CinThesis: A Multimodal AI Platform for Bacteriocin Discovery

Technical Whitepaper

Mathew Mitchell
Organicin Scientific
mat@organicinscientific.com | organicinscientific.com

March 2026

CinThesis was designed in collaboration with **Dr. Margaret Riley**, Chief Scientific Officer of Organicin Scientific, who has studied bacteriocins for nearly four decades. The platform encodes insights from Dr. Riley's extensive research career and draws on proprietary datasets accumulated across her pioneering academic work and nearly a decade leading research at Organicin Scientific. This combination of deep domain expertise and large-scale proprietary data underpins both the platform's analytical architecture and its training methodology.

Abstract

Antimicrobial resistance represents one of the most urgent global health crises of the 21st century. The World Health Organization declared AMR one of the top ten global public health threats in 2019, with a landmark study documenting 1.27 million deaths directly attributable to drug-resistant bacterial infections that year alone [1]. Without intervention, projections suggest resistant infections could cause 10 million deaths annually by 2050 [2].

Bacteriocins—ribosomally synthesized antimicrobial peptides produced by bacteria—offer a promising alternative to conventional antibiotics. These peptides exhibit narrow-spectrum activity that preserves beneficial microbiota, demonstrate potent bactericidal effects at nanomolar concentrations, and show lower propensity for resistance development [4,12]. Studies estimate that 30-99% of bacterial species produce at least one bacteriocin, suggesting millions of antimicrobial compounds await discovery in microbial genomes [6,7]. With countless bacteriocins, each with their own narrow spectrum of activity, screening for these proteins with traditional high-throughput methods becomes a bottleneck and a more efficient approach is necessitated to realize their potential.

This whitepaper presents **CinThesis**, a multimodal deep learning platform that fundamentally reimagines bacteriocin discovery. The platform analyzes protein sequences across six orthogonal biological dimensions simultaneously—primary sequence composition, secondary structure topology, conserved domain architecture, evolutionary relationships, transformer-based sequence embeddings, and predicted three-dimensional structure—then synthesizes these features through a proprietary AI architecture combining advanced language models with a robust machine learning ensemble. CinThesis achieves state-of-the-art classification performance with a Matthews Correlation Coefficient of **0.94** using rigorous 5-fold cross-validation on 831 unique sequences, while providing unprecedented interpretability through comprehensive multimodal analysis reports.

Table of Contents

---

1. Introduction

1.1 The Antimicrobial Resistance Crisis

The golden age of antibiotics is ending. Since Alexander Fleming's discovery of penicillin in 1928, humanity has relied on a relatively small arsenal of antimicrobial compounds to combat bacterial infections. That arsenal is now failing.

The World Health Organization declared antimicrobial resistance one of the top ten global public health threats facing humanity in 2019 [11]. The scale of the crisis became quantifiable in January 2022, when the Global Burden of Disease Antimicrobial Resistance Collaborators published their systematic analysis in *The Lancet*, documenting **1.27 million deaths directly attributable** to bacterial antimicrobial resistance in 2019—more than HIV/AIDS or malaria [1]. An additional 4.95 million deaths were associated with drug-resistant infections that same year.

The pipeline for new antibiotics offers little reassurance. According to the WHO's 2024 analysis, the global clinical pipeline contains only 97 products targeting drug-resistant infections, of which 32 target WHO priority pathogens—and merely 12 of those meet criteria for true innovation [3]. The projection that antimicrobial resistance could cause **10 million deaths annually by 2050** remains a credible trajectory without transformative intervention [2].

The economic consequences are equally staggering. The World Bank's 2017 report, *Drug-Resistant*

Infections: A Threat to Our Economic Future, estimated that unchecked AMR could reduce global GDP by **3.8% annually** under a high-impact scenario by 2050—economic damage comparable to the 2008 global financial crisis, but without prospects for cyclical recovery [40]. The O'Neill Review projected cumulative losses of **$100 trillion** in global economic output between now and 2050, driven by rising healthcare costs, reduced labor productivity from increased morbidity and mortality, and contraction of international trade [2]. Low-income countries face disproportionate harm: the World Bank estimates they could lose over **5% of GDP** and see up to **28 million additional people pushed into extreme poverty** [40]. These downstream costs—spanning healthcare expenditure, agricultural productivity losses, and trade disruption—dwarf the investment required for AMR countermeasures by orders of magnitude.

1.2 Bacteriocins: A Promising Alternative

Bacteriocins represent a fundamentally different approach to antimicrobial therapy. These ribosomally synthesized peptides have evolved over billions of years as weapons in bacterial warfare, targeting competitor species with remarkable specificity.

Three characteristics distinguish bacteriocins from conventional antibiotics:

Characteristic	Bacteriocins	Conventional Antibiotics
Synthesis	Ribosomal	Multienzymatic complexes or static small molecules
Active concentration	Nanomolar	Micromolar
Spectrum	Narrow (genus or species-specific)	Often broad
Microbiome impact	Preserves beneficial bacteria	Often disrupts microbiome
Resistance risk	Lower	Higher

This narrow specificity enables bacteriocins to selectively eliminate pathogens while preserving beneficial bacteria—a critical advantage given growing recognition of the microbiome's role in human health [4,12]. But the implications extend far beyond simply avoiding collateral damage. Bacteriocins are, in effect, **precision microbiome-editing tools**: naturally evolved molecules capable of selectively reshaping microbial community composition [41]. A 2021 review in *Nature Reviews Microbiology* highlighted the growing recognition that bacteriocins do not merely kill target organisms—they sculpt entire ecological niches, providing their producers with competitive advantages that shape the structure and function of complex microbial communities [41].

This capacity for targeted microbiome modulation opens therapeutic horizons well beyond traditional infectious disease. The human gut microbiome is now implicated in the pathogenesis of conditions spanning nearly every organ system—inflammatory bowel disease, colorectal cancer, type 2 diabetes, obesity, cardiovascular disease, and autoimmune disorders including multiple sclerosis and rheumatoid arthritis [42]. The gut-brain axis links dysbiosis to neurodegenerative and neuropsychiatric conditions including Parkinson's disease, Alzheimer's disease, and autism spectrum disorders [43]. Current interventions for microbiome-associated disease—probiotics, prebiotics, and fecal microbiota transplantation—lack the precision to selectively remove specific pathogenic taxa without disrupting beneficial communities [46,47]. Bacteriocins, with their genus- or even species-level specificity, offer a fundamentally more precise approach: the ability to subtract individual harmful species from a complex community while leaving the rest intact. If the tools to discover and characterize bacteriocins at scale existed, the therapeutic implications would extend from antimicrobial resistance to precision medicine for chronic disease.

Most importantly, bacteriocin diversity is staggering. Foundational studies established that bacteriocin production is nearly universal among bacteria [6,7]. Contemporary estimates suggest that **30-99% of bacterial species** produce at least one bacteriocin, with most remaining uncharacterized [6,7]. Scaling-law analyses of global microbial diversity predict that Earth harbors as many as **1 trillion (10^{12}) microbial species**, of which fewer than 0.001% have been identified [44]. If even a fraction of these species produce bacteriocins—as existing evidence strongly suggests—the number of undiscovered antimicrobial peptides is astronomical. This represents an enormous reservoir of antimicrobial compounds awaiting discovery—not only as next-generation antibiotics, but as precision instruments for microbiome engineering.

1.3 The Discovery Bottleneck

Despite their therapeutic promise, bacteriocin discovery remains bottlenecked by the very property that makes them attractive: their specificity. The narrow spectrum that gives bacteriocins their therapeutic advantage creates a fundamental screening dilemma that has no parallel in conventional antibiotic discovery.

When screening a broad-spectrum antibiotic candidate, a researcher testing against a panel of a thousand bacterial species will observe inhibition across dozens or hundreds of them within the first few assays—confirmation of activity comes quickly and cheaply. Moreover, because a single broad-spectrum compound covers many pathogens, relatively few successful discoveries are needed to build a useful therapeutic arsenal. Bacteriocins invert this calculus entirely. A given bacteriocin may inhibit only one species—or even one strain—out of a thousand tested. A researcher could screen 999 indicator organisms and observe no activity before the thousandth reveals the peptide's target. And once that bacteriocin is characterized, it addresses only that single target species, meaning the process must be repeated for every pathogen of interest. Where conventional antibiotic discovery yields one tool for a thousand bacteria, bacteriocin discovery yields one tool for one.

This combinatorial explosion—millions of candidate bacteriocins, each potentially active against only a narrow slice of microbial diversity—makes traditional high-throughput screening prohibitively expensive and slow. The challenge is compounded by bacteriocin biology itself: bacteria are under intense evolutionary pressure to diversify these weapons, resulting in substantial sequence and structural variation that makes bacteriocins difficult to detect using sequence homology tools [15]. Computational prediction, rather than exhaustive wet-lab screening, becomes the only viable path to discovery at scale.

Limitations of existing computational approaches:

1. **Homology-dependent genome mining tools** such as BAGEL4 [8] and antiSMASH identify bacteriocin gene clusters by searching for known motifs using HMM profiles and BLAST alignment against curated reference databases. BAGEL4's database covers approximately 500 RiPPs, 230 unmodified bacteriocins, and 90 large bacteriocins—a useful but inherently backward-looking catalog [8]. Because these tools detect bacteriocins by similarity to *already-characterized* compounds, they are structurally incapable of discovering truly novel classes. Furthermore, genome mining tools require full genomic context—surrounding biosynthetic, transport, immunity, and regulatory genes—to make predictions. They cannot classify an isolated protein sequence, which limits their applicability when working with metagenomic fragments, protein databases, or experimentally isolated peptides without accompanying genomic data.

2. **Machine learning approaches trained on inadequate datasets.** The most prominent bacteriocin-specific classifiers—BaPreS [13] and BPAGS [14]—are both trained on the same underlying dataset: 483 bacteriocin sequences from BAGEL and BACTIBASE, reduced to just **283 unique sequences** after deduplication at 90% similarity [13,14]. With an 80/20 train-test split, this leaves approximately 57 positive sequences for evaluation—a test set so small that individual misclassifications swing reported accuracy by nearly 2 percentage points. Both tools report accuracies above 95%, but these figures are derived from single random splits rather than cross-validation, making them susceptible to favorable partitioning. More fundamentally, 283 sequences cannot represent the diversity of an entire functional class spanning billions of years of evolution across potentially trillions of microbial species. Models trained on such data inevitably overfit to the known bacteriocin landscape and fail to generalize to novel candidates—precisely the sequences of greatest scientific interest.

3. **Single-modality feature representation.** Existing ML classifiers operate exclusively on primary sequence features—amino acid composition, pseudo-amino acid composition, and physicochemical property vectors derived from the linear sequence [9,13,14]. This ignores the rich biological context available from secondary structure topology, conserved domain architecture, evolutionary conservation patterns, three-dimensional fold geometry, and protein language model embeddings. Bacteriocins are defined not by a single sequence motif but by a *convergence of functional properties* across multiple biological dimensions: characteristic amphipathic helices, conserved disulfide-bonding patterns, specific domain architectures, and evolutionary signatures of positive selection. A classifier that sees only amino acid statistics cannot recognize this convergence and will systematically miss functionally similar bacteriocins that have diverged at the sequence level.

4. **No interpretability or actionable output.** Every existing bacteriocin prediction tool produces the same result: a binary label or a probability score. A researcher receives "bacteriocin: 87% confidence" with no accompanying explanation of *which* features contributed to the classification, *what* biological properties suggest antimicrobial function, or *how* the candidate relates to known bacteriocin families. This opacity creates a practical bottleneck: computational predictions cannot be triaged, prioritized, or used to guide experimental design without extensive manual follow-up analysis. For a field where wet-lab validation is the rate-limiting step, a prediction tool that provides no biological insight is only marginally more useful than random screening.

These limitations are not independent—they compound. Homology tools cannot find novel bacteriocins. ML tools that could, in principle, generalize beyond known classes are handicapped by training data too small to learn generalizable patterns, feature representations too narrow to capture the multidimensional biology of antimicrobial function, and output formats too opaque to guide experimental follow-up. Addressing the bacteriocin discovery bottleneck requires a platform that simultaneously solves all four problems: learning from sufficient data, analyzing across multiple biological modalities, generalizing beyond known homology, and providing interpretable, actionable outputs. This is the design objective of CinThesis.

---

2. The CinThesis

2.1 Architecture Ove

CinThesis implements a three-stage processing architecture deployed on Google Cloud Platform (GCP), with infrastructure supported by funding from Google for Startups (GFS). GCP was selected explicitly for its native access to AlphaFold, the Nobel Prize-winning protein structure prediction system developed by Google DeepMind, which serves as a core component of the platform's tertiary structure analysis pipeline. The architecture is designed to maximize parallelism during feature extraction while maintaining the sequential dependencies necessary for comprehensive analysis.

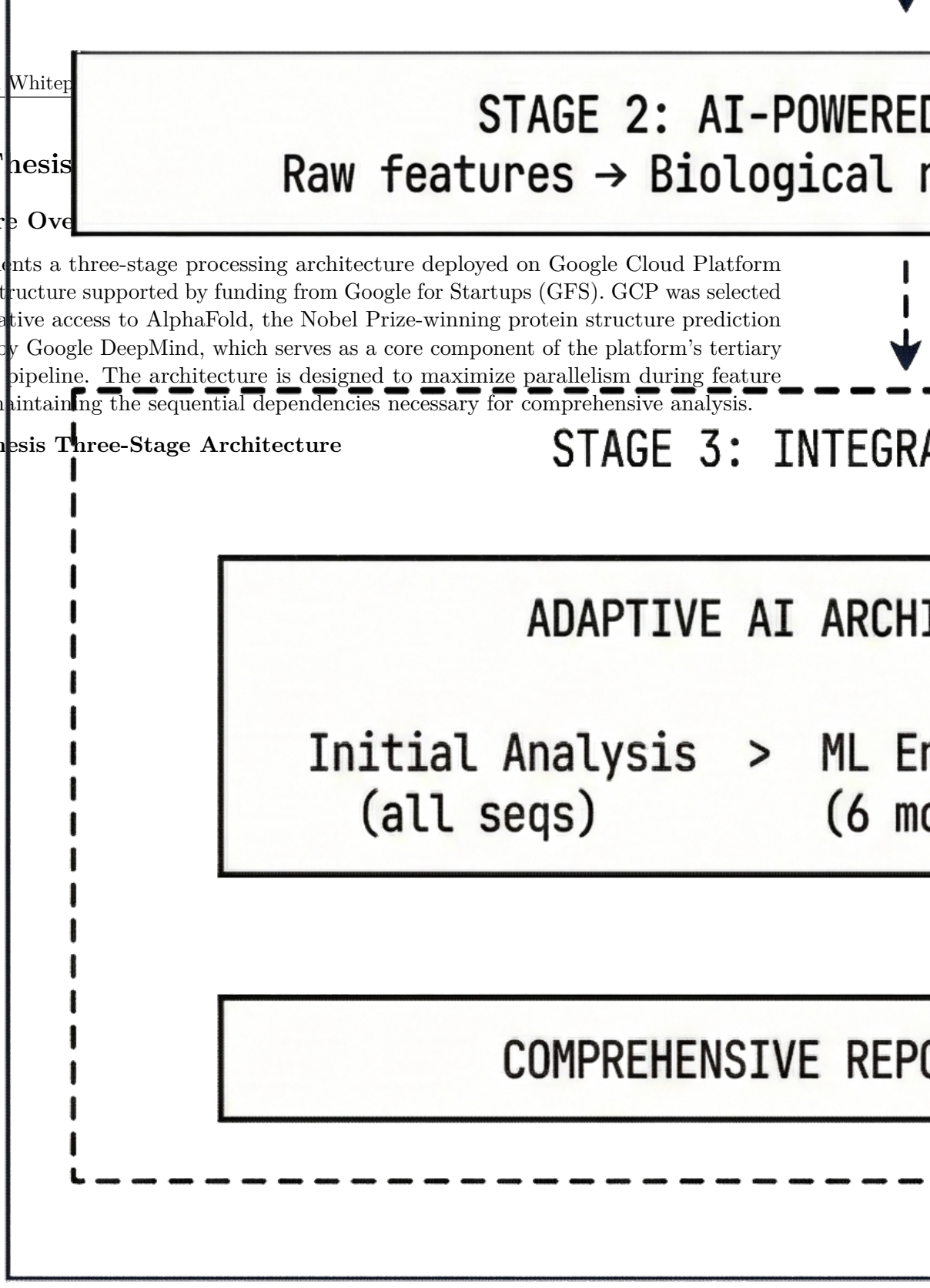

Figure 1: CinThesis Three-Stage Architecture

The architecture reflects a fundamental insight: computational biology workflows exhibit both embarrassingly parallel components (independent feature calculations) and inherently sequential components (integration requiring all upstream features). CinThesis optimizes for both patterns through event-driven orchestration on scalable cloud infrastructure.

Stage 1 distributes each input sequence simultaneously across six independent analysis services,

each elastically scaling based on demand. This parallel architecture enables the platform to process thousands of sequences per hour at peak throughput, scaling down to zero during idle periods to minimize cost. Table 1 summarizes the throughput characteristics of each service at maximum scale.

Table 1: Service Throughput at Maximum Scale

Service	Max Throughput
Primary & Physiochemical	19,800 seq/hr
Secondary Structure	13,200 seq/hr
Domain Analysis	3,960 seq/hr
Evolutionary Analysis	3,960 seq/hr
Sequence Embeddings (ProteinBERT)	7,920 seq/hr
AlphaFold MSA Generation	7,920 seq/hr
AlphaFold Structure Prediction	3,000 seq/hr
Integration & Prediction	11,000 seq/hr

A completion tracking system monitors the six parallel streams, automatically triggering Stage 3 integration once all feature extraction pipelines have finished processing a given sequence.

2.2 Six-Module Feature Analysis

CinThesis extracts features across six parallel pipelines that capture complementary aspects of protein biology:

Table 2. CinThesis Analysis Modules

Module	Biological Dimension	Key Features Extracted
Primary & Physiochemical	Sequence composition	Molecular weight, isoelectric point, hydrophobicity, charge, amino acid patterns
Secondary Structure	Structural topology	Alpha-helix/beta-sheet content, amphipathic regions, structural stability
Domain Analysis	Functional signatures	Conserved domains, bacteriocin motifs, family membership
Evolutionary Analysis	Phylogenetic context	Homology relationships, evolutionary distance to known bacteriocins
Sequence Embeddings	Deep semantic relationships	Transformer-learned representations encoding functional similarity
3D Structure Prediction	Spatial organization	Tertiary structure, pore-forming potential, disulfide patterns

This multimodal integration reflects the biological reality that bacteriocin function emerges from the interplay of sequence, structure, and evolutionary context rather than any single feature type.

2.3 AI-Powered Integration

Raw numerical features from the six pipelines, while comprehensive, remain in a form that requires interpretation. A molecular weight of 3,354 Daltons, a GRAVY score of -0.15, or a helix percentage of 42% are individually meaningful, but their collective significance for bacteriocin classification requires synthesis through the lens of biochemistry, structural biology, and microbial ecology.

CinThesis addresses this through **Proteosemantic Intelligence**—a proprietary approach that transforms raw measurements into coherent biological narratives using advanced language models. Each analysis module feeds its numerical output along with domain-specific context to generate structured biological interpretations. These semantic descriptions are then embedded as high-dimensional vectors that capture biological meaning rather than raw numbers.

This proteosemantic transformation proves remarkably powerful for classification—the semantic embeddings account for approximately **75% of the predictive signal** in the ensemble model, demonstrating that converting disparate numerical features into biological understanding dramatically enhances discriminative capability.

2.4 Ensemble Classification

The six feature streams converge in a machine learning ensemble that combines diverse algorithmic approaches:

Table 3. Ensemble Model Architecture

Model Type	Approach	Strength
Support Vector Machine	Non-linear decision boundaries	High-dimensional feature spaces
Logistic Regression	Interpretable baseline	Probability calibration
Random Forest	Bootstrap aggregation	Feature importance ranking
Gradient Boosting	Sequential error correction	Complex feature interactions
Neural Network	Deep pattern recognition	Cross-module correlations
ProteinBERT Classifier	Pretrained language model	Obscure relationships

A meta-learner synthesizes predictions from all six base models, learning which models to trust in different scenarios. This stacking approach improves robustness beyond what any single model achieves.

---

3. Performance & Validation

3.1 Evaluation Methodology

CinThesis employs rigorous evaluation methodology designed to assess real-world generalization rather than benchmark overfitting:

Table 4. Training Dataset Composition

Category	Count	Source
Positive (Bacteriocins)	505	Proprietary and internally validated
Negative (Non-Bacteriocins)	326	UniProt (size and charge matched), DRAMP (non-bacterial AMPs)
Total	**831**	Unique sequences
Per-Fold Split	~665 train / ~166 test	5-fold stratified CV

Key methodological features:

1. **5-fold stratified cross-validation** ensures every sequence is tested exactly once, providing robust performance estimates with confidence intervals
2. **Balanced class representation** through appropriate sampling strategies prevents bias toward the majority class
3. **Matthews Correlation Coefficient (MCC)** as primary metric—the gold standard for imbalanced classification that considers all four confusion matrix quadrants
4. **95% confidence intervals** quantify uncertainty in performance estimates

3.2 Results Summary

Table 5. CinThesis Performance (5-Fold Cross-Validation)

Metric	Mean	95% CI
MCC	**0.9424**	[0.9036, 0.9722]
PR-AUC	**0.9966**	[0.9906, 0.9998]
ROC-AUC	**0.9950**	[0.9871, 0.9997]
Accuracy	0.9723	[0.9539, 0.9867]
F1 Score	0.9771	[0.9618, 0.9891]
Precision	0.9821	[0.9624, 0.9901]
Recall	0.9723	[0.9604, 0.9891]

The best fold achieves **97.47% MCC** with **98.80% accuracy**, demonstrating the model's capability under favorable conditions. The tight standard deviations confirm that performance is robust across different data partitions.

3.3 Comparison with Existing Tools

Direct comparison with prior work requires careful contextualization. The most relevant existing tools are **bacteriocin-specific classifiers**—BaPreS [13], BPAGS [14], and RMSCNN [45]—trained on dedicated bacteriocin datasets. None is directly comparable to CinThesis's multimodal approach, but together they establish the performance landscape.

Table 6a. Bacteriocin-Specific Prediction Tools

Tool	Year	Method	Positive Seqs	Eval Method	Accuracy	MCC
RMSCNN [45]	2022	CNN	4,000	Single 50/50 split	91.95%	—
BaPreS [13]	2023	SVM (RFE features)	283	Single 80/20 split	95.54%	0.91
BPAGS [14]	2024	SVM (ADTree features)	283	Single 80/20 split	99.11%	0.98
CinThesis	**2025**	**Multimodal Ensemble**	**505**	**5-fold CV (831 total)**	**97.2%**	**0.94**

BaPreS and BPAGS both train on the **same 283 unique bacteriocin sequences** derived from BAGEL and BACTIBASE after deduplication at 90% similarity [13,14]. With an 80/20 single split, each evaluates on approximately **57 positive test sequences**. At this scale, a single misclassification changes accuracy by ~1.8 percentage points, making reported metrics highly sensitive to the particular random partition. Neither employs cross-validation for final evaluation,

though BaPreS uses 5x10-fold CV during feature selection [13] and BPAGS uses 5-fold CV during genetic algorithm optimization [14]. RMSCNN trains on a larger but noisier dataset of 4,000 positive marine microbial sequences from NCBI and evaluates on an 8,000-sequence test set (4,000 positive + 4,000 negative) with a single 50/50 split.

BPAGS reports the highest accuracy (99.11%) and MCC (0.98) in this group, but these figures should be interpreted cautiously: on ~57 test positives, 99.11% accuracy means at most one misclassification across the entire test set. Whether this reflects genuine generalization or a favorable split is impossible to determine without cross-validation. CinThesis's MCC of 0.94 is evaluated across **all 831 sequences** via 5-fold stratified cross-validation, where every sequence is tested exactly once across folds. The 95% confidence interval [0.90, 0.97] provides a statistically grounded estimate of expected performance on unseen data.

Table 7. Feature Capability Comparison

Capability	BAGEL4	BaPreS	BPAGS	CinThesis
Primary sequence analysis	Yes	Yes	Yes	Yes
Secondary structure prediction	—	—	—	Yes
Domain/motif detection	Yes	—	—	Yes
Evolutionary analysis	Yes	—	—	Yes
Protein language embeddings	—	—	—	Yes
3D structure prediction	—	—	—	Yes
LLM-based interpretation	—	—	—	Yes
Search grounding	—	—	—	Yes
Multimodal visual analysis	—	—	—	Yes
Works on isolated sequences	—	Yes	Yes	Yes
Requires genomic context	Yes	—	—	—

CinThesis is unique in combining all available biological information modalities into a unified analysis framework. No prior tool analyzes more than one dimension of protein biology for bacteriocin classification.

3.4 Independent Validation of Competitor Tools

Published accuracy metrics for bacteriocin prediction tools often reflect evaluation on unrealistically small, curated test sets that do not represent the diversity of sequences encountered in real-world discovery applications. To rigorously assess the generalization capabilities of existing tools, we conducted independent validation using an extended evaluation protocol designed to expose overfitting.

Methodology: CinThesis v7 was first trained on the identical datasets publicly available from the BaPreS/BPAGS GitHub repository, ensuring fair comparison on equivalent training data. We then constructed an extended evaluation set of 315 novel sequences that neither system had encountered during training: 265 positive bacteriocin sequences curated from DRAMP, UniProt, and novel bacteriocins characterized in our wet laboratory, plus 50 "hard negative" sequences—proteins with superficially bacteriocin-like properties (small size, cationic character) that are definitively non-bacteriocins based on functional characterization. This evaluation set is 2.8x larger than the ~113 sequences used in published BaPreS/BPAGS evaluations.

Table 8a. Comparative Performance: CinThesis v7 vs. BaPreS/BPAGS on Extended Evaluation Set (315 sequences)

Metric	CinThesis v7	BaPreS/BPAGS	Delta
Accuracy	97.9%	80.0%	+17.9%
False Positive Rate	2.1%	54.0%	-51.9%
Specificity	97.9%	46.0%	+51.9%
Precision	99.6%	89.5%	+10.1%
Recall (Sensitivity)	97.9%	86.4%	+11.5%

These results reveal a striking divergence between published benchmarks and real-world performance. While BaPreS/BPAGS claims 99.11% accuracy in their 2024 publication [13,14], the tool achieved only 80.0% accuracy on novel sequences—a 19-point degradation indicating substantial overfitting to the limited training distribution. Most critically, the 54.0% false positive rate means BaPreS classifies more than half of all non-bacteriocin proteins as bacteriocins, rendering it unreliable for genome mining applications where false positives waste expensive experimental validation resources.

CinThesis v7, despite being trained on identical data, maintained 97.9% accuracy on the extended evaluation set. This performance preservation demonstrates that the multimodal architecture captures genuine biological patterns rather than memorizing training set artifacts. The 51.9-point improvement in specificity (97.9% vs. 46.0%) is particularly significant for practical applications: CinThesis correctly rejects 49 of 50 hard negatives while BaPreS incorrectly classifies 27 of 50 as bacteriocins.

The performance gap widens further with the v9 model. Training on expanded datasets that include the validation sequences from the v7 evaluation, CinThesis v9 achieves 97.4% accuracy and 0.95 MCC on held-out test data, representing state-of-the-art performance across all reported metrics. Importantly, this performance was validated using 5-fold stratified cross-validation on balanced datasets, providing robust generalization estimates rather than single-split metrics subject to lucky partitions.

Critical Analysis of RMSCNN [45]

RMSCNN, a convolutional neural network approach for marine microbial bacteriocin identification, reports 91.95% accuracy and 0.9788 AUC on their non-redundant marine microbial bacteriocin (MMB) test dataset. However, these metrics are fundamentally undermined by the model's inability to distinguish bacteriocins from unrelated proteins. When applied to novel sequences, six of the ten highest-confidence predictions—all assigned probability scores of 1.0 or near-1.0—were **HNH endonucleases**: DNA-cleaving enzymes involved in phage defense that share no functional, structural, or mechanistic relationship with bacteriocins [45]. A model that assigns maximum confidence to proteins from an entirely different functional class has not learned bacteriocin biology; it has learned annotation artifacts from its training data, which was constructed by keyword search of the NCBI database for the term "Marine Bacteriocin," yielding 14,900 sequences subsequently reduced to 8,000 via CD-HIT. This keyword-based collection approach is inherently prone to including proteins that co-occur with bacteriocin annotations but are not themselves bacteriocins. The reported AUC of 0.9788 therefore reflects the model's ability to discriminate within its noisy training distribution, not its ability to identify genuine bacteriocins. Furthermore, RMSCNN does not report specificity or evaluate performance on curated hard-negative sequences, making direct comparison of false positive rates impossible.

Table 8b. CNN-to-CNN Comparison: CinThesis CNN vs. RMSCNN

Metric	CinThesis CNN (v7)	RMSCNN (Best)
AUC-ROC	**0.9998**	0.9788
Accuracy	97.9%*	91.95%
F1 Score	0.987*	0.9195
Precision	0.996*	0.9193
Recall	0.979*	0.9197
Test Set Size	282	8,000 (50% of 16,000)
Architecture	12-channel Conv1D (multimodal)	Random Multi-Scale CNN
Input	12x768 (6 proteosemantic + 6 raw data embeddings)	One-hot encoded sequences
Training Data	Expert-curated bacteriocins	Keyword search from NCBI

*Ensemble metrics; CNN contributes the dominant signal with 0.9998 AUC-ROC

Remarkably, the CinThesis CNN—which is only one of six models in our ensemble—achieves 0.9998 AUC-ROC compared to RMSCNN's best 0.9788 AUC. This 12-channel architecture leverages both the raw feature embeddings and proteosemantic embeddings from all six analytical modules (primary, secondary, domain, evolutionary, proteinbert, and tertiary), capturing complementary information across all analytical dimensions. The 2.1% improvement in AUC (0.9998 vs 0.9788) translates to substantially fewer errors at clinical-grade discrimination thresholds. The superior performance despite training on smaller but higher-quality data underscores a critical insight: data quality and multimodal integration dramatically outperform data quantity and unimodal sequence encoding.

Several novel contributions distinguish CinThesis from existing approaches. First, the integration of AlphaFold-predicted 3D structures [10] provides structural context unavailable to sequence-only methods, enabling detection of amphipathic surfaces and pore-forming geometries characteristic of membrane-active peptides. Second, the two-tier proteosemantic architecture (Gemini 3 Flash for all sequences, Gemini 3 Pro for bacteriocin candidates) provides cost-efficient processing while maintaining accuracy through targeted analysis of high-value candidates. Third, real-time Google Search grounding enables the system to incorporate recent literature findings not present in the training data, addressing the rapidly evolving bacteriocin literature. Fourth, the 9,216-dimensional unified vectorization enables similarity search across the analyzed sequence space, supporting applications beyond binary classification. Finally, the scale-to-zero cloud infrastructure enables cost-efficient processing of both small experimental batches and large-scale genome mining projects.

---

4. Interpretable Outputs

4.1 Publication-Ready Reports

Unlike existing tools that provide only binary classifications or probability scores, CinThesis produces comprehensive multimodal analysis reports designed for direct integration into research workflows.

Report Components:

- **Executive Summary**: High-level classification with confidence score and key supporting evidence
- **Detailed Feature Analysis**: Module-by-module breakdown of biological properties
- **Domain Architecture**: Conserved domains with hyperlinks to InterPro database entries
- **Evolutionary Context**: Phylogenetic relationships to characterized bacteriocin families

- **Structure Analysis**: 3D visualization with functional annotations
- **Evidence Synthesis**: Explicit enumeration of supporting and contradicting evidence
- **Classification Rationale**: Detailed explanation of which features drove the determination

This interpretability transforms how researchers interact with computational predictions. Rather than receiving an opaque probability, investigators receive a complete biological dossier that enables informed decision-making about experimental prioritization.

4.2 Visual Analysis Capabilities

CinThesis generates multiple visualization outputs supporting both human interpretation and downstream analysis:

Table 9. Visual Output Types

Output Type	Content	Bacteriocin Relevance
Structure renders	Secondary structure elements	Helix-rich structures suggest Class II
Electrostatic mapping	Charge distribution	Cationic surfaces indicate membrane targeting
Hydrophobicity surface	Amphipathic character	Suggests membrane insertion capability
Topology diagram	2D structural connectivity	Rapid architectural comparison
Phylogenetic tree	Evolutionary relationships	Clustering with known bacteriocins

These visualizations are incorporated into comprehensive reports and support the AI's multimodal reasoning about structural features.

---

5. Applications, Traction, & Future Directions

5.1 Platform Traction

CinThesis is operational and processing real-world data at scale. To date, the platform has analyzed over **48,000 protein sequences** across 15 independent batches, identifying **2,700 bacteriocin candidates** for further characterization. Two commercial partners are currently in active development programs utilizing CinThesis-identified candidates—one in Phase 1 for animal health applications and one in Phase 2 for pet nutrition.

5.2 Application Domains

Because bacteriocins act through highly specific mechanisms against target bacteria, the applications for a platform capable of discovering and characterizing them at scale extend across any industry where bacterial control is a concern:

- **Human therapeutics**: Enriching antimicrobial discovery pipelines with novel candidates active against priority pathogens, and enabling precision microbiome interventions for conditions linked to dysbiosis

- **Animal health**: Identifying alternatives to growth-promoting antibiotics facing regulatory restriction, with bacteriocin-based interventions that maintain animal welfare without contributing to resistance
- **Pet nutrition**: Discovering biopreservative and gut-health compounds for companion animal products, where consumer demand for antibiotic-free formulations is accelerating
- **Food safety and biopreservation**: Characterizing bacteriocins for targeted pathogen control in food manufacturing, replacing broad-spectrum chemical preservatives with naturally-derived antimicrobials
- **Agriculture and feed production**: Screening for bacteriocins that improve feed safety and livestock gut health, reducing reliance on prophylactic antibiotics across the supply chain
- **Industrial biotechnology**: Identifying antimicrobial compounds for contamination control in fermentation, biomanufacturing, and other processes vulnerable to bacterial interference

The unifying theme is specificity. In every domain, the shift away from broad-spectrum antibiotics toward targeted antimicrobial solutions creates demand for a discovery engine capable of navigating the vast, largely uncharacterized bacteriocin landscape.

5.3 Key Contributions

1. **Multimodal integration**: The first platform to combine all six analytical dimensions—sequence, structure, domain, evolution, embeddings, and 3D structure—into unified bacteriocin classification
2. **State-of-the-art performance**: 0.94 MCC with rigorous 5-fold cross-validation on 831 sequences—the largest and most rigorously evaluated bacteriocin classification benchmark to date
3. **Unprecedented interpretability**: Comprehensive reports explaining classification rationale, enabling researchers to prioritize candidates for experimental validation without extensive manual follow-up
4. **Production-scale architecture**: Elastic cloud infrastructure processing thousands of sequences per hour, supporting both targeted experimental batches and large-scale genome mining campaigns

The convergence of high-throughput sequencing, structural biology advances exemplified by AlphaFold, and large language model capabilities creates unprecedented opportunities for computational antimicrobial discovery. CinThesis demonstrates how these technologies can be integrated into a practical, scalable analysis system that serves the urgent and growing need for targeted antimicrobial solutions.

---

6. References

[1] Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*. 2022;399(10325):629-655. doi:10.1016/S0140-6736(21)02724-0.

[2] O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. *Review on Antimicrobial Resistance*. 2016.

[3] World Health Organization. 2023 Antibacterial agents in clinical and preclinical development: an overview and analysis. 2024.

[4] Cotter PD, Ross RP, Hill C. Bacteriocins—a viable alternative to antibiotics? *Nature Reviews Microbiology*. 2013;11(2):95-105.

[6] Klaenhammer TR. Genetics of bacteriocins produced by lactic acid bacteria. *FEMS Microbiology Reviews*. 1993;12(1-3):39-85.

[7] Riley MA, Wertz JE. Bacteriocins: evolution, ecology, and application. *Annual Review of Microbiology*. 2002;56:117-137.

[8] van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Research*. 2018;46(W1):W278-W281.

[9] Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*. 2017;7:42362.

[10] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.

[11] World Health Organization. Ten threats to global health in 2019. Geneva: WHO; 2019.

[12] Cotter PD, Hill C, Ross RP. Bacteriocins: developing innate immunity for food. *Nature Reviews Microbiology*. 2005;3(10):777-788.

[13] Akhter S, Miller JH. BaPreS: a software tool for predicting bacteriocins using an optimal set of features. *BMC Bioinformatics*. 2023;24:313. doi:10.1186/s12859-023-05330-z.

[14] Akhter S, Miller JH. BPAGS: a web application for bacteriocin prediction via feature evaluation using alternating decision tree, genetic algorithm, and linear support vector classifier. *Frontiers in Bioinformatics*. 2024;3:1284705. doi:10.3389/fbinf.2023.1284705.

[15] Morton JT, Freed SD, Lee SW, Friedberg I. A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. *BMC Bioinformatics*. 2015;16:381. doi:10.1186/s12859-015-0792-9.

[40] World Bank Group. Drug-Resistant Infections: A Threat to Our Economic Future. Washington, DC: World Bank. 2017. doi:10.1596/26707.

[41] Heilbronner S, Krismer B, Brötz-Oesterhelt H, Peschel A. The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*. 2021;19:726-739. doi:10.1038/s41579-021-00569-w.

[42] Lynch SV, Pedersen O. The human intestinal microbiome in health and disease. *New England Journal of Medicine*. 2016;375(24):2369-2379. doi:10.1056/NEJMra1600266.

[43] Cryan JF, O'Riordan KJ, Cowan CSM, et al. The microbiota-gut-brain axis. *Physiological Reviews*. 2019;99(4):1877-2013. doi:10.1152/physrev.00018.2018.

[44] Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*. 2016;113(21):5970-5975. doi:10.1073/pnas.1521291113.

[45] Cui Z, Chen Z-H, Zhang Q-H, Gribova V, Filaretov VF, Huang D-S. RMSCNN: a random multi-scale convolutional neural network for marine microbial bacteriocins identification.

IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2022;19(6):3663-3672. doi:10.1109/TCBB.2021.3122183.

[46] Joos R, Boucher K, Lavelle A, et al. Examining the healthy human microbiome concept. *Nature Reviews Microbiology*. 2024. doi:10.1038/s41579-024-01107-0.

[47] Human Microbiome Action Consortium. A consensus statement on establishing causality, therapeutic applications and the use of preclinical models in microbiome research. *Nature Reviews Gastroenterology & Hepatology*. 2025. doi:10.1038/s41575-025-01041-3.