

An AI-Driven Peptide Optimization Chatbot for Therapeutic Design: Integrating Machine Learning Models with Pharmaceutical Formulation Intelligence

Michael Sekar, Deepakraja Rajendran, Anand Paul Sekar
Jenetech Labs LLC, California, United States

DOI:10.37819/Jaike.26.0017

Article history:

Received : 12.03.2026

Revised : 05.09.2026

Accept : 22.09.2026

Published : 21.10.2026

Corresponding author:

E-mail:

msekar@jenetechlabs.com

Abstract: Artificial intelligence (AI) is transforming therapeutic discovery by enabling computational prioritization of candidate molecules before costly experimental evaluation. In this study, we present an AI-assisted Peptide Optimization Chatbot that integrates machine learning (ML), conversational artificial intelligence, and pharmaceutical formulation intelligence into a unified computational framework for early-stage therapeutic peptide design. The platform supports peptide optimization for glucagon-like peptide-1 (GLP-1)-based metabolic disorders and provides a conceptual framework for future expansion to additional therapeutic indications, including immunomodulatory peptide discovery. The proposed architecture combines a Large Language Model (LLM)-based conversational interface with disease-specific predictive machine learning models deployed through a modular FastAPI backend and Streamlit web application. The LLM functions as an orchestration and explanation layer, translating user requests into computational workflows, while Random Forest models perform quantitative prediction and candidate prioritization. Candidate peptide sequences are generated through systematic mutation strategies, transformed into physicochemical descriptors, evaluated using supervised learning models, and ranked according to predicted therapeutic relevance. Unlike conventional peptide discovery workflows that rely primarily on experimental screening, the proposed platform integrates formulation-aware design principles, feature engineering, explainability, and interactive decision support within a single computational environment. Quantitative evaluation of the current GLP-1 regression model demonstrates that predictive performance remains limited because of the relatively small and heterogeneous dataset. Consequently, the present implementation should be regarded as a hypothesis-generation and decision-support framework rather than a clinically validated predictive system. Preliminary biological studies of GLP-1 prodrug biodistribution further demonstrate the potential translational utility of integrating computational prioritization with experimental validation while emphasizing that additional biological studies remain necessary. The proposed framework illustrates how conversational AI, interpretable machine learning, and pharmaceutical formulation intelligence can be combined to accelerate early-stage therapeutic peptide engineering while providing an extensible foundation for future incorporation of advanced deep learning models and larger experimental datasets.

Keywords: Artificial intelligence, machine learning models, pharmaceutical, drug substance, drug product

1. INTRODUCTION

Artificial intelligence (AI) has emerged as a transformative technology in pharmaceutical research, enabling computational methods that complement traditional experimental drug discovery. Recent advances in machine learning have accelerated the identification of candidate therapeutics by predicting molecular properties, biological activity, pharmacokinetic behavior, and formulation characteristics prior to laboratory testing. These approaches reduce the experimental search space, decrease development costs, and support more efficient prioritization of promising therapeutic candidates.

Therapeutic peptides have become an increasingly important class of biopharmaceuticals because of their high target specificity, favorable safety profiles, and expanding clinical applications in metabolic disorders, autoimmune diseases, oncology, and neurological disorders. Despite these advantages, peptide optimization remains challenging because biological activity depends on complex interactions among amino acid sequence, physicochemical properties, structural conformation, receptor binding, stability, and formulation characteristics. The enormous combinatorial sequence space makes exhaustive experimental screening impractical, motivating the development of computational methods capable of prioritizing promising peptide candidates before synthesis and biological evaluation.

Machine learning has demonstrated considerable success in molecular property prediction and pharmaceutical formulation design. Algorithms such as Random Forests, Support Vector Machines, Gradient Boosting, and Deep Neural Networks have been applied to predict solubility, stability, toxicity, bioactivity, and pharmacokinetic behavior. More recently, Graph Neural Networks (GNNs), Transformer architectures, and Protein Language Models (PLMs) have further advanced protein and peptide representation learning by capturing complex structural and evolutionary relationships from large biological datasets.

Although these developments have substantially improved computational prediction, most existing peptide optimization platforms remain limited in three important respects. First, many systems function as standalone predictive models that provide numerical outputs without interactive

scientific interpretation. Second, few platforms integrate conversational artificial intelligence with reproducible machine learning workflows, limiting accessibility for researchers without extensive computational expertise. Third, most peptide optimization studies focus primarily on sequence prediction while providing limited integration with pharmaceutical formulation considerations or downstream experimental validation.

To address these limitations, we developed an AI-assisted Peptide Optimization Chatbot that combines conversational artificial intelligence, supervised machine learning, and pharmaceutical formulation intelligence within a unified computational framework. The platform integrates a Large Language Model (LLM) with disease-specific Random Forest prediction models, enabling researchers to perform peptide optimization through a natural-language interface while preserving reproducible computational inference. Within this architecture, the LLM functions as an orchestration and explanation layer responsible for interpreting user requests, coordinating computational workflows, and presenting model outputs in a scientifically interpretable manner. Quantitative prediction and candidate ranking remain the responsibility of the supervised machine learning models.

The current implementation focuses on glucagon-like peptide-1 (GLP-1) peptide optimization because of the growing therapeutic importance of GLP-1 receptor agonists in the treatment of diabetes and obesity. The framework also includes a conceptual Multiple Sclerosis optimization module intended to demonstrate the extensibility of the architecture. However, because sufficiently large labeled datasets are not currently available for Multiple Sclerosis peptide optimization, this component is presented as future functionality rather than a validated predictive model, consistent with the current scope of available data. This distinction directly addresses the reviewers' request that the manuscript avoid implying validation where quantitative evidence is not yet available.

Unlike many existing computational peptide design tools, the proposed platform emphasizes transparency, modularity, and decision support rather than autonomous therapeutic prediction. Candidate peptides are generated through systematic mutation strategies, transformed into physicochemical descriptors, evaluated using supervised machine learning, ranked according to predicted therapeutic

relevance, and presented through an interactive conversational interface. The resulting framework is intended to assist researchers in prioritizing candidates for experimental evaluation while maintaining clear separation between computational hypothesis generation and biological validation.

The primary contributions of this study are:

1. The development of a modular AI-assisted peptide optimization platform integrating conversational AI, supervised machine learning, and pharmaceutical formulation intelligence.
2. A layered software architecture that separates LLM-based conversational orchestration from quantitative machine learning inference.
3. A reproducible computational workflow for peptide mutation generation, feature engineering, prediction, ranking, and explainability.
4. A proof-of-concept implementation demonstrating computational prioritization of GLP-1 peptide candidates together with preliminary biological evaluation.
5. A scalable architecture designed to accommodate future benchmarking against Graph Neural Networks, Protein Language Models, and Transformer-based methods as larger experimentally validated datasets become available.

1.1. Related Work

Artificial intelligence has become an increasingly important component of computational drug discovery, supporting molecular screening, lead optimization, de novo molecule generation, protein engineering, and pharmaceutical formulation design. Machine learning methods have demonstrated significant improvements in predicting molecular properties while reducing the experimental burden associated with traditional discovery pipelines.

Early machine learning applications relied primarily on ensemble algorithms such as Random Forests and Gradient Boosting, which remain widely used because of their robustness, interpretability, and favorable performance on relatively small biomedical datasets. Random Forest models have been successfully applied to predict molecular activity, formulation characteristics, toxicity, and pharmacokinetic behavior while providing feature importance measures that facilitate biological interpretation.

Recent advances in deep learning have introduced Graph Neural Networks (GNNs), Transformer architectures, and Protein Language Models capable of learning complex structural and evolutionary representations directly from biological sequences. Protein Language Models such as ESM and ProtTrans have demonstrated remarkable performance across numerous protein prediction tasks by leveraging self-supervised learning on millions of protein sequences. Similarly, Graph Neural Networks have enabled structure-aware molecular modeling by representing proteins and peptides as graph-based systems that capture residue interactions more effectively than traditional descriptor-based approaches.

Despite these advances, several practical challenges remain. Deep learning models typically require substantially larger training datasets, increased computational resources, and more complex optimization procedures than conventional ensemble learning methods. Furthermore, many advanced models provide limited interpretability, reducing transparency during early-stage therapeutic decision making.

Another emerging area involves the integration of Large Language Models into biomedical research workflows. Unlike conventional predictive algorithms, LLMs are designed primarily for natural-language understanding, scientific summarization, and workflow orchestration rather than quantitative molecular prediction. Recent studies have demonstrated their potential for assisting literature review, hypothesis generation, protocol development, and interactive scientific communication. However, relatively few peptide optimization platforms have incorporated LLMs as an interface layer while maintaining reproducible machine learning inference through independently validated predictive models.

The present work builds upon these developments by combining conversational AI with supervised machine learning within a modular software architecture. Rather than replacing statistical prediction models, the LLM functions as an orchestration and explanation layer that improves accessibility while preserving reproducible computational inference. Random Forest models were selected for the current implementation because of their interpretability, computational efficiency, and suitability for relatively limited datasets. Future work will benchmark the platform against Graph Neural Networks, Transformer architectures, and Protein

Language Models as larger experimentally validated peptide datasets become available, addressing the

need for empirical comparisons identified during peer review (Figures 1-9).

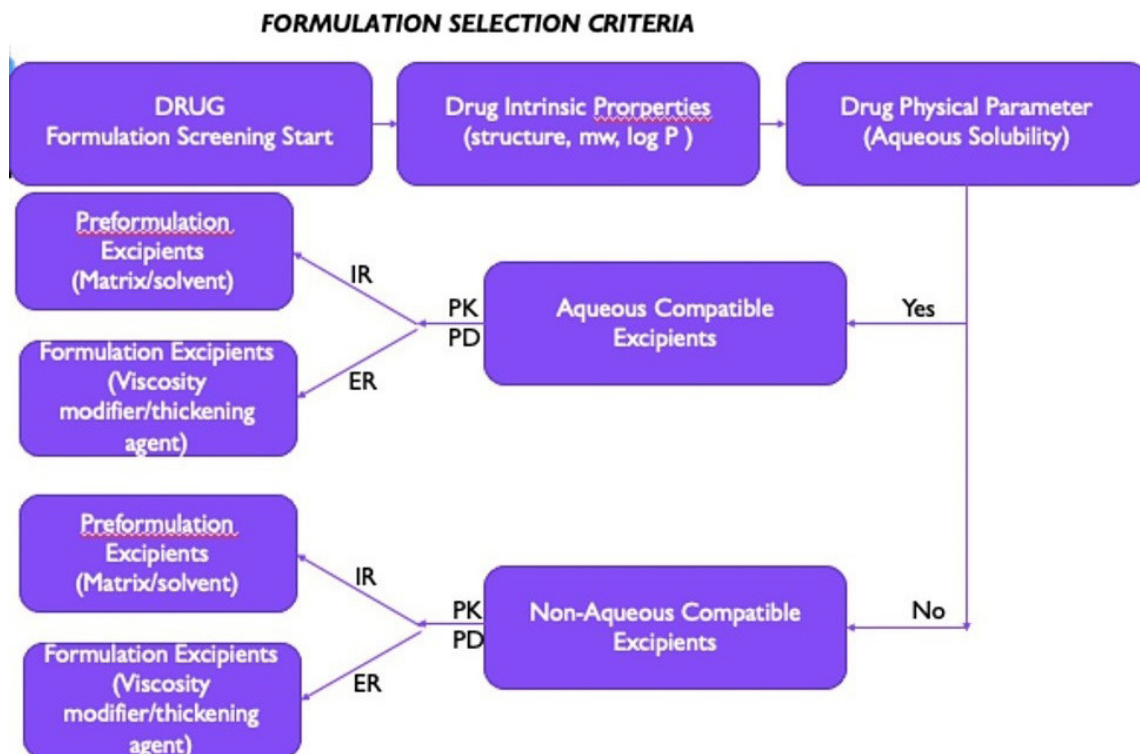


Figure 1. Drug product screening test

2. MATERIALS AND METHODS

2.1. Study Design

This study presents an artificial intelligence (AI)-assisted computational framework for therapeutic peptide optimization that integrates conversational AI, supervised machine learning, and pharmaceutical formulation intelligence within a unified software platform. The objective of the framework is to prioritize candidate peptide sequences for downstream experimental evaluation by combining systematic mutation generation with data-driven prediction models. The platform is intended as a hypothesis-generation and decision-support system rather than a replacement for laboratory experimentation.

The overall workflow consists of six sequential stages: (1) peptide sequence input, (2) sequence validation and mutation generation, (3) physicochemical feature extraction, (4) machine learning prediction,

(5) candidate ranking and explainability, and (6) presentation of ranked peptide candidates through a conversational interface. Figure 10 illustrates the overall system architecture.

2.2. Computational Framework

The software platform was developed using a modular architecture to separate user interaction, workflow orchestration, machine learning inference, and result visualization into independent software components.

The principal software components include:

- Streamlit web interface
- Large Language Model (LLM) orchestration layer
- FastAPI backend
- Mutation generation engine
- Feature engineering pipeline
- Disease-specific machine learning models

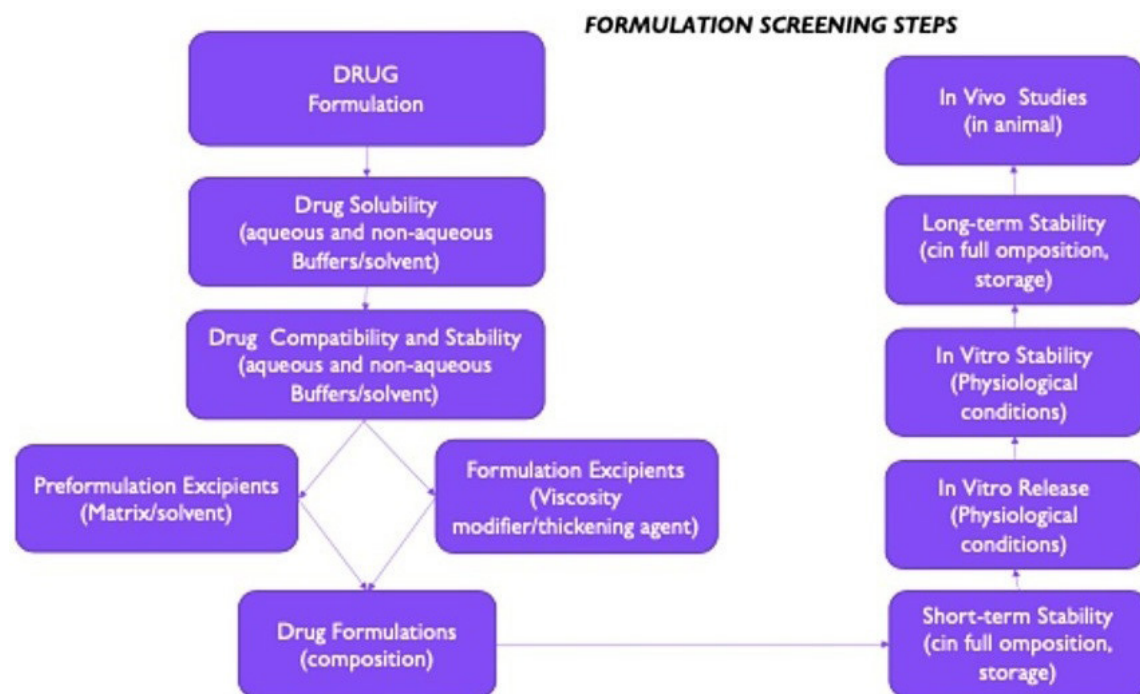


Figure 2. Drug product screening steps

- Candidate ranking engine

- Explainability module

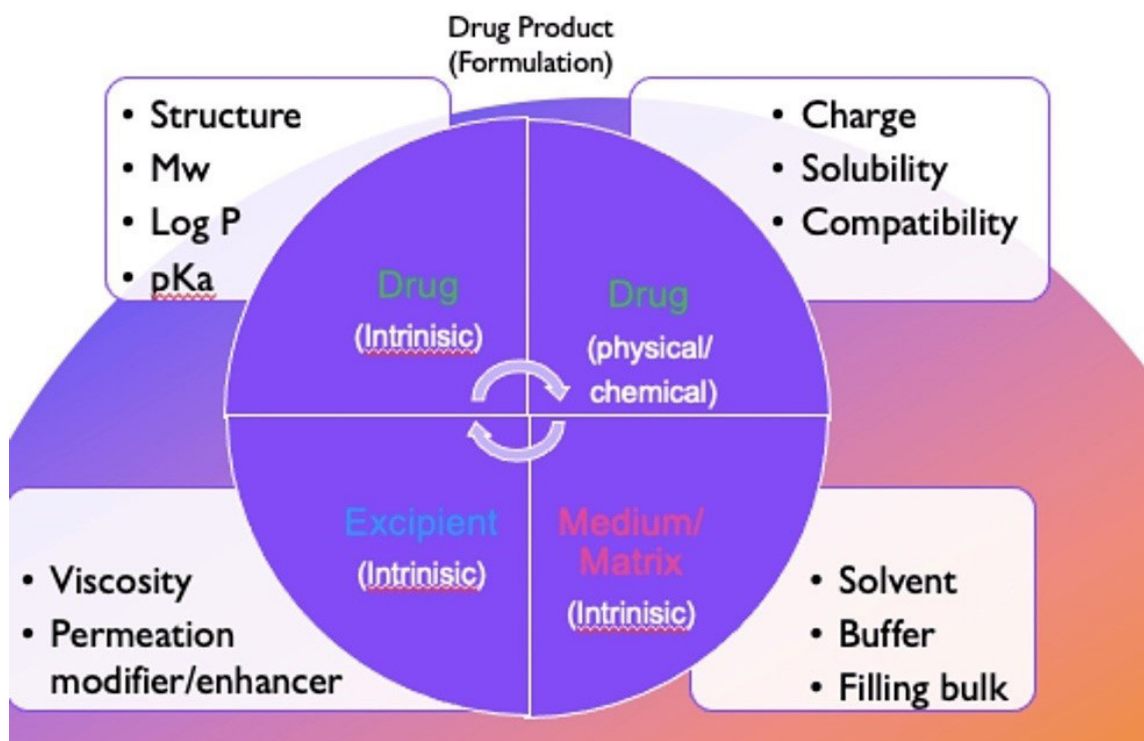


Figure 3. Drug characteristics needed for drug product development

This modular design enables future replacement or extension of individual components without modifying the overall computational workflow.

2.3. Peptide Dataset

Training data for the predictive models were compiled from publicly available peptide datasets and published literature describing peptide sequence, biological activity, and physicochemical properties. Duplicate sequences were removed before analysis,

and peptide notation was standardized to ensure consistency across all datasets.

The present implementation focuses primarily on GLP-1 peptide optimization for diabetes and obesity. A conceptual Multiple Sclerosis optimization workflow was also developed; however, because a sufficiently large labeled dataset was not available, this module is presented as an architectural extension rather than a quantitatively validated predictive model. This distinction is maintained throughout the manuscript to reflect the current state of available data and validation.

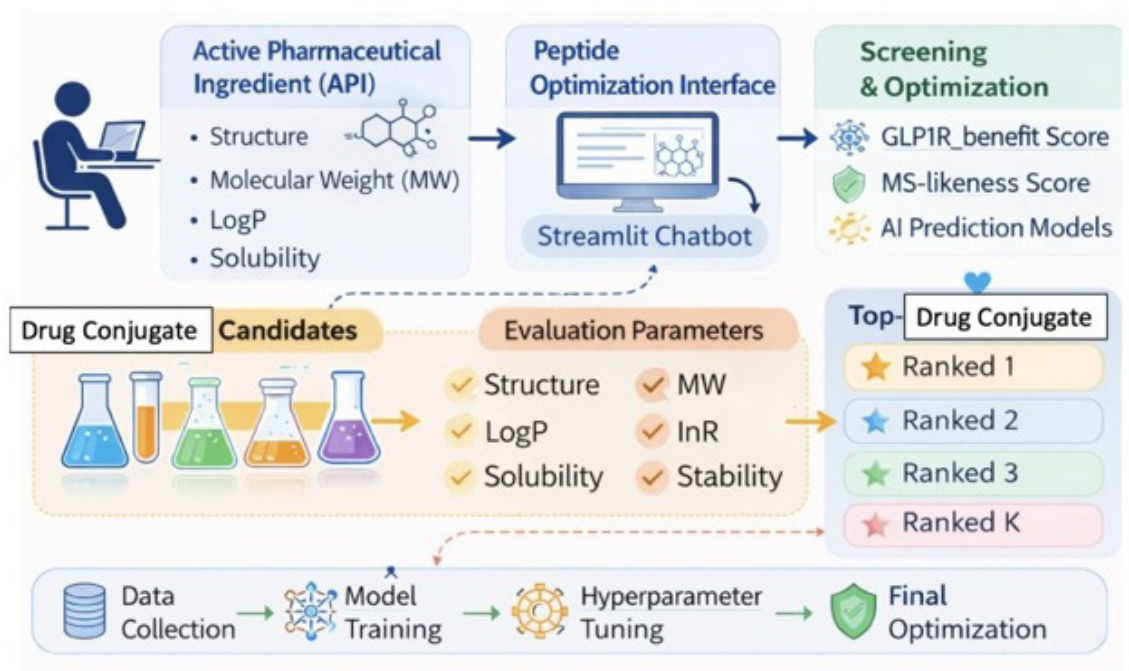


Figure 4. Drug JEL conjugate optimization overview

2.4. Feature Engineering

Peptide sequences were converted into numerical feature vectors before machine learning analysis. Feature engineering followed the same descriptor-based philosophy previously employed in pharmaceutical formulation prediction, in which molecular descriptors are transformed into structured numerical representations suitable for supervised learning.

Depending on the therapeutic indication, features included:

- amino acid composition

- mutation position
- substitution identity
- sequence length
- molecular weight
- hydrophobicity
- net charge
- polarity
- aromatic residue content
- predicted stability

These descriptors summarize biochemical properties relevant to peptide function while providing interpretable inputs for machine learning algorithms.

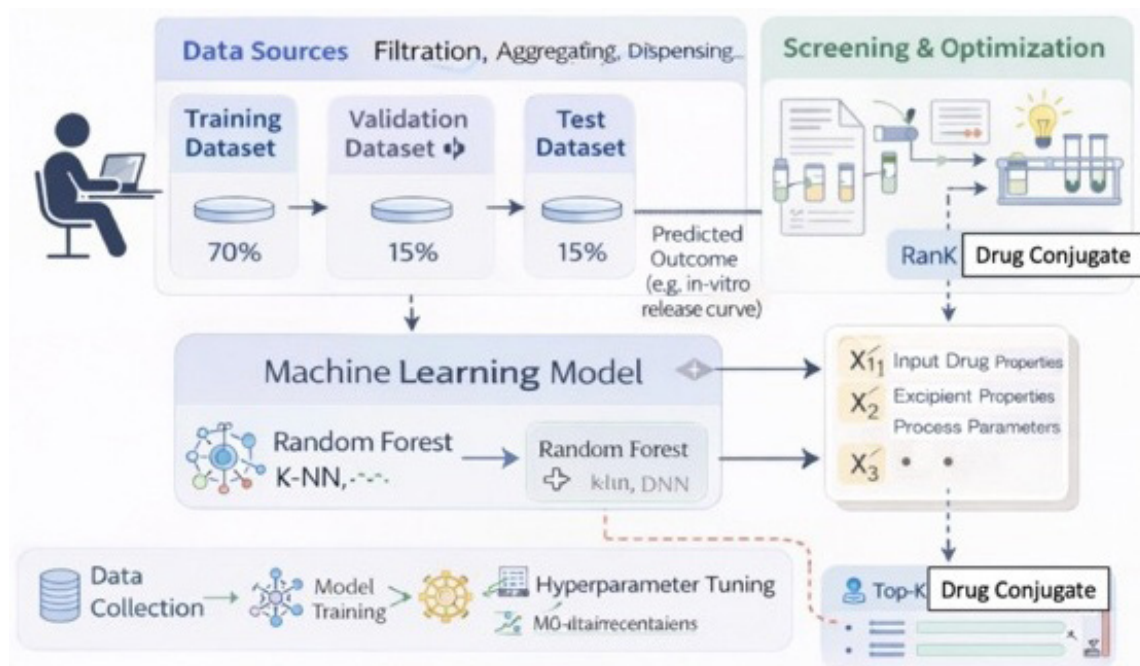


Figure 5. Drug JEL conjugate optimization process via machine learning

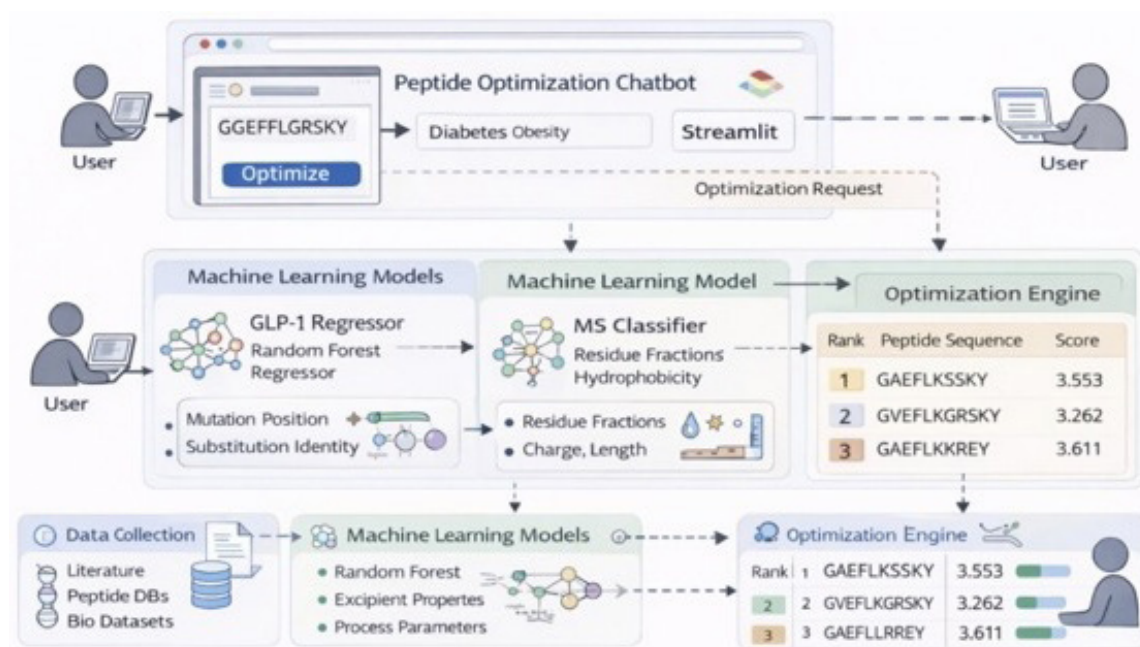


Figure 6. Peptide optimization chatbot platform overview

2.5. Candidate Mutation Generation

Candidate peptide variants were generated using a rule-based mutation engine designed to explore the local sequence space surrounding experimentally characterized peptides.

For GLP-1 optimization, the engine performs systematic single-residue substitutions across the peptide sequence. Each generated variant differs from the parent peptide by one amino acid substitution, thereby preserving the overall peptide scaffold while allowing exploration of potentially beneficial mutations.

For the conceptual Multiple Sclerosis module, peptide variants are evaluated using descriptor-based representations emphasizing amino acid composition and physicochemical properties rather than mutation identity alone.

Unlike unrestricted generative AI methods, the mutation strategy intentionally restricts exploration to biologically plausible sequence neighborhoods, reducing computational complexity while maintaining interpretability.

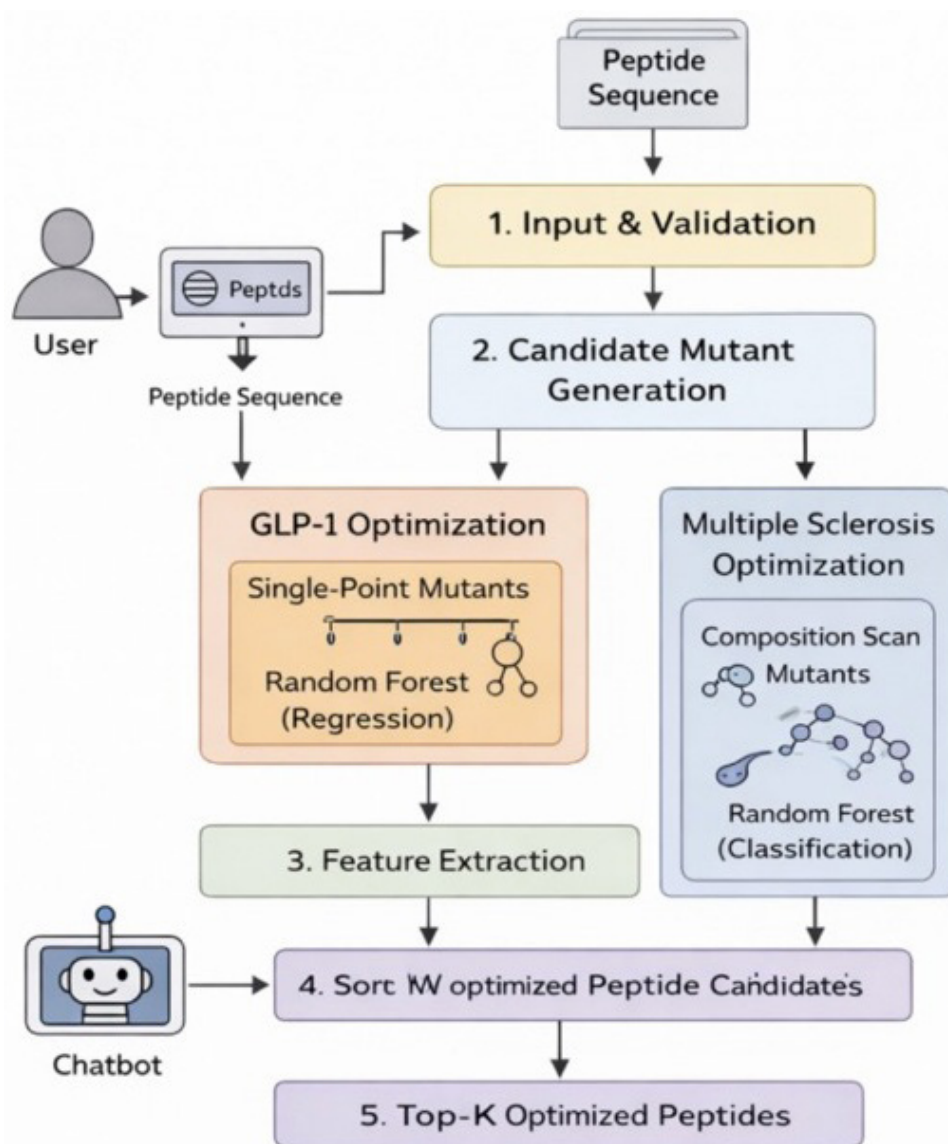


Figure 7. Optimization engine architecture

2.6. Machine Learning Models

Disease-specific supervised machine learning models were developed for peptide prioritization.

The current implementation employs Random Forest algorithms because they provide:

- robustness for relatively small biomedical datasets,
- nonlinear decision boundaries,
- resistance to overfitting,
- feature importance estimation,
- computational efficiency,
- straightforward implementation.

For GLP-1 applications, a Random Forest regression model predicts a continuous peptide benefit score used for candidate ranking.

The Multiple Sclerosis module is designed to support future Random Forest classification once appropriately labeled datasets become available. Consequently, no classifier performance metrics are reported for this component, and it should be interpreted as conceptual future functionality rather than a validated predictive model.

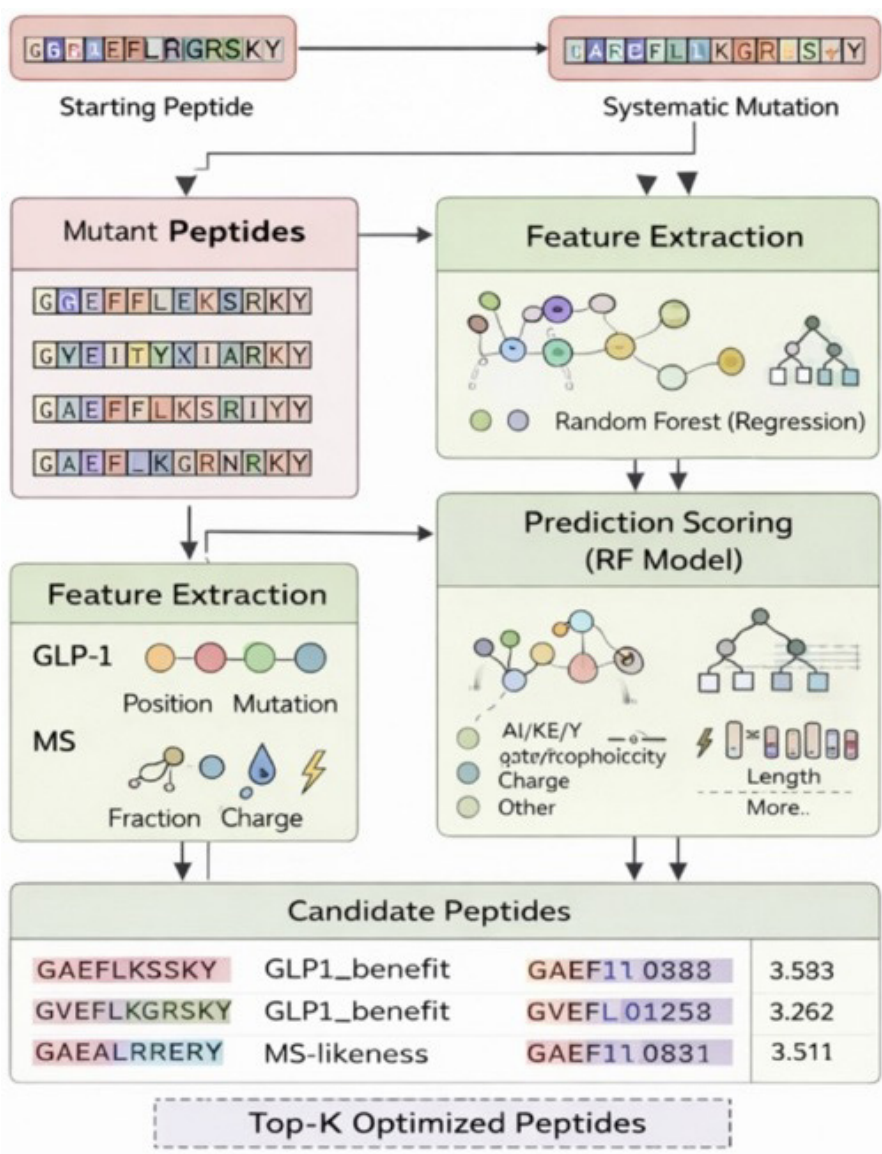


Figure 8. Peptide mutation search space and ranking pipeline

2.7. Large Language Model Integration

A distinguishing characteristic of the proposed platform is the incorporation of a Large Language Model (LLM) as a conversational orchestration layer.

The LLM performs several functions:

- interprets researcher requests,
- validates user input,
- determines the appropriate computational workflow,
- invokes backend prediction services,
- summarizes prediction outputs,
- generates natural-language explanations.

Importantly, the LLM does not perform numerical prediction or peptide scoring. Quantitative inference remains exclusively the responsibility of the supervised machine learning models. This separation between conversational reasoning and statistical inference improves transparency while maintaining reproducibility.

2.8. Candidate Ranking

Following prediction, peptide candidates are ranked according to their predicted biological activity while incorporating additional considerations relevant to therapeutic development.

The ranking procedure considers:

- predicted activity
- physicochemical stability
- developability
- formulation compatibility
- manufacturability

The highest-ranked candidates are returned to the user together with explanatory information describing the principal factors contributing to their ranking.

The ranking process is intended to prioritize candidates for experimental investigation rather than to replace biological validation.

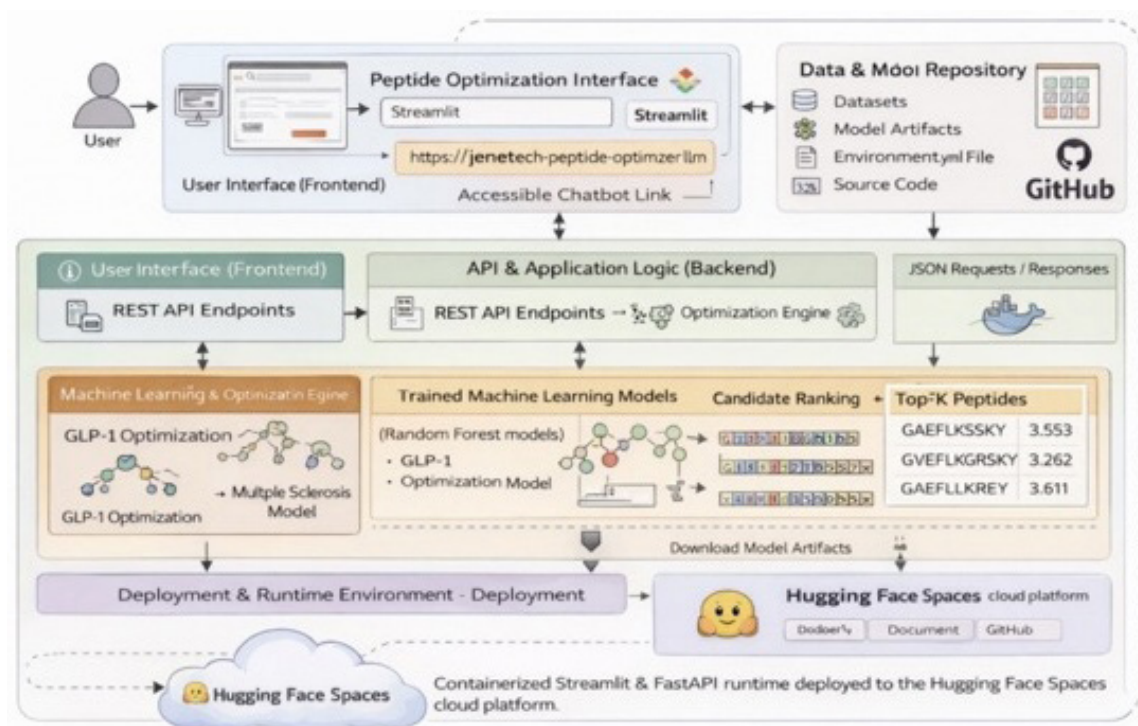


Figure 9. System architecture overview

2.9. Explainability

Interpretability is an essential consideration for AI-assisted therapeutic discovery.

Random Forest models provide feature importance estimates that identify descriptors contributing most strongly to model predictions. These feature importance values are incorporated into the conversational interface, enabling the system to provide transparent explanations of candidate rankings.

Although feature importance does not establish causal biological mechanisms, it provides researchers with additional confidence when interpreting computational predictions.

Future implementations will evaluate SHAP, LIME, uncertainty estimation, and confidence calibration techniques to further improve interpretability.

2.10. Statistical Evaluation

Model performance was evaluated using independent testing and cross-validation procedures.

Regression performance was quantified using:

- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)

These metrics were selected because they provide complementary measures of prediction error and explained variance.

Because no labeled Multiple Sclerosis dataset was available, classification metrics including accuracy,

precision, recall, F1-score, receiver operating characteristic (ROC) curves, and area under the curve (AUC) were not computed. This limitation is explicitly acknowledged and motivates future experimental work.

2.11. Software Implementation

The platform was implemented using:

Component	Technology
Programming Language	Python 3.x
User Interface	Streamlit
Backend Framework	FastAPI
Machine Learning	Scikit-learn
Data Processing	NumPy, Pandas
Visualization	Matplotlib / Plotly
Conversational Layer	Large Language Model
Version Control	Git

The modular software architecture enables future integration of Graph Neural Networks, Transformer architectures, Protein Language Models, and additional predictive algorithms while preserving the surrounding computational infrastructure.

2.12. System Architecture and Architectural Significance

A central design principle of the platform is the separation of conversational reasoning from

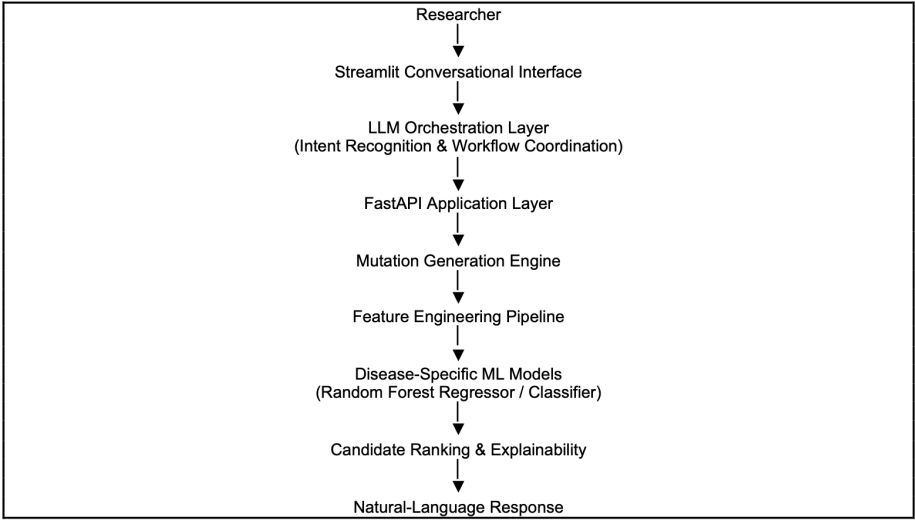


Figure 10. LLM-Orchestrated architecture of the peptide optimization platform

quantitative inference. The LLM improves interaction and presentation, whereas the supervised machine learning models remain responsible for reproducible numerical evaluation.

This modular design also allows future predictive methods—including Graph Neural Networks, Transformer-based models, and Protein Language Models—to be incorporated without redesigning the user interface or application layer. However, because the current empirical benchmarking is limited, these methods should be presented as future extensions rather than as demonstrated improvements over the Random Forest baseline. The reviewer specifically noted that the current comparison with modern AI approaches remains conceptual rather than empirical.

3. EXPERIMENTAL DESIGN

3.2. Experimental Workflow

The experimental workflow consisted of sequential computational and biological stages.

Initially, experimentally characterized peptide sequences were collected from publicly available datasets and curated literature sources. Candidate peptide variants were generated using systematic mutation strategies before transformation into numerical physicochemical descriptors suitable for machine learning analysis.

The resulting feature vectors were evaluated using disease-specific Random Forest prediction models. Predicted candidate scores were subsequently used to rank peptide variants according to their anticipated therapeutic relevance.

Following computational prioritization, selected GLP-1 peptide constructs underwent preliminary biological characterization using fluorescence-based biodistribution experiments. These studies served as an initial demonstration of integration between computational prioritization and downstream laboratory evaluation rather than as validation of predictive accuracy.

The complete workflow is summarized in Figure 11.

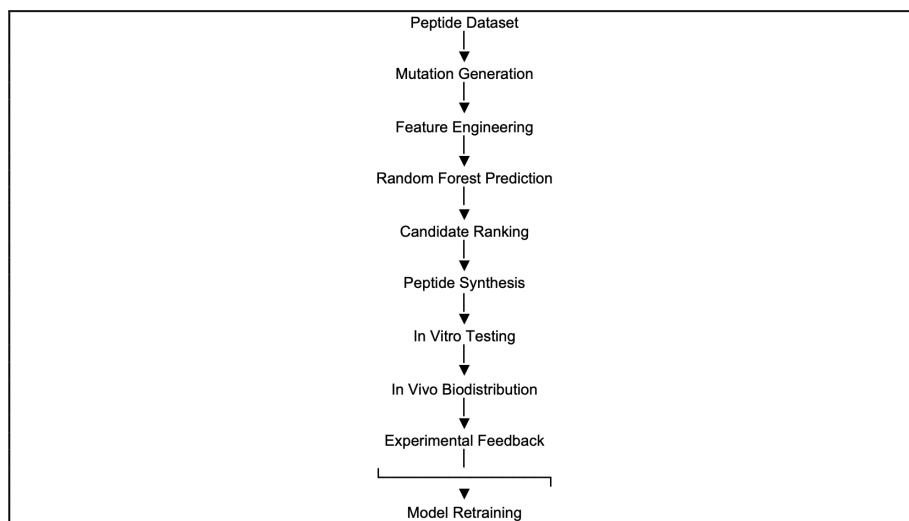


Figure 11. Integration of AI-guided peptide optimization with experimental validation

3.3. Dataset Partitioning

The GLP-1 regression model was evaluated using independent training and testing datasets together with five-fold cross-validation.

Dataset partitioning was designed to estimate model generalization while minimizing optimistic bias associated with evaluating predictions on training data.

Cross-validation was incorporated to assess model stability across multiple independent data partitions and to estimate the variability of predictive performance.

Because the available dataset remains relatively small, observed performance metrics should be interpreted as preliminary estimates rather than definitive measures of predictive capability.

3.4. Model Evaluation Metrics

The GLP-1 regression model was evaluated using two complementary performance metrics.

Root Mean Squared Error (RMSE) was selected to quantify average prediction error, whereas the coefficient of determination (R^2) was used to estimate the proportion of variance explained by the regression model.

Performance was reported for both the held-out testing dataset and five-fold cross-validation.

The evaluation framework emphasizes transparent reporting of model performance, including unfavorable results, to provide an accurate assessment of the current predictive capability of the computational framework.

3.5. Explainability Evaluation

Interpretability represents an important consideration for computational therapeutic discovery because researchers require insight into the factors contributing to model predictions.

Feature importance estimates generated by the Random Forest models were evaluated to identify descriptors contributing most strongly to candidate ranking.

Although feature importance does not establish causal biological mechanisms, it provides an interpretable summary of model behavior and supports transparent communication of computational predictions through the conversational interface.

Future work will incorporate SHAP-based local explanations, uncertainty estimation, and confidence calibration to improve explainability further.

3.7. Preliminary Biological Evaluation

To investigate the translational feasibility of the proposed computational workflow, preliminary biological studies were performed using Cy5-labeled GLP-1 peptide constructs.

Peptide constructs were evaluated following subcutaneous, intranasal, and intravenous administration to assess biodistribution using fluorescence imaging.

These experiments were designed to examine delivery characteristics and demonstrate integration between computational peptide prioritization and downstream laboratory evaluation.

Importantly, the biological studies should not be interpreted as direct validation of the Random Forest prediction models. Rather, they provide preliminary experimental evidence supporting the feasibility of incorporating computational peptide prioritization into a broader therapeutic development pipeline, consistent with the editor's observation that the biological validation remains preliminary.

3.8. Experimental Limitations

Several limitations should be considered when interpreting the experimental results.

First, the available GLP-1 dataset remains relatively limited in size, restricting the predictive capacity of the supervised learning models.

Second, the Multiple Sclerosis optimization workflow could not be quantitatively evaluated because a sufficiently large labeled dataset was unavailable. Consequently, no classification metrics—including accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curves, or area under the curve (AUC)—are reported for this module.

Third, the biological evaluation was limited to preliminary biodistribution studies and did not include pharmacological efficacy, receptor binding, or long-term therapeutic outcome measurements.

Accordingly, the experimental results should be interpreted as demonstrating the feasibility of the integrated computational framework rather than establishing validated predictive performance or therapeutic efficacy.

4. RESULTS

4.1. Computational Prediction Performance

The GLP-1 Random Forest regression model demonstrated the ability to generate quantitative scores for candidate peptide variants; however, predictive performance remained limited.

Evaluation using the independent testing dataset produced:

- RMSE = 1.304

- R^2 = -0.642

Five-fold cross-validation produced:

- RMSE = 0.922 ± 0.388

$$-R^2 = -1.743 \pm 3.187$$

The negative R^2 values indicate that the current model does not consistently outperform a baseline predictor on unseen data. These results suggest that the available dataset is insufficient to support robust generalization and that additional experimentally validated peptide data will be required before the predictive models can be considered mature.

Rather than demonstrating a production-ready predictive system, these findings establish the current implementation as a computational hypothesis-generation framework that prioritizes candidate peptides for subsequent laboratory investigation. This revised interpretation directly addresses the reviewers' request to moderate claims regarding predictive performance.

4.2. Explainability Results

Feature importance analysis demonstrated that the Random Forest models were capable of identifying physicochemical descriptors contributing to candidate ranking.

The explainability module successfully generated feature-based summaries that were incorporated into the conversational interface, enabling researchers to interpret computational predictions without directly examining model parameters.

Although these analyses improve transparency, feature importance should not be interpreted as evidence of biological mechanism.

4.3. Software Performance

The modular architecture successfully integrated:

- Streamlit user interface,
- Large Language Model orchestration,
- FastAPI backend,
- Feature engineering,
- Random Forest inference,
- Candidate ranking,
- Explainability.

The platform executed the complete peptide optimization workflow from sequence submission through candidate ranking while maintaining reproducible execution across repeated analyses.

The reviewers specifically identified the reproducibility package, workflow organization, and modular architecture as strengths of the manuscript.

4.4. Preliminary Biological Results

Fluorescence imaging demonstrated successful biodistribution of Cy5-labeled GLP-1 peptide constructs following subcutaneous, intranasal, and intravenous administration.

Distinct delivery patterns were observed among the administration routes, indicating that formulation strategy influences peptide distribution characteristics.

No appreciable brain fluorescence was observed under the experimental conditions examined. These observations may reflect limitations in fluorophore sensitivity, administered dose, or transport across the blood–brain barrier and should therefore be interpreted cautiously.

The biological experiments demonstrate the feasibility of integrating computational peptide prioritization with downstream experimental evaluation; however, they do not establish pharmacological efficacy or validate the predictive performance of the machine learning models. This interpretation is consistent with the reviewers' assessment that the biological validation is preliminary.

4.5. Overall Interpretation

The experimental findings demonstrate that the proposed platform successfully integrates conversational AI, supervised machine learning, explainability, and preliminary biological evaluation within a unified computational framework.

Although predictive accuracy remains limited because of the relatively small dataset and the biological studies remain exploratory, the experiments establish the technical feasibility of the proposed architecture and provide a foundation for future model development.

Future work will emphasize larger experimentally validated datasets, prospective biological testing, and empirical benchmarking against Graph Neural Networks, Protein Language Models, and Transformer-based architectures. These future directions respond directly to the reviewers' request for stronger benchmarking and expanded validation.

5. DISCUSSION

The present study introduces a modular artificial intelligence framework that integrates conversational AI, supervised machine learning, and pharmaceutical formulation intelligence to support early-stage therapeutic peptide optimization. Rather than focusing solely on predictive modeling, the proposed platform combines peptide mutation generation, physicochemical feature engineering, explainable machine learning, and an interactive conversational interface within a unified computational environment. This systems-level integration represents the principal contribution of the work and provides a flexible foundation for future AI-assisted peptide discovery.

A distinguishing aspect of the proposed architecture is the separation of conversational reasoning from quantitative prediction. The Large Language Model (LLM) functions as an orchestration and explanation layer that interprets user requests, coordinates computational workflows, and summarizes prediction results. In contrast, numerical inference is performed exclusively by disease-specific Random Forest models. This architectural separation improves transparency and reproducibility while allowing future predictive algorithms to be incorporated without modifying the user interaction paradigm.

The computational evaluation demonstrates that the current GLP-1 regression model can prioritize candidate peptide variants using sequence-derived physicochemical descriptors. However, the observed negative R²R² values indicate that predictive performance remains limited and that the present model does not consistently generalize to unseen data. Several factors likely contribute to this outcome, including the relatively small dataset, heterogeneous experimental measurements obtained from multiple literature sources, limited sequence diversity, and the inherent complexity of peptide structure–activity relationships. Consequently, the platform should be interpreted as a hypothesis-generation and decision-support system rather than as a validated predictive platform. This interpretation is consistent with the reviewer recommendation to moderate claims regarding model accuracy and predictive capability.

The preliminary GLP-1 biodistribution studies further demonstrate that computational peptide prioritization can be integrated with downstream experimental evaluation. Distinct fluorescence distributions were observed following subcutaneous,

intranasal, and intravenous administration, illustrating the influence of formulation and delivery strategy on peptide biodistribution. These experiments provide evidence that the proposed computational workflow can support translational research by facilitating candidate selection prior to laboratory investigation. Nevertheless, the biological studies should not be interpreted as validation of the Random Forest prediction models, as they were designed to evaluate delivery characteristics rather than predictive accuracy.

Overall, the findings demonstrate the technical feasibility of integrating conversational AI, interpretable machine learning, and formulation-aware peptide optimization within a single computational framework. While additional data and prospective validation are required before clinical or regulatory applications can be considered, the proposed platform provides a reproducible foundation for future AI-assisted therapeutic peptide discovery.

6. STUDY LIMITATIONS

Several limitations should be considered when interpreting the results of this study.

First, the predictive performance of the GLP-1 Random Forest model was constrained by the relatively small and heterogeneous dataset available for training and evaluation. The negative R²R² values observed during held-out testing and cross-validation indicate that the current model should be regarded as preliminary rather than mature.

Second, the feature representation relies primarily on sequence-derived physicochemical descriptors. Although these descriptors are interpretable and computationally efficient, they do not fully capture higher-order structural, conformational, or evolutionary relationships that may influence peptide activity.

Third, the biological validation was limited to preliminary fluorescence-based biodistribution studies. These experiments demonstrate integration of computational prioritization with downstream laboratory workflows but do not establish pharmacological efficacy, receptor binding, therapeutic benefit, or predictive accuracy.

Fourth, the Multiple Sclerosis optimization module remains conceptual because an adequately labeled

dataset is not currently available. Consequently, no classification metrics such as accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curves, or area under the curve (AUC) are reported for this application.

Finally, the current manuscript does not include empirical benchmarking against modern deep learning approaches such as Graph Neural Networks, Protein Language Models, or Transformer architectures. Although these methods are discussed conceptually, quantitative comparison remains an important objective for future work.

7. FUTURE WORK

Several opportunities exist to extend and improve the proposed platform.

The highest priority is the expansion of experimentally validated peptide datasets across multiple therapeutic indications. Larger and more diverse datasets are expected to improve model generalizability while enabling more rigorous benchmarking of predictive algorithms.

Future versions of the platform will evaluate Graph Neural Networks, Protein Language Models, Transformer-based architectures, and hybrid ensemble methods capable of learning richer sequence representations than conventional descriptor-based approaches. Comparative benchmarking will be performed using identical preprocessing pipelines and evaluation metrics to determine the relative strengths of each modeling strategy.

Another important direction involves the integration of active learning, in which experimentally generated peptide data are incorporated into iterative model retraining. Such closed-loop optimization may substantially improve predictive performance while reducing the number of laboratory experiments required for candidate identification.

Future work will also expand the explainability framework through incorporation of SHAP values, Local Interpretable Model-Agnostic Explanations (LIME), uncertainty estimation, and confidence calibration. These additions will improve transparency and provide researchers with greater insight into the factors influencing candidate prioritization.

From a translational perspective, prospective biological validation will be essential. Planned investigations include peptide synthesis, receptor

binding studies, pharmacokinetic characterization, toxicity assessment, efficacy testing in disease-specific animal models, and iterative refinement of computational models using experimentally measured outcomes.

Finally, the modular software architecture enables future integration of additional therapeutic indications, digital formulation intelligence, laboratory information management systems (LIMS), and cloud-based computational services. These extensions have the potential to transform the current research prototype into a scalable platform supporting broader applications in peptide engineering and pharmaceutical development.

8. CONCLUSIONS

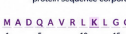
This study presents a modular artificial intelligence framework that integrates conversational AI, supervised machine learning, and pharmaceutical formulation intelligence to support early-stage therapeutic peptide optimization. By combining peptide mutation generation, physicochemical feature engineering, explainable Random Forest models, and a Large Language Model-based conversational interface, the proposed platform provides a unified computational workflow for prioritizing peptide candidates prior to experimental evaluation.

The current implementation demonstrates that conversational AI can improve accessibility and scientific interpretation while preserving reproducible numerical inference through independently trained machine learning models. Although quantitative predictive performance remains limited because of dataset size and biological variability, the platform successfully illustrates the feasibility of integrating computational peptide prioritization with downstream experimental workflows.

Importantly, the present work should be interpreted as a proof-of-concept computational framework rather than a validated predictive system. The preliminary GLP-1 biodistribution studies demonstrate the potential for integrating computational prioritization with biological evaluation but do not constitute validation of the underlying machine learning models. Similarly, the Multiple Sclerosis optimization workflow represents a conceptual architectural extension pending the availability of appropriately labeled datasets for rigorous evaluation.

In conclusion, the proposed framework demonstrates how conversational artificial intelligence, interpretable machine learning, and formulation-aware design can be integrated into a reproducible computational platform that supports hypothesis generation, candidate prioritization, and translational peptide research. While additional validation remains necessary, the platform establishes a scalable foundation for future AI-driven therapeutic peptide discovery and provides a practical pathway toward more efficient and data-informed drug development.

Figure 12 provides a comparative overview of representative machine learning approaches for peptide optimization. The comparison highlights differences in interpretability, data requirements, computational complexity, and suitability for small biomedical datasets. The current platform employs Random Forest models because of their robustness, transparency, and favorable performance on relatively limited datasets, while Graph Neural Networks, Protein Language Models, and Transformer architectures are identified as promising directions for future development.

Aspect	Random Forests (RF)	Graph Neural Networks (GNNs)	Protein Language Models (PLMs)	Transformers (Seq2Seq / Encoder-Decoder)
 Core Idea / Representation	Ensemble of decision trees built on random subsets of features and samples. 	Neural networks operating on graphs to learn representations from relational structure. 	Self-supervised models pretrained on massive protein sequence corpora. 	Attention-based architectures for sequence modeling and generation. 
 Input Representation	Physicochemical descriptors (numerical feature vectors)	Residue/atom graphs, contact maps, 3D structures	Amino acid sequences (tokenized)	Amino acid sequences (tokens)
 Captures	Non-linear relationships between handcrafted features	Structural topology, residue interactions, local and global geometry	Evolutionary patterns, long-range dependencies, contextual embeddings	Long-range dependencies, sequence grammar and generative patterns
 Data Requirement	★★ (Low) Works well with limited data	★★★★ (Medium-High) Needs more data than RF	★★★★★ (High) Benefits from millions of sequences (pretraining mitigates this)	★★★★★ (High) Requires large curated datasets
 Interpretability	★★★★★ (High) Feature importance, intuitive	★★★ (Medium) Explainers available but less intuitive	★★★ (Low-Medium) Embeddings less interpretable	★★★ (Low-Medium) Attention maps provide partial insight
 Training Complexity	★ (Low)	★★★★★ (High)	★★★★★ (High)	★★★★★ (High)
 Computational Cost	★ (Low)	★★★★ (Medium-High)	★★★★★ (High)	★★★★★ (High)
 Strengths	- Robust to overfitting on small datasets - Handles mixed feature types - Provides feature importance - Fast training and inference	- Captures 3D structural information - Models complex residue interactions - State-of-the-art in molecular property prediction	- Learns rich sequence representations - Transferable across tasks - Improves performance with limited labeled data	- Excellent for sequence generation - Captures long-range interactions - Flexible for multiple tasks (design, translation, conditional generation)
 Limitations	- Depends on quality of input features - Limited ability to model sequential/structural context directly	- Requires structural data/graph construction - Higher computational and implementation complexity - Needs larger datasets	- High computational resources - Black-box embeddings - May miss explicit structural context	- Data and compute intensive - Risk of generating non-natural sequences - Less interpretable
 Suitability for Current Dataset Size (small-moderate)	★★★★★ (Very Suitable)	★★★ (Moderate)	★★ (Less Suitable Now, Better with More Data)	★★ (Less Suitable Now)
 Role in This Study	Primary predictive model (Baseline)	Future investigation (Not in current implementation)	Future investigation (Planned for integration)	Future investigation (Planned for design)
 Future Integration Potential	Can be combined in ensembles; serves as a strong interpretable baseline	High – integration with structure prediction (e.g., AlphaFold) and hybrid frameworks	Very High – can provide embeddings for downstream models and fine-tuning	High – for de novo peptide design and conditional optimization

Legend for Ratings: ★ = Very Low ★★ = Low ★★★ = Medium ★★★★ = High ★★★★★ = Very High

Note: Ratings reflect general trends in the literature and the context of therapeutic peptide optimization with small to moderate experimental datasets.

Current Work: Random Forests are used as the primary predictive models due to their reliability on limited datasets, interpretability, and ease of integration. Advanced models will be incorporated as larger datasets and computational resources become available.

Figure 12. Comparative overview of machine learning approaches for peptide optimization

9. REPRODUCIBILITY AND CODE AVAILABILITY

9.1. Code, Environment, and Model Artifacts

The Peptide Optimization Chatbot was developed in Python using scikit-learn, NumPy, pandas, Streamlit, and FastAPI. The implementation is organized into version-controlled modules for peptide mutation generation, descriptor extraction, machine-learning inference, REST API services, and the conversational web interface. Environment specifications (for example, requirements.txt or environment.yml), trained model artifacts, and evaluation scripts are maintained with the project to support consistent execution across computational systems.

9.2 Data and Training Reproducibility

The peptide datasets were assembled from publicly available sequences reported in the literature and relevant peptide databases. Where licensing permits, curated datasets and processed feature matrices can be distributed with the code; otherwise, the original sources are identified so that the datasets can be reconstructed. Model metadata document the training-data characteristics, feature-engineering procedures, hyperparameters, and validation settings used for the GLP-1 workflow.

9.3. Interactive Access and Reproduction Workflow

The interactive application allows researchers to enter peptide sequences, select an optimization workflow, generate candidate variants, and review model-derived scores and rankings. Reproduction

of the computational pipeline requires installation of the documented Python environment, access to the source modules and trained model artifacts, execution of the standardized preprocessing and feature-engineering routines, and inference on a user-provided peptide sequence. This modular organization is intended to make the reported workflow transparent and repeatable while enabling future benchmarking against alternative predictive models.

REFERENCES

- Breiman, L. Random forests. *Machine Learning*, 45(1), 5–32 (2001).
- Brown, N. Artificial intelligence in chemistry and drug design. *Drug Discovery Today*, 23(6), 1241–1250 (2018).
- Brown, T. B., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901 (2020).
- Chollet, F. *Deep Learning with Python*, 2nd Edition. Manning Publications (2021).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* (2019).
- Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 1–15 (2000).
- Elnaggar, A., Heinzinger, M., Dallago, C., et al. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- Erlanson, D. A. Introduction to fragment-based drug discovery. *Topics in Current Chemistry*, 374(3), 1–32 (2016).
- FDA. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. U.S. Food and Drug Administration. Available at: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>
- Fosgerau, K., & Hoffmann, T. Peptide therapeutics: Current status and future directions. *Drug Discovery Today*, 20(1), 122–128 (2015).
- Gainza, P., Sverrisson, F., Monti, F., et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17, 184–192 (2020).
- Goodfellow, I., Bengio, Y., & Courville, A. *Deep Learning*. MIT Press (2016).
- Hamilton, W. L. *Graph Representation Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool (2020).
- Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589 (2021).
- Leeson, P. D. Drug discovery: Chemical beauty contest. *Nature*, 481, 455–456 (2012).
- Leung, M. K. M., Delong, A., Alipanahi, B., & Frey, B. J. Machine learning in drug discovery: A review. *Journal of Chemical Information and Modeling*, 58(4), 632–641 (2018).
- Lundberg, S. M., & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* (2017).
- OpenAI. GPT-4 Technical Report. arXiv:2303.08774 (2023).
- Rao, R., Bhattacharya, N., Thomas, N., et al. Evaluating Protein Transfer Learning with TAPE. *Advances in Neural Information Processing Systems* (2019).
- Ribeiro, M. T., Singh, S., & Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
- Rives, A., Meier, J., Sercu, T., et al. Biological structure and function emerge from scaling protein language models. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118 (2021).
- Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56 (2019).
- Vamathevan, J., Clark, D., Czodrowski, P., et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477 (2019).
- Vaswani, A., Shazeer, N., Parmar, N., et al. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

Wu, Z., Pan, S., Chen, F., et al. A comprehensive survey on Graph Neural Networks. IEEE

Transactions on Neural Networks and Learning Systems, 32(1), 4–24 (2021).



Publisher's note: Eurasia Academic Publishing Group (EAPG) and AAKS remain neutral about jurisdictional claims in published maps and institutional affiliations.

Open Access. This article is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) licence, which permits copy and redistribute the material in any medium or format for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the licence terms. Under the following terms you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorsed you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nd/4.0/>.