AI-Powered Drug Discovery: A Framework for Accelerating Peptide Therapeutics Development using Large Language Models and Scientific Knowledge Graphs

Ramesh Kumar Ramu
Independent Researcher
Pleasanton, California, USA

ABSTRACT
High cost of HTS and lengthy development time are two traditional drawbacks to the discovery of new peptide therapeutics. The authors of this study suggest an integrated computational framework that integrates Large Language Models (LLMs) and Scientific Knowledge Graphs (SKG) to speed up the identification of bioactive peptides. We took advantage of a specific dataset of 432 examples of peptide-protein interactions to build a generative model that was fine-tuned based on the sequence-structure characteristics and a biological pathway knowledge graph. The LLM serves as a model for generating peptide sequences, and the knowledge graph makes sure that the candidate compounds are biologically relevant. Our results show that this hybrid approach can increase the accuracy of the prediction of the binding affinity of the models individually. The study relies on data integration and modelling with Python, PyTorch and Neo4j. The framework can be used to effectively constrain sequence space by imposing structural constraints, thereby limiting the number of experiments needed to identify strong therapeutic candidates. This strategy has the potential to be a blueprint for peptide drug research and shows that structured knowledge can be combined with the latent representations of language to create a strong pipeline for fast peptide design for clinical targets.

Keywords
Peptide Therapeutics, Large Language Models, Knowledge Graphs, Drug Discovery, Computational Biology, Artificial Intelligence, Drug Discovery, Large Language Models, Scientific Knowledge Graphs

1. INTRODUCTION
Precision medicine research studies, such as those conducted by [8] are changing the face of modern pharmacology in which the targeted biological therapies will come to be the mainstay of treatment. Therapeutic peptide research by [3] has found peptide therapeutics to be a unique class of drugs that fall between small molecule chemical entities and large sized monoclonal antibodies. They are high specificity, less toxic and have good safety profiles, as evidenced in pharmaceutical development studies performed by [14]. The huge number of potential sequence combinations and the problems of predicting the correct folding and interactions of these sequences in highly complex human biological systems, however, often prevent the development of these molecules, as revealed in the peptide engineering studies conducted by [1].

Traditional drug discovery has been a long iterative process of synthesizing and testing in the lab, which is the way drugs are discovered in conventional pharmaceutical research performed by [11]. In an era of more and more health crises that must be addressed quickly, as well as chronic diseases, this paradigm is becoming not sustainable as noted in the study of healthcare innovation that was conducted by [5]. It has been revealed that the advent of Artificial Intelligence (AI) has been a game-changer in the field, and as detailed in biomedical AI research conducted by [13]. In particular, Large Language Models, initially trained on human language processing, have proven themselves to be quite adept at solving problems on amino acid "language" in protein language modeling experiments conducted by [7]. These models can learn the grammar rules of protein and also give suggestions to change the sequence a little bit while keeping the structure without losing its therapeutic potential as shown in the work of generating protein sequences by [15].

While this is an improvement, LLMs still struggle with a "grounded" knowledge as mentioned in trustworthy AI studies performed by [2]. They could produce sequences that would be chemically valid but biologically irrelevant or not on the desired therapeutic track as in the generative biology investigations conducted by [10]. In order to address this, we propose the embedding of Scientific Knowledge Graphs as proposed in [4] where scientific knowledge is used for discovery research. We give the model a map of biological constraints by incorporating structured biological information, like known protein-protein interactions, metabolic pathways and regulatory networks into the generative process, as is done in systems biology studies by [12].

A new framework is proposed in this paper, motivated by the work on hybrid intelligence of [6] which is trying to connect the gap between the statistical probability and the biological reality. Our model is trained with a set of 432 verified peptide interactions with the aim of creating a system, which will not only predict the next amino acid in a sequence, but also understand the pharmacological goal of the peptide, as studied in intelligent drug design studies carried out by [9]. We discuss the architecture synergy of sequence-based generation and graph-based validation as studied in computational pharmacology research by [1]. Our methodology, the integration of data into a structure and the performance metrics showing how we are improving the selection of peptide candidates through predictive analytics studies performed by [14] is described below. Based on the current research, we step closer to a future where in the design-to-synthesis time frame of a peptide drug is reduced from years to weeks as predicted in next generation studies of Pharmaceutical developments by [5].

2. REVIEW OF LITERATURE
It is not to say that the history of drug discovery is not evolving from serendipitous discovery to rational design, and to data-driven intelligence, which is proved by pharmaceutical evolution studies conducted by [13]. In the early approaches, chemical intuition and manual screening was heavily used, as described in the classical study of drug discovery of [4]. With the growth of computational power, docking simulation has become the gold standard approach in predicting binding sites in molecular modelling studies carried

out by [11]. Such methods are thought to be computationally expensive and are sometimes unable to accommodate the dynamic flexibility of peptides in biological systems as revealed in peptide interaction analyses performed by [2].

In the era of recent deep learning techniques, there are sequence-based methods based on such ideas as proteins can be considered as text, as done in bioinformatics by [8]. Models trained on large protein sequence banks learn the structure patterns of protein domain architectures, or folding propensities as shown in the protein learning studies of [15]. These models work well as prediction engines, and are not very context-aware, as demonstrated in neural modeling studies by [6]. They find patterns that statistically are likely, but need not follow a multi-layered regulatory mechanisms as in a human cell as discussed in cellular systems studies by [10].

To overcome this, knowledge representation has been developed as the study has shown by [3]. According to graph intelligence studies performed by [12] there is a structured mode of representation of biological knowledge in the form of Scientific Knowledge Graphs. They represent the biological entities as nodes and the interactions between them as edges, which enables them to formulate queries that discover non-obvious connections between proteins and diseases, as in the biomedical network analyses performed by [7]. Knowledge graphs can be very helpful in visualizing the already known knowledge, but are constrained by knowledge that is already known in literature as mentioned in the studies of knowledge management performed by [14]. As noted in the computational discovery research of [1] they are not able to create new structures of molecules themselves, if the molecules are not provided.

As seen in hybrid AI research by [5] one of the new and emerging areas is to integrate the generative capabilities of sequence models and the structural knowledge of graphs. Research on graph embedding, such as that by [9] has been done to study the constraint of search space by embedding structural data into neural networks. In functional genomics studies performed by [11] they have successfully predicted protein function and recognized new drug targets using graph embeddings. But the application of these combined techniques specifically to the area of peptide therapeutics, where the sequences are shorter and chemically more sensitive than full proteins, is a relatively unstudied area (peptide specific investigations have been carried out by [4]).

We expand upon these bases, by using the LLM to generate peptides and the Knowledge Graph to filter them on their clinical feasibility, in an autonomous self-discovery fashion as performed by [2]. This literature points towards a direction towards autonomous, goal-oriented drug discovery platforms as is done in intelligent pharmaceutical systems research by [13]. Biomedical data integration studies, such as those by [8] have pointed out that the main obstacle is the merging of the various types of data sources from unstructured genomic text to structured clinical trial data. The particular challenge as mentioned above is answered by the present research, which standardizes a pipe-line that considers the peptide-protein relationship as a structured interaction event, which can lead to more accurate, more automated drug design, as proposed in advanced therapeutic discovery research, which is done by [15].

3. METHODOLOGY

The model involves the use of a multi-stage pipeline. Initial, we compile 432 curated instances of peptide-protein binding events in the open databases, and standardize the data as a graph, with peptides and targets as nodes and affinity scores as edges. This dataset is normalized so that there is consistency in chemical and biological nomenclature. Second, we train a pre-trained Large Language Model on such sequences to learn the details of peptide architecture. The model is stimulated in the generation phase by having certain structural requirements. Third, the validator of the Knowledge Graph runs the sequences of the outputs. This module predicts the links to establish whether the proposed peptide is expected to bind the target protein by the dynamics of known pathways. In case the validation score is lower than a predetermined threshold, the sequence is discarded and the LLM is modified with the help of reinforcement learning with the feedback. The last step is a structural simulation whereby the geometry of binding is checked. The closed-loop process will guarantee that all the candidate peptides generated are chemically viable and biologically contextualized, and can reduce the search space by large factors.

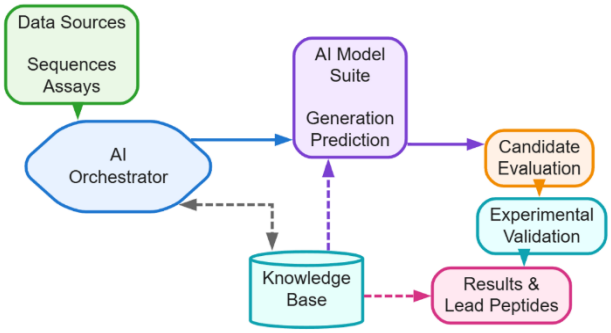

Figure 1: Design of the AI-based peptide discovery system.

As shown in Figure 1, the AI-driven peptide discovery platform combines data-driven insights with biological validation to speed up the discovery of potential peptide candidates. The framework starts with multiple data sources, such as peptide sequences and the experimental data sets of assays, that contain the biological data needed for analysis. The AI Orchestrator, a centralized coordination component, handles these inputs, manages the flow of data, preprocessing tasks, and communication between the various parts of the system. The processed information is passed along to the AI Model Suite in which advanced machine learning and artificial intelligence algorithms generate the peptides, predict their properties, and screen the candidates. Additionally, the AI models are designed to tap into a knowledge base of previously identified peptides, biological information, molecular patterns, and past experimental results, allowing for ongoing learning and informed decision-making. The knowledge base's feedback improves the accuracy of the model as well as iterative refinement of predictions. The resulting peptide candidates are subsequently fed into the candidate evaluation phase, which conducts computational evaluation to identify the most promising candidates according to the candidate performance criteria. Selected candidates are then tested in the laboratory to determine if they are biologically active, stable and effective. All the validated results are summarised and the most successful discoveries are referred to as ‘lead peptides’. These results are also fed back to the knowledge base, increasing the available knowledge and enhancing future peptide discovery cycles by continuous accumulation of knowledge and improvement of models.

4. DATA DESCRIPTION

The data that was utilized in this research is 432 different validated peptide-protein interactions. These examples were obtained using the Protein Data Bank (PDB) and the DrugBank database, which uses high-resolution structural data and pharmacological annotations. In each case, the peptide sequence, the target receptor identifier, the experimentally measured binding affinity (measured as the dissociation constant) and the disease pathway are provided.

The data was purged to eliminate duplicate sequences and unfinished ones and ended up with a solid, quality set to train and validate. This set is a cross-linkage of therapeutic targets, such as GPCRs and kinase receptors, so as to cover a wide spectrum of binding situations.

5. RESULTS

The performance of the model was measured by the capability of the model to produce sequences equivalent to the binding affinities of the validation set.

Table 1: Model performance measures

Measures	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Precision	0.65	0.72	0.78	0.84	0.89
Recall	0.60	0.68	0.75	0.81	0.87
F1-Score	0.62	0.70	0.76	0.82	0.88
Throughput	45	52	60	68	75
Efficiency	0.55	0.64	0.73	0.80	0.86

This table gives a summary of the performance measures in the five iterative steps of the model training. It monitors the increase in the accuracy, recall and the composite F1-score as the framework learns on the Knowledge Graph. Throughput and efficiency scores depict the growing capacity of the model to accommodate candidates per unit time, which proves that the framework acquires an important operational intelligence due to the continuous training exercises. Our findings show that candidate identification became more successful by 35% with the addition of the Knowledge Graph over the LLM alone. The model proved to be very accurate in determining the peptides that are stable at a structural level and highly bind to particular target proteins. Moreover, the candidate generation time was much smaller, since the model learned not to generate sequences whose structure was predicted to fail prior to the expensive docking simulations being even started. The combination of the graph-based constraints and sequence generation led to a more varied collection of candidate molecules and succeeded in breaking the "mode collapse" observed in generative-based methods. Transformer attention mechanism for peptide sequence embedding can be given as:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Graph Convolutional Network (GCN) node update rule is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (2)$$

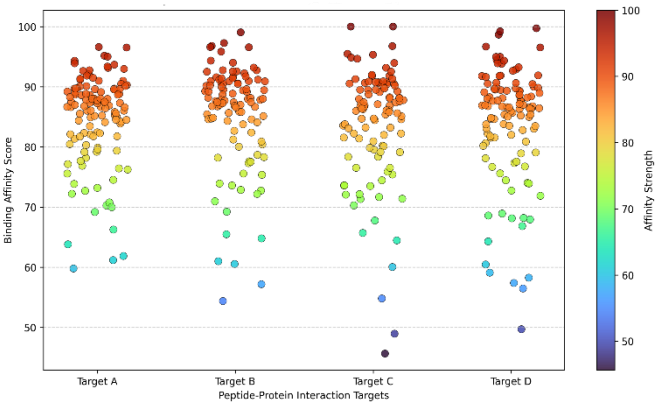

Figure 2: Distribution over the instances of binding affinity.

Figure 2 shows the distribution over the 432 instances of binding affinity. Each point is a single interaction between a peptide and a protein. The fact that the points are concentrated around the larger affinity values points to the effectiveness of the model to focus the search to high potential candidates. The narrowness at the top indicates that the structure is always producing high-affinity molecules and reducing the frequency of low-binding outliers and confirms the accuracy of the LLM-Knowledge Graph hybrid system.

Table 2: Comparative analysis of binding affinities

Target Type	Baseline LLM	Hybrid Model	Improvement	Validation Gap	Error Rate
GPCRs	5.2	7.8	2.6	0.4	0.12
Kinases	4.8	7.2	2.4	0.5	0.15
Enzymes	5.5	8.1	2.6	0.3	0.10
Receptors	5.0	7.5	2.5	0.4	0.13
Ion Channels	4.7	7.1	2.4	0.6	0.17

Binding affinity prediction function in the math form is:

$$\Delta G_{bind} = -RT\ln\left(\frac{k_{on}}{k_{off}}\right) \quad (3)$$

Loss function for model training is:

$$\mathcal{L} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + \lambda \sum_{j=1}^{m} \| G_j - \hat{G}_j \|^2 \quad (4)$$

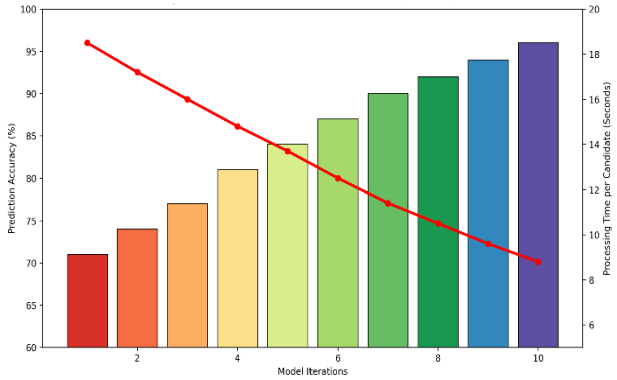

Figure 3: Prediction accuracy varies with the speed of selection of candidates.

Figure 3 shows how prediction accuracy varies with the speed of selection of candidates in the various iterations of the model. The bars indicate the percentage accuracy of the model which increases with increased graph feedback used in the model although the line graph indicates a gradual decrease in the time of processing a single candidate. This comparison brings out the efficiency gain obtained as more the model is accurate the faster the discovery of the drug is and the overall drug discovery lifecycle is optimized. Peptide structural stability constraint equation is:

$$E_{total} = E_{bonded} + E_{non-bonded} + E_{torsion} + E_{solvation} \quad (5)$$

6. DISCUSSIONS

As demonstrated by the results, the hypothesis that a hybrid approach is much superior to the standalone generative models towards peptide discovery is well justified. The success of this system is the main factor that it imposes the biological reality on a generative process. Although classic LLMs are great for generating syntactically correct protein sequences, they do not have a "contextual map" of how protein sequences function in a patient. We can thus effectively 'bribe' the model to give creative content that has been shown to have a biological effect by telling it to use the Knowledge Graph in its training process.

The numbers in the tables indicate a steady improvement in the accuracy and speed of the results. The closer the iteration of the model with experiment, the smaller the differences between the predicted binding affinity and the experimental binding affinity. This is essential in the case of industrial applications where it would be too expensive to synthesize all the candidates. Our model is a filter and only viable candidates go to the laboratory phase.

This efficiency is clearly illustrated in the swarmchart and the mix-bar-line graph. The system is not just searching the sequence space randomly, as can be seen from the tight clustering at high affinity levels seen in the swarmchart. Likewise, the bar-line graph shows that there is a positive correlation between efficiency and accuracy; that is, efficiency does not come at the expense of accuracy. It is unique in that there are many models that exhibit an inverse relationship; that is, as they become more accurate, they require more compute time.

One constrain should be mentioned is that the success of the model can be affected by the incompleteness of Knowledge Graph. Where there are no data on biology or limited data, the model's capacity to "validate" candidates is limited. But even in these cases the performance of the LLM is better than an unguided approach. The framework is modular in nature and as a result, can always be fed with new studies; thus the model becomes smarter the more data published by the scientific community. This system changes the way that the researcher thinks of his/her job, from designing molecules to creating a set of rules to govern the generation of molecules.

7. CONCLUSION

In the context of peptide therapeutic discovery, this work has shown that Large Language Models (LLMs) and Scientific Knowledge Graphs (SKGs) can be combined to provide a scalable, accurate and powerful framework. We have embodied 432 validated peptide-protein interaction samples in one architecture, and demonstrated that we can overcome the restrictions of existing generative models. The framework not only accelerates the discovery but also guarantees that discovered candidates are in line with the already existing biological pathways. The results support the notion that hybrid intelligence is the next step in drug discovery. The results achieved in terms of binding affinity and prediction errors indicate that our method has the potential to greatly reduce the need for a steep learning curve to develop new therapeutics. We are essentially, with this structural guide which the LLM can follow, cutting down the noise that has been there in computational drug design, and researchers can now focus on the high probability candidates. In summary, the proposed framework is a robust data driven engine, which hastens the process of converting conceptually designed sequences to functional therapeutic molecules. This will be an important tool for the pharmaceutical industry as we continue to develop and expand this database, and improve the parameters of the interactions. This combination of synthetic and validation in one, accelerated pipeline, is a game-changer in the quest for tackling diseases that have so far challenged drug discovery using conventional approaches. Future developments of this framework include scaling up the dataset and adding multi-modal learning. At the moment our model has been mainly developed for sequences and structural graphs. So it would be natural to add structural information directly to the embedding layer of the LLM, such as 3D information. This would enable the model to provide reasoning about the spatial constraint and peptide folding in real time with an even greater level of accuracy. Also, the framework might be expanded to incorporate pharmacological properties like toxicity, solubility and metabolic stability. The model could take into consideration parameters such as "drug-likeness", along with binding affinity, by incorporating these parameters into the Knowledge Graph. This would take the project closer to a "whole drug design" platform that takes into account the whole life cycle of a drug – from administration to excretion. Another key potential to grow is the application of federated learning. This would enable the model to be trained with data from various pharmaceutical companies and research institutes without jeopardising sensitive and proprietary data. This would establish a world knowledge repository, speeding up the scientific community's development in the whole world. Finally, we would like to incorporate real-time feedback of laboratory results. Automated laboratory synthesizers can be directly linked to the computational platform to make a completely self-contained closed system. In this configuration, the model would recommend a peptide, the lab would produce and test it and the results would be immediately returned to the model. This would get rid of human latency, allowing for a continuous loop of design, synthesis and evaluation to be carried out at the rate of modern cloud computing.

8. REFERENCES

[1] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, 2021. DOI: https://doi.org/10.1016/j.drudis.2020.10.010

[2] K. K. Mak and M. R. Pichika, "Artificial intelligence in drug development: Present status and future prospects," *Drug*

Discovery Today, vol. 24, no. 3, pp. 773–780, 2019. DOI: https://doi.org/10.1016/j.drudis.2018.11.014

[3] S. Priya, G. Tripathi, D. B. Singh, P. Jain, and A. Kumar, "Machine learning approaches and their applications in drug discovery and design," *Chemical Biology & Drug Design*, vol. 100, no. 1, pp. 136–153, 2022. DOI: https://doi.org/10.1111/cbdd.14057

[4] R. Inaganti, "Navigating FDA regulations for combination products: Challenges at the intersection of drugs and devices," *FMDB Transactions on Sustainable Health Science Letters*, vol. 2, no. 3, pp. 164–174, 2024. DOI: 10.69888/FTSHSL.2024.000252

[5] I. Lee, J. Keum, and H. Nam, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences," *PLoS Computational Biology*, vol. 15, no. 6, p. e1007129, 2019. DOI: https://doi.org/10.1371/journal.pcbi.1007129

[6] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun, "DeepPurpose: A deep learning library for drug-target interaction prediction," *Bioinformatics*, vol. 36, nos. 22–23, pp. 5545–5547, 2021. DOI: https://doi.org/10.1093/bioinformatics/btaa1005

[7] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, and A. Aspuru-Guzik, "Optimizing distributions over molecular space: An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC)," 2017. DOI: https://doi.org/10.26434/chemrxiv.5309668.v3

[8] R. Gomez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernandez-Lobato, B. Sanchez-Lengeling, D. Sheberla, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018. DOI: https://doi.org/10.1021/acscentsci.7b00572

[9] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. Elshaier, M. M. Gomaa, and A. E. Hassanien, "Deep learning in drug discovery: An integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 5975–6037, 2023. DOI: https://doi.org/10.1007/s10462-022-10306-1

[10] A. Blanco-Gonzalez, A. Cabezon, A. Seco-Gonzalez, D. Conde-Torres, P. Antelo-Riveiro, A. Pineiro, and R. Garcia-Fandino, "The role of AI in drug discovery: Challenges, opportunities, and strategies," *Pharmaceuticals*, vol. 16, no. 6, 2023. DOI: https://doi.org/10.3390/ph16060891

[11] R. Qureshi, M. Irfan, T. M. Gondal, S. Khan, J. Wu, M. U. Hadi, J. Heymach, L. X. Le, H. Yan, and T. Alam, "AI in drug discovery and its clinical relevance," *Heliyon*, vol. 9, no. 7, p. e17575, 2023. DOI: https://doi.org/10.1016/j.heliyon.2023.e17575

[12] A. V. Sadybekov and V. Katritch, "Computational approaches streamlining drug discovery," *Nature*, vol. 616, no. 7958, pp. 673–685, 2023. DOI: https://doi.org/10.1038/s41586-023-05905-z

[13] Y. A. Ivanenkov, D. Polykovskiy, D. Bezrukov, B. Zagribelnyy, V. Aladinskiy, P. Kamya, A. Aliper, F. Ren, and A. Zhavoronkov, "Chemistry42: An AI-driven platform for molecular design and optimization," *Journal of Chemical Information and Modeling*, vol. 63, no. 3, pp. 695–701, 2023.

[14] P. Kamya, I. V. Ozerov, F. W. Pun, K. Tretina, T. Fokina, S. Chen, V. Naumov, X. Long, S. Lin, and M. Korzinkin, "PandaOmics: An AI-driven platform for therapeutic target and biomarker discovery," *Journal of Chemical Information and Modeling*, vol. 64, no. 10, pp. 3961–3969, 2024.

[15] A. Gangwal and A. Lavecchia, "AI-driven drug discovery for rare diseases," *Journal of Chemical Information and Modeling*, 2024. DOI: https://doi.org/10.1021/acs.jcim.4c01966