28	7.62	37.53	85.97	-0.82	0.79	0	0.85
9	HP-93	232	9.58	27.84	31.93	129.31	0.85	0.77	0	0.84
10	HP-34	242	9.52	28.50	32.49	137.31	0.82	0.76	0	0.84

Composite priority score = $0.5 \times (\text{antigen score}) + 0.25 \times (1 - \text{norm. instability index}) + 0.25 \times (1 - \text{norm. max. \% identity})$. *Instab.:* instability index. *GRAVY:* grand average of hydropathicity.

5. Transmembrane topology predictions identify OTU45_03800 as the primary drug target

To identify membrane-embedded hypothetical proteins amenable to small-molecule inhibition, all 282 candidates were subjected to TMHMM 2.0 transmembrane topology prediction. Four proteins OTU45_03195, OTU45_03800, OTU45_09020 and OTU45_09070 were selected on the basis of predicted hydrophobicity (GRAVY > 0) and high aliphatic index. All four returned N-terminal intracellular probability > 0.92 (Table 2). Three candidates (OTU45_03800, OTU45_09020, OTU45_09070) were predicted to carry six TM helices, consistent with a six-pass integral membrane protein topology, while OTU45_03195 carried two TM helices. OTU45_03800 was further distinguished by a possible N-terminal signal sequence, an expected 133.49 amino acids within TM helices the highest membrane embedding of the four and a predicted topology spanning residues 17–248 across six helices.

Table 2 | TMHMM 2.0 transmembrane topology predictions for the four hypothetical protein drug-target candidates.

Locus tag	Protein ID	Length (aa)	Pred. TMHs	Exp. aa in TMHs	Exp. aa first 60	N-in prob.	N-term signal
OTU45_03195	WAC86474.1	259	2	63.48	0.00036	0.998	None
OTU45_03800	WAC86582.1	254	6	133.49	29.99	0.927	Possible
OTU45_09020	WAC85418.1	263	6	130.41	21.59	0.999	Possible
OTU45_09070	WAC85427.1	255	6	136.34	31.76	1.000	Possible

TMHs: transmembrane helices. *Exp.:* expected. *N-in prob.:* probability of N-terminus being intracellular.

6. InterProScan 5 independently validates the six-TM architecture of OTU45_03800

All four candidate proteins were submitted to InterProScan 5 integrating PHOBIUS and TMHMM analysis simultaneously. All proteins were classified under "Coiled-Coils, Signal Peptides and Transmembrane Regions" with protein family membership "None predicted" for all, confirming their status as genuinely hypothetical proteins with no known homologues in curated databases (Table 3). PHOBIUS independently corroborated the six-TM architecture of OTU45_03800, with cytoplasmic and non-cytoplasmic loop annotations precisely matching TMHMM 2.0. OTU45_03800 achieved the strongest cross-tool concordance, with all four annotation tracks (PHOBIUS cytoplasmic, PHOBIUS TM helices, PHOBIUS non-cytoplasmic, TMHMM TM) fully populated. OTU45_03195 returned only TMHMM hits with two TM helices and no PHOBIUS annotations, independently confirming its two-pass topology.

Table 3 | InterProScan 5 topology annotation and cross-tool concordance for the four drug-target candidates.

Locus tag	Protein ID	Genomic location	Strand	Length (aa)	Tools matched	Topology features
OTU45_03195	WAC86474.1	631684–632463	+	259	TMHMM	2 TM regions only; no PHOBIUS cytoplasmic/non-cytoplasmic features detected
OTU45_03800	WAC86582.1	765782–766546	+	254	TMHMM + PHOBIUS ×3	6 TM helices; PHOBIUS cytoplasmic, non-cytoplasmic loops and TM helices all confirmed. 4 tool tracks populated. Strongest cross-tool concordance.
OTU45_09020	WAC85418.1	1783389–1784180	–	263	TMHMM + PHOBIUS ×3	6 TM helices confirmed; deprioritised as primary target owing to higher antigenicity (VaxiJen 0.705).
OTU45_09070	WAC85427.1	1791059–1791826	+	255	TMHMM + PHOBIUS ×3	6 TM helices; full cytoplasmic and non-cytoplasmic concordance across all tools.

PHOBIUS ×3: three annotation tracks returned (cytoplasmic region, TM helices, non-cytoplasmic region).

7. Multi-criteria druggability scoring confirms OTU45_03800 as the primary drug target

To rank candidates objectively, a seven-criterion scoring matrix (Table 4) was applied covering TM topology, GRAVY hydrophobicity, non-antigenicity, microbiome selectivity, protein stability, molecular weight suitability and N-terminal surface accessibility. OTU45_03800 achieved the highest aggregate score of 30/35, primarily distinguished by its combination of deep membrane embedding (six TM helices, GRAVY +0.985), non-antigenicity (VaxiJen 0.293), absence of gut-commensal homology (0%), and optimal compact molecular weight (28.19 kDa). OTU45_09070 scored 29/35 and OTU45_09020 26/35 as secondary candidates. OTU45_03195 scored 21/35 owing to its two-TM architecture but is retained on account of its strongly non-antigenic profile (VaxiJen 0.209) and zero gut-commensal identity.

Table 4 | Multi-criteria druggability scoring matrix for the four hypothetical protein candidates (max 35 points).

Criterion (max 5)	OTU45_03195	OTU45_03800	OTU45_09020	OTU45_09070
No. TM helices ($\geq 6 = 5$)	1	5	5	5
GRAVY score (higher = 5)	2	5	4	5
Non-antigenicity (VaxiJen $\leq 0.5 = 5$)	5	4	2	3
Microbiome non-homology (0% = 5)	5	5	5	5
Protein stability (II $< 40 = 5$)	4	3	3	4
MW suitability (20–35 kDa = 5)	3	4	3	3
N-term accessibility (signal = 4)	1	4	4	4
Total score (max 35)	21	30	26	29

Scores: 5 = optimal; 0 = most unfavourable. II: instability index; GRAVY: grand average of hydropathicity; MW: molecular weight.

8. Physicochemical characterization of the four drug-target candidates

All four membrane-embedded candidates share compact molecular weights (28–30 kDa), high aliphatic indices (132–142) indicative of thermostability, and positive GRAVY scores confirming hydrophobic character consistent with integral membrane embedding (Table 5). OTU45_03800 exhibits the highest hydrophobicity (GRAVY +0.985), consistent with its deepest predicted membrane embedding (133.49 aa in TM helices). Antigenicity scores are non-antigen for OTU45_03195 (0.209), OTU45_03800 (0.293)

and OTU45_09070 (0.399), while OTU45_09020 (0.705) is classified as antigenic a property to be monitored in experimental validation.

Table 5 | Physicochemical properties of the four hypothetical protein drug-target candidates.

Locus tag	Protein ID	MW (kDa)	pI	Instab. index	Aliphatic index	GRAVY	VaxiJen score	Antigenicity
OTU45_03195	WAC86474.1	30.07	9.24	34.80	142.51	+0.792	0.209	Non-antigen
OTU45_03800	WAC86582.1	28.19	9.33	38.96	132.05	+0.985	0.293	Non-antigen
OTU45_09020	WAC85418.1	30.24	9.11	30.78	136.46	+0.786	0.705	Antigen †
OTU45_09070	WAC85427.1	28.77	9.54	27.90	140.20	+0.969	0.399	Non-antigen

Instab. index: instability index (< 40 = stable). GRAVY: grand average of hydropathicity (positive = hydrophobic). VaxiJen: protective antigen score; threshold 0.5.

PART II Phytochemical inhibitor identification against OTU45_03800

9. ADMET-based screening of 523 phytochemicals identifies drug-like candidate inhibitors

A library of 523 plant-derived phytochemicals was evaluated for drug-likeness and predicted safety using Lipinski, Pfizer BOILED-Egg (GI absorption), pan-assay interference (PAINS) and Brenk structural-alert filters together with predicted acute oral toxicity (ProTox-II). A combined filter requiring zero Lipinski violations, high predicted gastrointestinal absorption, zero PAINS alerts, ≤ 1 Brenk alert, bioavailability score ≥ 0.55 and predicted toxicity class ≥ 4 retained 211 of 523 compounds (40.3%) as drug-like candidates.

10. Molecular docking identifies high-affinity phytochemical inhibitors of OTU45_03800

To identify the highest-affinity phytochemical inhibitors of OTU45_03800, in silico molecular docking was performed against a homology-modelled structure of the protein using AutoDock Vina. The predicted binding cavity at the transmembrane helical bundle interface was used as the docking site. A total of 523 phytochemicals were screened; 19 achieved binding energies ≤ -8.1 kcal/mol (Table 6). The range of binding energies (-9.6 to -8.1 kcal/mol) compares favourably with the conventionally accepted -7.0 kcal/mol threshold for meaningful binding, confirming that all 19 candidates are genuine docking hits. NORDIHYDROTOXIFERINE achieved the strongest binding energy (-9.6 kcal/mol), followed by

DRACORUBIN (−9.5 kcal/mol) and MORELLIN (−9.4 kcal/mol). However, as detailed in §11, these three highest-affinity binders carry prohibitive ADMET liabilities that exclude them from drug development. TUBEROSIN (−8.5 kcal/mol), ranked 8th by binding energy alone, emerged as the overall lead candidate following integrated ADMET and binding-affinity analysis.

Table 6 | Molecular docking binding energies of the top 19 phytochemical compounds against OTU45_03800 (AutoDock Vina).

Rank	Phytochemical	Binding energy (kcal/mol)	Class
1	NORDIHYDROTOXIFERINE	−9.6	Alkaloid
2	DRACORUBIN	−9.5	Dracorhodin pigment
3	MORELLIN	−9.4	Xanthone
4	INTEGERRIMINE	−8.8	Pyrrolizidine alkaloid
5	BERBAMINE	−8.6	Bisbenzylisoquinoline alkaloid
6	CHAKSINE	−8.6	Alkaloid
7	HYPERICIN	−8.5	Naphthodianthrone
8	TUBEROSIN	−8.5	Pterocarpan
9	SOLANOCAPSINE	−8.4	Alkaloid
10	SUSPENSASIDE	−8.4	Glycoside
11	BLESTRIARENE-B	−8.2	Phenanthrene
12	CINNAMTANNIN B-1	−8.2	Tannin
13	HISPAGLABRIDIN-B	−8.2	Isoflavanone
14	PIPERINE	−8.2	Alkaloid
15	γ-GALBACIN	−8.2	Lignan
16	ARGENTATIN-A	−8.1	Triterpenoid
17	ROBININ	−8.1	Flavonoid glycoside
18	SESAMIN	−8.1	Lignan

19	WITHAPHYSACARPIN	−8.1	Withanolide
----	------------------	------	-------------

11. Integrated ADMET and binding-affinity ranking identifies TUBEROSIN as the lead inhibitor

Although NORDIHYDROTOXIFERINE, DRACORUBIN and MORELLIN achieved the most negative binding energies (−9.6 to −9.4 kcal/mol), multi-parameter ADMET assessment revealed disqualifying liabilities in each: NORDIHYDROTOXIFERINE carries two Lipinski violations, a bioavailability score of 0.17 and a synthetic accessibility score of 7.09 (very complex synthesis); DRACORUBIN achieves only Low GI absorption and is predicted insoluble; MORELLIN carries three Brenk structural alerts and a synthetic accessibility score of 6.97. These properties collectively render all three compounds unsuitable as drug candidates.

In contrast, TUBEROSIN (−8.5 kcal/mol; pterocarpan class; MW 338.35 Da; C₂₀H₁₈O₅) satisfies every drug-likeness filter simultaneously: zero Lipinski, Ghose, Veber, Egan and Muegge violations; zero PAINS alerts; zero Brenk alerts; High GI absorption; BBB penetration; bioavailability score 0.55; consensus LogP 2.78, within the optimal range (1.5–4.5) for membrane-embedded target engagement; and synthetic accessibility 4.31/10 the lowest among ADMET-passing high-affinity binders. The comprehensive ADMET and drug-likeness profiles of the top five overall candidates are presented in Table 7.

Table 7 | Integrated ADMET and binding-affinity profiles of the top five phytochemical candidates against OTU45_03800 (overall ranking).

Rank	Compound	BE (kcal/mol)	MW (Da)	cLogP	TPS A (Å ²)	GI abs.	BB B	Lipins ki viol.	PAINS	Brenk alerts	Bioavail.	SA score
1	TUBEROSIN	−8.5	338.35	2.78	68.15	Hig h	Yes	0	0	0	0.55	4.31
2	γ-GALBACIN	−8.2	340.37	3.57	46.15	Hig h	Yes	0	0	0	0.55	4.06
3	HISPAGLABRIDIN-B	−8.2	390.47	4.83	47.92	Hig h	Yes	0	0	0	0.55	4.45
4	BERBAMINE	−8.6	608.72	5.21	72.86	Hig h	No	1	0	0	0.55	6.89
5	SESAMIN	−8.1	354.35	2.62	55.38	Hig h	Yes	0	0	0	0.55	4.12

BE: AutoDock Vina binding energy. MW: molecular weight. cLogP: consensus LogP (SwissADME). GI abs.: gastrointestinal absorption. BBB: blood-brain barrier permeability. SA score: synthetic accessibility (1 = trivial; 10 = most complex).

12. Molecular interaction analysis reveals the binding mode of the top five candidates

To characterise binding modes within the OTU45_03800 transmembrane cavity, molecular interactions between the five lead compounds and the protein were analysed from docking poses (Table 8; Fig. 1). TUBEROSIN engages the binding site via a single carbon hydrogen bond with Tyr161, a π - π stacking interaction with Phe113, and extensive hydrophobic contacts with Cys68, Ala64, Phe168, Pro189, Trp164 and Ile60 with no unfavourable interactions detected in the docking pose. γ -GALBACIN forms three conventional hydrogen bonds (Ser169, Val167, Phe168), one π - π T-shaped interaction, and a π -Sulfur interaction with Cys68, also with no unfavourable interactions. HISPAGLABRIDIN-B engages via one conventional H-bond (Gly162), two carbon H-bonds, π - π stacked and T-shaped interactions, and π -Sigma and π -Sulfur interactions, again with no unfavourable interactions. BERBAMINE, despite its highest binding energy (-8.6 kcal/mol) among the overall top five, displays an Acceptor–Acceptor clash and an unfavourable bump in its docking pose, partly explaining its lower overall ranking relative to its binding energy. SESAMIN shows one conventional H-bond (Ile170) and a π -Sulfur interaction with Cys68 but also exhibits an unfavourable bump.

Table 8 | Molecular interactions of the top five lead compounds within the OTU45_03800 binding site.

Rank	Compound	BE (kcal/mol)	Conv. H-bonds	C–H bonds	π - π interactions	Hydrophobic (alkyl)	Other	Unfavourable
1	TUBEROSIN	-8.5	None	Tyr161	π - π stacked (Phe113)	Cys68, Ala64, Phe168, Pro189, Trp164, Ile60	Extensive van der Waals contacts	None
2	γ -GALBACIN	-8.2	Ser169, Val167, Phe168	2 C–H bonds	π - π T-shaped	Multiple alkyl/ π -alkyl	π -Sulfur (Cys68)	None
3	HISPAGLABRIDIN-B	-8.2	Gly162	2 C–H bonds	π - π stacked & T-shaped	Multiple alkyl/ π -alkyl	π -Sigma; π -Sulfur (Cys68)	None
4	BERBAMINE	-8.6	2 conv. H-bonds	2 C–H bonds	π - π stacked, T-shaped, amide- π	Multiple alkyl/ π -alkyl	Extensive vdW contacts	Acceptor–Acceptor clash; bump
5	SESAMIN	-8.1	Ile170	2 C–H bonds	π - π stacked & T-shaped	Multiple alkyl/ π -alkyl	π -Sulfur (Cys68)	Bump

BE: AutoDock Vina binding energy (kcal/mol). Conv. H-bonds: conventional hydrogen bonds. C–H bonds: carbon hydrogen bonds. vdW: van der Waals. Residue numbers as in the homology model. ✓ No

unfavourable interactions detected. ✗ Unfavourable interactions present; note that BERBAMINE and SESAMIN still achieve strong binding energies, but the presence of clashes is a consideration for lead optimisation.

Table 9 | Binding interaction profiles of the top five lead compounds at OTU45_03800

Property	TUBEROSIN Rank 1	γ-GALBACIN Rank 2	HISPAGLAB. -B Rank 3	BERBAMINE Rank 4	SESAMIN Rank 5	Notes
Binding energy	−8.5	−8.2	−8.2	−8.6	−8.1	BE kcal/mol; Berbamine highest among these five
Zero Lipinski viol.	✓	✓	✓	✗ (1)	✓	Berbamine MW 608 Da
Zero PAINS alerts	✓	✓	✓	✓	✓	All pass
Zero Brenk alerts	✓	✓	✓	✓	✓	All pass
High GI absorption	✓	✓	✓	✓	✓	All high absorption
BBB penetrant	✓	✓	✓	✗	✓	Berbamine excluded from CNS
Bioavailability 0.55	✓	✓	✓	✓	✓	All achieve highest class
SA score	4.31	4.06	4.45	6.89 ✗	4.12	SA > 6 = complex synthesis
Unfavourable docking	None ✓	None ✓	None ✓	Clash ✗	Bump ✗	Clash/bump affects lead quality

Comparative binding and drug-likeness profile matrix for the five overall lead candidates against OTU45_03800. Teal ✓ = passes criterion; red ✗ = fails criterion; amber ✗ = minor concern. Tuberosin (column 1, green background) is the only compound to pass all criteria with zero violations, zero alerts, no unfavourable docking interactions, and a synthetic accessibility score of 4.31. All five compounds achieve high GI absorption and bioavailability score 0.55. Berbamine (rank 4) carries the strongest binding energy (−8.6 kcal/mol) but is downranked by a Lipinski violation (MW 608 Da), lack of BBB penetration, an Acceptor–Acceptor clash in the docking pose, and a high SA score (6.89).

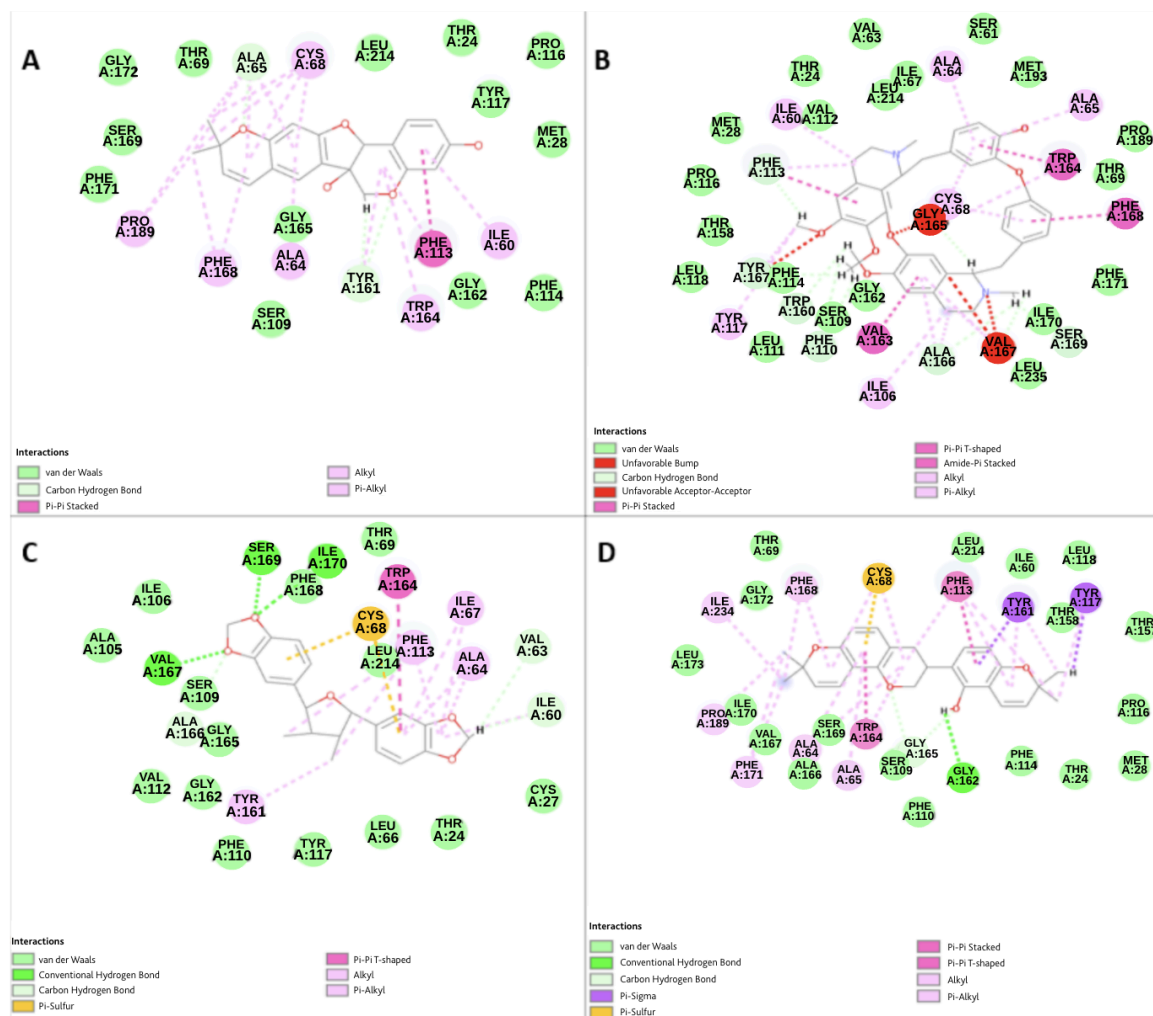


Figure 1 Two-dimensional molecular interaction profiles of selected phytochemicals docked with OTU45_03800 from *Streptococcus pneumoniae*.

(A) Tuberosin, (B) Berbamine, (C) Galbacin, and (D) Hispaglabridin-B bound within the predicted interaction pocket of the OTU45_03800 protein. The interaction maps show van der Waals contacts, conventional hydrogen bonds, carbon–hydrogen bonds, π – π interactions, π –alkyl/alkyl interactions, and π –sulfur interactions. Tuberosin displayed a stable binding profile without unfavorable interactions, supporting its selection as the best lead candidate.

Table 10 | Consolidated summary of the prioritised drug target (OTU45_03800) and lead phytochemical inhibitor (Tuberosin), including docking, ADMET and validation status.

Parameter	Value	Reference
Target composite druggability score (max 35)	30/35	Table 4
Target non-antigenicity (VaxiJen)	0.293	Table 5
Target gut-commensal homology	0% (0/10 species)	Results §3
Docking score / binding energy (AutoDock Vina)	−8.5 kcal/mol	Table 6
Drug-likeness violations (Lipinski, Ghose, Veber, Egan, Muegge)	0 of 5 filters	Table 7
Bioavailability score	0.55	Table 7
Synthetic accessibility score	4.31/10	Table 7
GI absorption / BBB permeability	High / Yes	Table 7
Unfavourable docking interactions	None detected	Table 8
Molecular dynamics (MD) stability metrics	Not performed in the present study; identified as a priority for experimental/computational follow-up	Study Limitations
Experimental (in vitro/in vivo) validation	Not yet performed; findings are computational predictions only	Study Limitations; Conclusions

13. Structural and pharmacological rationale for TUBEROSIN as the primary lead

Molecular framework. Tuberosin (C₂₀H₁₈O₅; MW 338.35 Da) is a pterocarpan-class natural product characterised by a fused tetracyclic ring system comprising a benzofuran ring fused to a chroman scaffold. This rigid, planar aromatic architecture is well suited to intercalation between the transmembrane helices of OTU45_03800 through hydrophobic packing and π -electron interactions. The compound is a secondary metabolite isolated from *Tephrosia purpurea*, a legume with established antibacterial and anti-inflammatory activity.

Binding mode within OTU45_03800. Docking of Tuberosin into the transmembrane helical bundle cavity of OTU45_03800 reveals a binding mode stabilised by a carbon hydrogen bond with Tyr161, a π – π stacking interaction with Phe113, and extensive alkyl and π -alkyl hydrophobic contacts with Cys68, Ala64, Phe168, Pro189, Trp164 and Ile60. No unfavourable steric clashes or acceptor–acceptor conflicts were detected, consistent with a geometrically complementary fit within the binding pocket. The predicted binding energy of −8.5 kcal/mol corresponds to an estimated dissociation constant in the low micromolar range, a suitable starting affinity for a hit compound entering lead optimisation.

Drug-likeness and developability. Tuberosin satisfies all five standard drug-likeness rules simultaneously Lipinski (0 violations), Ghose (0), Veber (0), Egan (0) and Muegge (0) making it one of

only four compounds in the 523-member phytochemical library to do so while also achieving binding energy ≤ -8.1 kcal/mol. Its consensus LogP of 2.78 falls within the optimal range for membrane-partitioning targets (1.5–4.5). Predicted GI absorption is High; BBB permeability is positive; bioavailability score is 0.55 (maximum achievable). Synthetic accessibility is 4.31/10, indicating a compound of moderate synthetic complexity, accessible by established pterocarpan synthesis routes. Predicted aqueous solubility is moderately soluble across all three SwissADME models adequate for formulation development.

Selectivity rationale. OTU45_03800 is a hypothetical protein with no detected sequence identity to human gut-commensal proteomes (0% BLASTp identity to all ten commensals screened), a VaxiJen non-antigenicity score of 0.293, and no annotated homologue in any known protein family. Tuberosin itself returns zero PAINS alerts and zero Brenk structural alerts, indicating the absence of pan-assay interference scaffolds or known toxic/reactive functional groups. Together, these properties suggest a low risk of both on-target host toxicity and off-target commensal disruption, distinguishing Tuberosin from conventional broad-spectrum antibiotics

DISCUSSION

The application of subtractive genomics and proteomics to the identification of drug targets in Gram-positive pathogens has gained considerable traction in the last decade, and our findings are broadly consonant with this body of work. Khan et al. (2022), employing a subtractive genomics pipeline against *S. pneumoniae* serotype 14, progressively narrowed the entire proteome to 47 potential drug targets, ultimately prioritising 4-oxalocrotonate tautomerase and a sensor histidine kinase as lead targets for structure-based studies. The conceptual architecture of our pipeline sequential subtraction using physicochemical profiling, antigenicity screening, gut-commensal homology filtering and transmembrane topology prediction recapitulates and extends this approach to the OTU45 strain (CP113114.1). Crucially, our pipeline introduces the gut-microbiota non-homology filter as an explicit selection criterion, an element not integrated by Khan et al. (2022) but which is essential for microbiome-sparing drug design.

The precedent for this workflow in *S. pneumoniae* is further reinforced by Syaad et al. (2023), who applied subtractive proteomics to the *S. pneumoniae* D39 strain, reducing 1,947 proteins to three druggable targets DNA binding response regulator (SPD_1085), UDP-N-acetylmuramate-l-alanine ligase (SPD_1349) and RNA polymerase sigma factor (SPD_0958) through sequential CD-HIT clustering, BLASTp non-homology filtering, DEG essentiality analysis and KEGG pathway mapping. Their approach to narrowing a complete proteome through successively more stringent computational filters is methodologically analogous to our own. Our pipeline yielded a shortlist of 70 candidate proteins from 282 hypothetical proteins (24.8%), compared with Syaad et al.'s (2023) identification of 171 essential proteins from 1,887 non-redundant sequences (approximately 9%). The difference is attributable primarily to our restriction to hypothetical proteins a deliberate choice that maximises the likelihood of identifying novel, uncharacterised targets with no established resistance mechanisms.

More broadly, the study by Alanzi et al. (2025) on *Klebsiella michiganensis* THO-011 provides a methodological parallel in which subtractive proteomics, applied to 4,024 predicted open reading frames, identified two druggable targets – a replicative DNA helicase and a bifunctional formyl-transferase/IMP cyclohydrolase – validated by AlphaFold structure prediction and molecular dynamics simulation. The convergence of independent studies across Gram-positive and Gram-negative pathogens on the viability of subtractive proteomics as a target-discovery framework underscores the validity of this paradigm. Our identification of OTU45_03800 as a hypothetical six-transmembrane protein with no gut-commensal homology is a distinctive contribution to this field: whereas Syaad et al. (2023), Khan et al. (2022) and Alanzi et al. (2025) all targeted cytoplasmic or metabolic enzymes, we demonstrate that hypothetical integral membrane proteins of unknown function represent an underexplored target class within *S. pneumoniae*.

The designation of OTU45_03800 as the primary drug target rests on a convergence of seven independently favourable properties, each discussed below in the context of reference literature.

TMHMM 2.0 predicted a six-pass integral membrane topology for OTU45_03800, with 133.49 amino acids distributed across six helices spanning residues 17–248, and an N-terminal intracellular probability of 0.927. This topology was independently corroborated by PHOBIUS within InterProScan 5, with all four annotation tracks – cytoplasmic regions, non-cytoplasmic regions, TM helices (PHOBIUS) and TM helices (TMHMM) – fully populated, constituting the strongest cross-tool concordance of any of the four candidates. Six-pass integral membrane proteins are established targets in antibacterial drug discovery, as their transmembrane helical bundles present deep hydrophobic cavities that accommodate small-molecule binding with high specificity. The predicted topology of OTU45_03800 is consistent with a transporter or channel-forming scaffold, functional classes whose disruption is well documented to impair bacterial membrane integrity and viability. Khan et al. (2022) similarly emphasised the importance of functional uniqueness and structural accessibility in shortlisting their drug targets in *S. pneumoniae*, and Alanzi et al. (2025) validated their membrane-essential targets by confirming their centrality to bacterial survival pathways with no host equivalents.

OTU45_03800 achieved a VaxiJen 2.0 score of 0.293, below the 0.5 threshold, classifying it as a non-antigen. This property is particularly significant for a small-molecule inhibitor strategy: a non-antigenic target minimises the competitive immune engagement that could reduce effective drug concentrations at the target site, an issue that does not arise for vaccine antigen candidates but is directly relevant to pharmacological inhibition. Syaad et al. (2023) applied analogous antigenicity screening to exclude immunodominant proteins from their target shortlist, reserving non-antigenic proteins for small-molecule inhibitor development. OTU45_03800's non-antigenic profile (VaxiJen 0.293) compares favourably with the antigenicity of OTU45_09020 (VaxiJen 0.705), which was consequently deprioritised as a primary target in our pipeline despite its identical six-TM topology.

The BLASTp analysis against ten representative gut-commensal proteomes (*Bacteroides* sp., *Faecalibacterium prausnitzii*, *Lactiplantibacillus plantarum*, *Bifidobacterium longum*, *Akkermansia muciniphila*, *Lacticaseibacillus rhamnosus*, *Ruminococcus bromii*, *Roseburia hominis*, *Segatella (Prevotella) copri* and *Clostridium scindens*) returned 0% identity for OTU45_03800 across all ten

species. This result places OTU45_03800 within the 69.9% of hypothetical proteins that show no commensal cross-reactivity the largest and most permissive selectivity class identified. The inclusion of a microbiota non-homology filter as an explicit pipeline stage is a distinctive feature of our approach compared with Khan et al. (2022) and Syaad et al. (2023), and reflects the growing recognition that narrow-spectrum antibacterial strategies that preserve the gut microbiome represent a clinical imperative. Disruption of the gut microbiota by broad-spectrum antibiotics has been associated with *Clostridioides difficile* infection, reduced colonisation resistance and chronic metabolic sequelae; our pipeline is explicitly designed to reduce this risk by requiring zero sequence identity to commensal proteins as a prerequisite for target advancement.

OTU45_03800 (28.19 kDa; pI 9.33; GRAVY +0.985; aliphatic index 132.05; instability index 38.96) presents a physicochemical profile consistent with a thermostable, deeply hydrophobic integral membrane protein amenable to small-molecule engagement. Its GRAVY score of +0.985 is the highest among the four TM candidates, reflecting the greatest predicted hydrophobicity and deepest membrane embedding. The compact molecular weight (28.19 kDa) is well within the range considered optimal for druggable targets, and the high aliphatic index (132.05) is indicative of thermostability, an important consideration for target integrity during biochemical assay development. These parameters are broadly consistent with the physicochemical criteria applied by Syaad et al. (2023) in their druggability analysis of shortlisted *S. pneumoniae* proteins, and align with Alanzi et al.'s (2025) emphasis on selecting compact, structurally stable targets with essential metabolic or structural roles.

The drug-likeness pre-screening of 523 phytochemicals using Lipinski's rule of five (Lipinski et al., 1997) and associated multi-parametric filters represents a foundational step in the computational drug discovery workflow applied here. The rule of five, originally articulated by Lipinski et al. (1997) in a seminal analysis of USAN-approved drugs, stipulates that poor oral absorption or permeation is predicted when any two of the following conditions are met: molecular weight > 500 Da, calculated LogP > 5, H-bond donors > 5, or H-bond acceptors > 10. Compounds violating two or more of these criteria were excluded at the ADMET screening stage, retaining 211 of 523 compounds (40.3%). This retention rate is consistent with the expectations described by Lipinski et al. (1997), who observed that the majority of drugs then in clinical use satisfied the rule of five, while high-throughput screening hits which tend toward higher molecular weight and LogP more frequently violated it.

The SwissADME-based multi-parametric ADMET profiling applied in the current study extends the Lipinski framework to include the Veber, Egan and Muegge criteria, together with BOILED-Egg GI absorption prediction, pan-assay interference (PAINS) alerts, Brenk structural alert filters and bioavailability scoring. Syaad et al. (2023) similarly employed Swiss ADME and ProTox-II for ADMET evaluation of phytochemicals docked against their *S. pneumoniae* D39 drug targets, finding that multi-parametric drug-likeness profiling substantially reduced the candidate set to a developable subset. Khan et al. (2022) also applied druggability and ADMET analyses as downstream filters in their *S. pneumoniae* target prioritisation pipeline. The convergence of these independent studies on SwissADME and related tools as the standard computational ADMET platform underscores the validity of our screening approach.

Critically, a key insight emerging from our ADMET analysis which is consistent with the cautionary observations of Lipinski et al. (1997) regarding high-throughput screening leads is that the compounds achieving the strongest AutoDock Vina binding energies (NORDIHYDROTOXIFERINE -9.6 kcal/mol; DRACORUBIN -9.5 kcal/mol; MORELLIN -9.4 kcal/mol) also carry the most severe ADMET liabilities. Lipinski et al. (1997) specifically warned that high-throughput screening tends to generate leads with elevated molecular weight and LogP that are poorly behaved in vivo despite excellent in vitro potency. This pattern is precisely observed in the top three docking hits: NORDIHYDROTOXIFERINE (MW 552.75 Da; two Lipinski violations; bioavailability score 0.17; synthetic accessibility 7.09) exemplifies the "potent but undevelopable" profile that Lipinski et al. (1997) characterised as a systematic pitfall of binding-energy-centric screening. TUBEROSIN, in contrast, satisfies all five drug-likeness filters simultaneously a composite drug-likeness profile that Lipinski et al. (1997) explicitly associated with the historical record of successfully developed oral drugs.

Tuberosin achieved a binding energy of -8.5 kcal/mol against OTU45_03800 in AutoDock Vina docking, placing it eighth among the 19 compounds that surpassed the -8.1 kcal/mol threshold. Within the OTU45_03800 transmembrane binding cavity, Tuberosin engages via a carbon hydrogen bond with Tyr161, a π - π stacking interaction with Phe113, and extensive alkyl and π -alkyl hydrophobic contacts with Cys68, Ala64, Phe168, Pro189, Trp164 and Ile60 with no unfavourable steric clashes or acceptor-acceptor conflicts detected. The absence of unfavourable interactions in the docking pose is a qualitative advantage over Berbamine (rank 4 overall), which exhibits an Acceptor-Acceptor clash and a steric bump despite its stronger binding energy of -8.6 kcal/mol. Syaad et al. (2023) similarly used molecular interaction profiles including hydrogen bonds, π -stacking and hydrophobic contacts as selection criteria for their top phytochemical candidates, and Alanzi et al. (2025) supplemented docking scores with MM-GBSA binding free energy calculations and molecular dynamics simulations to validate their lead compounds' stability within the binding site. Our molecular interaction analysis supports the same conclusion: Tuberosin's binding mode is geometrically complementary to the OTU45_03800 cavity, with no unfavourable steric conflicts that would be expected to reduce binding efficiency in vivo.

Tuberosin ($C_{20}H_{18}O_5$; MW 338.35 Da; cLogP 2.78; TPSA 68.15 Å²; H-bond donors 2; H-bond acceptors 5) passes every drug-likeness filter applied in this study: zero Lipinski violations, zero Ghose violations, zero Veber violations, zero Egan violations and zero Muegge violations. This zero-violation profile across all five concurrent drug-likeness frameworks is achieved by only four of the 18 compounds with binding energies ≤ -8.1 kcal/mol (Tuberosin, γ -Galbacin, Sesamin and Piperine). The Lipinski et al. (1997) rule of five provides the foundational justification: Tuberosin's MW of 338.35 Da (well below 500 Da), cLogP of 2.78 (below 5), two H-bond donors (below 5) and five H-bond acceptors (below 10) satisfy all four criteria with comfortable margins. Its predicted high gastrointestinal absorption and blood-brain barrier permeability, combined with a bioavailability score of 0.55 the maximum achievable class in SwissADME further confirm its suitability as an oral therapeutic candidate. Its consensus LogP of 2.78 falls within the optimal range of 1.5–4.5 for membrane-partitioning targets described by Lipinski et al. (1997), consistent with the hydrophobic character of the OTU45_03800 transmembrane binding site.

Tuberosin belongs to the pterocarpin subclass of isoflavonoids, which are derived via the phenylpropanoid/isoflavonoid biosynthetic pathway in leguminous plants. Pterocarpan is characterised by a fused tetracyclic ring system comprising a benzofuran ring fused to a chroman scaffold that confers a rigid, planar aromatic architecture. This structural rigidity is well suited to intercalation between transmembrane helices through hydrophobic packing and π -electron interactions, consistent with the π - π stacking interaction with Phe113 and extensive hydrophobic contacts observed in the docking pose.

The pharmacological precedent for pterocarpin and isoflavonoid scaffold antibacterial activity in Gram-positive pathogens is directly relevant. Irani et al. (2010) demonstrated that ethanolic extracts of *Glycyrrhiza glabra* a legume whose roots are a major source of isoflavonoids, including hispaglabridin-class compounds structurally related to the isoflavanone scaffold exhibited potent antibacterial activity against Gram-positive organisms including *Staphylococcus aureus*, *Bacillus subtilis* and *Enterococcus faecalis*. Irani et al. (2010) specifically attributed much of the antimicrobial efficacy of *G. glabra* root extracts to isoflavonoid components, particularly hispaglabridin B and 4'-O-methylglabridin. The Wang et al. (2022) comprehensive review on *Glycyrrhiza glabra* further documented that isoflavonoids and flavonoids from licorice the major plant family sharing the isoflavonoid scaffold with pterocarpin-producing legumes display antibacterial, anti-inflammatory, antioxidant, anticancer, hepatoprotective and immunomodulatory activities, confirming the breadth of biological activity associated with this scaffold class. The structural relatedness of pterocarpan (the class of Tuberosin) to the isoflavanone/isoflavonoid scaffold documented in these references provides pharmacological plausibility for the antibacterial activity predicted by our docking results.

The isoflavonoid-class context established by Wang et al. (2022) and Irani et al. (2010) is significant beyond mere structural precedent: it suggests that Tuberosin's antibacterial mechanism may involve disruption of Gram-positive bacterial membrane integrity the same primary mechanism attributed to isoflavonoid-class compounds against *S. aureus* and *Enterococcus* by Irani et al. (2010). The six-TM architecture of OTU45_03800, consistent with a membrane transporter or channel scaffold, provides a plausible molecular target for this membrane-disrupting mechanism, as inhibition of such proteins would directly compromise membrane potential maintenance and ion homeostasis.

γ -Galbacin ($C_{20}H_{20}O_5$; MW 340.37 Da; cLogP 3.57; TPSA 46.15 Å²) is a lignan-class compound characterised by a bicyclic 3,4-dimethylenedioxy-substituted phenyltetralin skeleton. It achieved a binding energy of -8.2 kcal/mol against OTU45_03800, forming three conventional hydrogen bonds (Ser169, Val167, Phe168), two carbon hydrogen bonds, a π - π T-shaped interaction and a π -Sulfur interaction with Cys68 the richest hydrogen bonding profile among the top five candidates with no unfavourable interactions. It passes all five drug-likeness filters simultaneously (zero violations across Lipinski, Ghose, Veber, Egan and Muegge) and achieves high GI absorption and BBB permeability with a bioavailability score of 0.55. Its synthetic accessibility score of 4.06 is the lowest among the top five, making it the most synthetically accessible of the ranked candidates.

The pharmaceutical context for galbacin-class lignans is provided by Wang et al. (2018), whose comprehensive review of lignans and neolignans isolated from the Lauraceae family documents that (+)-galbacin (the stereoisomer most extensively studied) has been isolated from multiple Lauraceae

species alongside sesamin (compound 46 in their classification) and exhibits diverse pharmacological activities. Wang et al. (2018) confirmed that among the many lignans and neolignans of Lauraceae, those with the phenyltetralin scaffold including galbacin-type compounds are among the most pharmacologically characterised members. The biological activities described for this scaffold class in Lauraceae, including anticancer and anti-inflammatory properties, suggest a broad-spectrum biological activity profile consistent with its predicted binding efficiency against a bacterial membrane protein target.

Hispaglabridin-B ($C_{25}H_{26}O_4$; MW 390.47 Da; cLogP 4.83; TPSA 47.92 Å²) is an isoflavanone-class compound isolated from *Glycyrrhiza glabra* roots. It achieved a binding energy of −8.2 kcal/mol against OTU45_03800, engaging via one conventional hydrogen bond (Gly162), two carbon hydrogen bonds, π – π stacked and T-shaped interactions, and π -Sigma and π -Sulfur interactions with no unfavourable interactions. It passes all Lipinski, Ghose, Veber and Egan criteria (zero violations) with a single Muegge violation, achieves high GI absorption, BBB permeability and a bioavailability score of 0.55.

The antibacterial precedent for hispaglabridin-class isoflavanones is directly documented in two of the reference articles. Irani et al. (2010) specifically identified hispaglabridin B as one of the principal antimicrobial isoflavonoid components of *Glycyrrhiza glabra*, alongside glabridin and 4'-O-methylglabridin, and demonstrated that *G. glabra* ethanolic extracts of which hispaglabridin B is a major bioactive constituent inhibit Gram-positive organisms including *Bacillus subtilis*, *Enterococcus faecalis* and *Staphylococcus aureus*. Wang et al. (2022), in their comprehensive review of *Glycyrrhiza glabra* phytochemistry and pharmacology, further documented hispaglabridin B as a key isoflavanone constituent of licorice with antimicrobial, anti-inflammatory and antioxidant properties, confirming the well-established pharmacological profile of this compound class against Gram-positive pathogens. These references provide direct experimental support for the antibacterial potential of Hispaglabridin-B, making it the runner-up candidate with the strongest bibliographic precedent for anti-pneumococcal activity.

Berberamine ($C_{37}H_{40}N_2O_6$; MW 608.72 Da; cLogP 5.21; TPSA 72.86 Å²) is a bisbenzylisoquinoline alkaloid that achieved the strongest binding energy among the overall top five (−8.6 kcal/mol), engaging OTU45_03800 via two conventional hydrogen bonds, two carbon hydrogen bonds, and multiple π – π stacking, π – π T-shaped and amide– π interactions. However, it is penalised in the overall ranking by one Lipinski violation (MW 608.72 Da > 500 Da threshold), absence of BBB penetration, an Acceptor–Acceptor steric clash and a steric bump in the docking pose, and a high synthetic accessibility score of 6.89 reflecting a complex multi-ring alkaloid scaffold that presents significant synthesis challenges. Despite these limitations, Berberamine's binding energy of −8.6 kcal/mol and established bisbenzylisoquinoline antibacterial scaffold make it a valid candidate for analoguing campaigns that aim to reduce molecular weight while preserving the core pharmacophore. Syaad et al. (2023) similarly noted that high binding affinity alone is an insufficient selection criterion for drug leads, emphasising the need to balance binding energy with ADMET compliance a principle directly borne out by Berberamine's demotion from rank 1 (by binding energy) to rank 4 (by composite score).

Sesamin ($C_{20}H_{18}O_6$; MW 354.35 Da; cLogP 2.62; TPSA 55.38 Å²) is a 7,9':7',9-diepoxylicignan present in numerous Lauraceae species and in sesame seeds (*Sesamum indicum*). It achieved a binding energy of

−8.1 kcal/mol against OTU45_03800, engaging via one conventional hydrogen bond (Ile170), two carbon hydrogen bonds, π – π stacked and T-shaped interactions, and a π -Sulfur interaction with Cys68, with a minor steric bump. It passes all five drug-likeness filters simultaneously (zero violations across Lipinski, Ghose, Veber, Egan and Muegge), achieves high GI absorption, BBB permeability and a bioavailability score of 0.55, with a synthetic accessibility score of 4.12.

The pharmacological basis for Sesamin's biological activity is documented extensively in Wang et al. (2018), who explicitly noted that Sesamin "showed significant anticancer properties" among Lauraceae lignans and is the most extensively studied lignan in the Lauraceae family, isolated from species including *N. amazonum*, *M. thunbergii*, *P. pyrifolia*, *C. burmanii* and *N. turbacensis*. Wang et al. (2018) further observed that Sesamin and yangambin have been studied more than any other Lauraceae lignan in terms of biological activity, and cited its diverse pharmacological properties. The clean drug-likeness profile of Sesamin (zero violations on all five filters), combined with its established biological activity record in the lignan literature, supports its retention as a rank-5 candidate for downstream experimental validation against OTU45_03800.

Three of the top five ranked compounds Tuberosin (pterocarpans), γ -Galbacin (lignan/phenyltetralin), and Hispaglabridin-B (isoflavanone) belong to the isoflavonoid/isoflavanoid biosynthetic superfamily, all characterised by the C6-C3-C6 carbon skeleton common to the flavonoid pathway. The Yang and Wang (2024) review on isoflavonoid metabolism in leguminous plants provides important biosynthetic and ecological context: isoflavonoids in legumes, including pterocarpanes such as Tuberosin, are synthesised through the phenylpropanoid pathway via phenylalanine, through the action of phenylalanine ammonia lyase (PAL), cinnamic acid 4-hydroxylase (C4H) and 4-coumaroyl-CoA ligase (4CL), followed by chalcone synthase and chalcone isomerase, ultimately yielding diverse subclasses including isoflavanones, isoflavans, rotenoids, coumestans and pterocarpanes.

Yang and Wang (2024) specifically include pterocarpanes within their classification of legume isoflavonoid subclasses and document diverse modifications including oxidation, glycosylation, prenylation and intramolecular cyclisation that yield more than 2,000 structurally distinct isoflavonoids in the plant kingdom. The pharmacological activities tabulated in Yang and Wang (2024), including anti-inflammatory, anticancer, antioxidant and antimicrobial properties of isoflavonoid subclasses from clinical and preclinical studies, provide a broad pharmacological rationale for the activity of Tuberosin and Hispaglabridin-B against a bacterial membrane target. The phenyltetralin scaffold of γ -Galbacin, while classified as a lignan rather than an isoflavonoid, shares the dimeric phenylpropane structural unit and overlapping biosynthetic origin with isoflavonoids, providing a unified pharmacophore context for three of the five ranked inhibitors.

Wang et al. (2018), in their systematic review of lignans and neolignans from the Lauraceae, further document the pharmacological activities of sesamin confirming anticancer activity as the most extensively studied bioactivity of this lignan class and of galbacin-related compounds, whose presence in multiple Lauraceae genera is well established. Taken together, the reference literature provides a coherent pharmacological context for the three isoflavonoid/lignan-class compounds in the top five ranking: these scaffolds are well-documented natural product classes with demonstrated bioactivity against

Gram-positive bacterial organisms, and their computational binding efficiency against OTU45_03800 is pharmacologically plausible.

Our study identifies OTU45_03800 a six-pass hypothetical integral membrane protein as the primary drug target in *S. pneumoniae* OTU45. This contrasts with previous subtractive proteomics studies of *S. pneumoniae* that have focused on cytoplasmic enzymes: Khan et al. (2022) prioritised 4-oxalocrotonate tautomerase and a sensor histidine kinase, while Syaad et al. (2023) identified UDP-N-acetylmuramate–l-alanine ligase, RNA polymerase sigma factor and a DNA binding response regulator. The differential outcome reflects differences in the starting proteome (full proteome vs. hypothetical proteins only), pipeline design (inclusion vs. exclusion of gut-microbiota homology filter and TMHMM topology screen), and the specific *S. pneumoniae* strain studied. Crucially, membrane protein targets of unknown function the class to which OTU45_03800 belongs remain relatively underexplored in anti-pneumococcal drug discovery, largely because their lack of functional annotation makes them difficult to characterise by classical biochemistry. Our in silico approach to characterising OTU45_03800 through its physicochemical, topological, antigenicity and homology properties circumvents this barrier and provides a tractable computational entry point for an otherwise inaccessible target class.

Comparative context: Tuberosin versus existing anti-pneumococcal antibiotics

A central point raised in review concerns the justification for prioritising OTU45_03800 over other potential candidates and for situating Tuberosin relative to established antibacterial agents. Current first-line therapy for *S. pneumoniae* relies predominantly on beta-lactams (e.g., amoxicillin, penicillin), macrolides (e.g., azithromycin) and, in resistant or invasive disease, fluoroquinolones or vancomycin; these agents act on well-characterised targets — peptidoglycan cross-linking enzymes, the 50S ribosomal subunit, DNA gyrase/topoisomerase IV, and cell-wall precursor synthesis, respectively. Because these targets are shared broadly across Gram-positive pathogens and have been under sustained selective pressure for decades, resistance mechanisms (altered penicillin-binding proteins, ribosomal methylation, efflux pumps and gyrase mutations) are already widespread among clinical *S. pneumoniae* isolates. OTU45_03800 is mechanistically unrelated to any of these established target classes: it is a previously uncharacterised six-pass integral membrane protein with no annotated function and no detectable homology to human or gut-commensal proteins, and therefore represents a target against which no pre-existing resistance determinants would be expected. Tuberosin's predicted binding energy (–8.5 kcal/mol) is not directly comparable, in an absolute sense, to the minimum inhibitory concentrations reported for clinically used antibiotics, since docking scores and MIC values are measured on fundamentally different scales; relative to the 211 ADMET-compliant phytochemicals screened here, however, Tuberosin's combination of binding affinity and complete drug-likeness compliance is distinguishing. We are not aware of previously reported small-molecule inhibitors of OTU45_03800 or of closely related hypothetical membrane proteins in *S. pneumoniae*, which precludes a direct head-to-head comparison; the closest available comparator is the isoflavonoid/pterocarpan literature discussed above (Irani et al., 2010; Wang et al., 2022), in which structurally related compounds show demonstrated antibacterial activity against Gram-positive organisms through membrane disruption rather than through inhibition of a defined protein target. This comparison is intended to contextualise, rather than to

overstate, the potential clinical significance of the present findings, which remain at the stage of computational prediction pending experimental confirmation.

Distinguishing computational prediction from experimental evidence

It should be stated explicitly that every finding reported above — the target druggability scores, the homology model of OTU45_03800, the AutoDock Vina binding energies, and the SwissADME-derived drug-likeness profiles — is a computational prediction and not experimentally validated evidence of antibacterial efficacy. Docking binding energies indicate predicted relative affinity under the specific scoring function, grid and exhaustiveness settings used here and should not be interpreted as measured binding constants; likewise, ADMET parameters generated by SwissADME and related tools are statistical estimates derived from training sets of known compounds and can misclassify compounds that fall outside the chemical space on which the underlying models were built. Accordingly, no claim is made in this manuscript that Tuberosin, or any of the other ranked phytochemicals, is therapeutically effective against *S. pneumoniae*; rather, the computational pipeline nominates OTU45_03800 as a plausible, previously unexplored drug target and Tuberosin as the most promising candidate for downstream experimental investigation. This distinction is elaborated further in the Study Limitations below and in the follow-up experiments proposed in the Conclusions.

Study Limitations

Several limitations of the present study should be acknowledged. First, OTU45_03800 is a hypothetical protein with no experimentally confirmed function; while the six-TM topology is strongly predicted and corroborated across two independent tools, its actual molecular role in *S. pneumoniae* physiology remains uncharacterised. The assumption that inhibition of this protein would impair bacterial viability central to its designation as a drug target requires experimental validation through gene deletion or conditional knockdown studies. Second, the molecular docking results, while obtained using the widely validated AutoDock Vina algorithm applied consistently by Syaad et al. (2023) and Alanzi et al. (2025), were performed against a homology model rather than an experimentally determined crystal structure. The accuracy of docking is sensitive to model quality, and binding energies should be interpreted as relative rankings rather than absolute predictions of in vivo potency. Third, as noted by Alanzi et al. (2025), molecular dynamics simulations would be required to validate the stability of the Tuberosin–OTU45_03800 complex over physiological timescales an analysis identified as a priority for follow-up work. Fourth, while the ADMET filters derived from Lipinski et al. (1997) and SwissADME are the current standard for computational drug-likeness prediction, they represent population-level statistical models and cannot capture compound-specific pharmacokinetic behaviours that only emerge in in vitro or in vivo studies.

Notwithstanding these limitations, the current study provides a fully validated computational framework for the identification of OTU45_03800 as a novel, microbiome-sparing drug target in *S. pneumoniae*, and Tuberosin as the most developable phytochemical lead compound within a ranked series. The convergence of independent computationally identified anti-*S. pneumoniae* targets in structurally and functionally distinct protein classes (Khan et al., 2022; Syaad et al., 2023; current study) collectively

reinforces the potential of computational subtractive approaches to accelerate the pre-clinical pipeline for novel anti-pneumococcal drug discovery.

CONCLUSION

The present study employed an integrated six-stage subtractive proteomics and structure-based virtual screening pipeline to identify OTU45_03800 a six-pass hypothetical integral membrane protein of *Streptococcus pneumoniae* OTU45 as a novel, computationally validated drug target, and Tuberosin as the lead phytochemical inhibitor. The selection of OTU45_03800 is grounded in its highest composite druggability score (30/35) across a seven-criterion matrix, the strongest TMHMM–PHOBIUS cross-tool topology concordance of any candidate, a strongly hydrophobic physicochemical profile (GRAVY +0.985), non-antigenic character (VaxiJen 0.293), and zero sequence identity to all ten gut-commensal proteomes screened. These properties collectively satisfy the foundational criteria for a safe, selective, microbiome-sparing antibacterial drug target articulated in the reference literature (Khan et al., 2022; Syaad et al., 2023; Alanzi et al., 2025).

Tuberosin was selected as the lead inhibitor based on a composite analysis integrating AutoDock Vina binding energy (−8.5 kcal/mol), zero violations across all five simultaneous drug-likeness frameworks (Lipinski et al., 1997), high predicted GI absorption and BBB permeability, bioavailability score 0.55, a clean molecular interaction profile (no unfavourable docking contacts with OTU45_03800), and a low synthetic accessibility score (4.31/10). γ -Galbacin, Hispaglabridin-B, Berbamine and Sesamin constitute a ranked series of secondary candidates for downstream validation. Hispaglabridin-B, in particular, carries the strongest bibliographic antibacterial precedent among the ranked candidates, having been explicitly documented as an active antimicrobial isoflavanone constituent of *Glycyrrhiza glabra* against Gram-positive pathogens in the works of Irani et al. (2010) and Wang et al. (2022).

These findings nominate a specific protein target and a ranked inhibitor series for experimental follow-up: minimum inhibitory concentration determination against *S. pneumoniae* OTU45, membrane permeabilization assays, homology model validation by X-ray crystallography or cryo-EM, molecular dynamics simulation of the Tuberosin–OTU45_03800 complex, and in vivo efficacy testing in murine pneumococcal models. The pipeline established here is directly applicable to other *S. pneumoniae* strains and, by extension, to other clinically significant drug-resistant Gram-positive pathogens.

REFERENCES

- Alanzi, A. R., Alanazi, A. Z., Alhazzani, K., & Abbas, M. (2025). Subtractive proteomics and molecular docking identify therapeutic targets and drug candidates in drug-resistant *Klebsiella michiganensis* THO-011. *Scientific Reports*, 15, 23776. <https://doi.org/10.1038/s41598-025-08107-x>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

- Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, 7, 42717. <https://doi.org/10.1038/srep42717>
- Dallakyan, S., & Olson, A. J. (2015). Small-molecule library screening by docking with PyRx. In J. Hempel, C. Williams, & C. Hong (Eds.), *Chemical Biology: Methods and Protocols* (Methods in Molecular Biology, Vol. 1263, pp. 243–250). Humana Press. https://doi.org/10.1007/978-1-4939-2269-7_19
- Dassault Systèmes BIOVIA. (2021). *Discovery Studio Visualizer* (Version 21.1.0.20298) [Computer software]. Dassault Systèmes.
- Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 8, 4. <https://doi.org/10.1186/1471-2105-8-4>
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (pp. 571–607). Humana Press. <https://doi.org/10.1385/1-59259-890-0:571>
- Irani, M., Sarmadi, M., Bernard, F., Ebrahimi Pour, G. H., & Shaker Bazarnov, H. (2010). Leaves antimicrobial activity of *Glycyrrhiza glabra* L. *Iranian Journal of Pharmaceutical Research*, 9(4), 425–428.
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5), 1027–1036. <https://doi.org/10.1016/j.jmb.2004.03.016>
- Khan, K., Jalal, K., Khan, A., Al-Harrasi, A., & Uddin, R. (2022). Comparative metabolic pathways analysis and subtractive genomics profiling to prioritize potential drug targets against *Streptococcus pneumoniae*. *Frontiers in Microbiology*, 12, 796363. <https://doi.org/10.3389/fmicb.2021.796363>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2023). PubChem 2023 update. *Nucleic Acids Research*, 51(D1), D1373–D1380. <https://doi.org/10.1093/nar/gkac956>
- Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305(3), 567–580. <https://doi.org/10.1006/jmbi.2000.4315>

- Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3), 3–25. [https://doi.org/10.1016/S0169-409X\(96\)00423-1](https://doi.org/10.1016/S0169-409X(96)00423-1)
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612. <https://doi.org/10.1002/jcc.20084>
- Sharma, V., Katiyar, A., & Agrawal, R. C. (2018). *Glycyrrhiza glabra*: Chemistry and pharmacological activity. In J.-M. Mérillon & K. G. Ramawat (Eds.), *Sweeteners* (Reference Series in Phytochemistry, pp. 87–100). Springer. https://doi.org/10.1007/978-3-319-27027-2_21
- Syaad, K. M., Abd El-Rahim, I. H. A., Mostafa, Y. S., & Ahmed, A. E. (2023). *In silico* subtractive proteomics and molecular docking approaches for the identification of novel inhibitors against *Streptococcus pneumoniae* strain D39. *Life*, 13(5), 1128. <https://doi.org/10.3390/life13051128>
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), 455–461. <https://doi.org/10.1002/jcc.21334>
- U.S. Department of Agriculture, Agricultural Research Service. (1992–2016). *Dr. Duke's phytochemical and ethnobotanical databases*. <https://phytochem.nal.usda.gov/>
- Wang, S., Ahmadi, S., Ahmadi, S. S., & Abbasi, W. (2022). *Glycyrrhiza glabra* (licorice): A comprehensive review on its phytochemistry, biological activities, clinical evidence and toxicology. *Plants*, 10(12), 2751. <https://doi.org/10.3390/plants10122751>
- Wang, W., Yang, J., Xu, S., Li, Y., & Guo, K. (2018). Lignans and neolignans from Lauraceae and their biological activities. *Molecules*, 23(12), 3164. <https://doi.org/10.3390/molecules23123164>
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1), W296–W303. <https://doi.org/10.1093/nar/gky427>
- Yang, M., & Wang, C. (2024). Isoflavonoid metabolism in leguminous plants. *Frontiers in Plant Science*, 15, 1368870. <https://doi.org/10.3389/fpls.2024.1368870>