Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity?

Markus Orsi,[a] and Jean-Louis Reymond[a]*

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

e-mail: *jean-louis.reymond@unibe.ch*

Abstract

Antimicrobial peptides (AMPs) are naturally occurring or designed peptides up to a few tens of amino acids which may help address the antimicrobial resistance crisis. However, their clinical development is limited by toxicity to human cells, a parameter which is very difficult to control. Given the similarity between peptide sequences and words, large language models (LLMs) might be able to predict AMP activity and toxicity. To test this hypothesis, we fine-tuned LLMs using data from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP). GPT-3 performed well but not reproducibly for activity prediction and hemolysis, taken as a proxy for toxicity. The later GPT-3.5 performed more poorly and was surpassed by recurrent neural networks (RNN) trained on sequence-activity data or support vector machines (SVM) trained on MAP4C molecular fingerprint-activity data. These simpler models are therefore recommended, although the rapid evolution of LLMs warrants future re-evaluation of their prediction abilities.

Keywords: large language models, LLM, GPT, hemolysis, activity prediction, antimicrobial peptides

Introduction

Antimicrobial peptides (AMPs) have gained significant attention in the field of drug discovery due to their potential therapeutic applications in the fight against antimicrobial resistance.[1–3] However, the vast number of possible peptide sequences and their complex structure-activity relationship landscape mean that it is difficult to rationally design peptides with the desired biological activity, in particular tuning their activity versus toxicity to human cells, which is often measured as hemolysis of human red blood cells.[4,5]

To address this issue, several machine-learning models have been developed for the *de novo* design of antimicrobial peptides.[6–21] Because property prediction from a peptide sequence can be framed as a natural language processing problem, many of these models use architectures specifically designed for language processing tasks.[22–24] Furthermore, the emergence of large language models (LLMs), such as OpenAI's GPT models,[25] has opened new possibilities for leveraging powerful language processing capabilities in drug discovery applications. Recent attempts by Jablonka *et al*. to explore the capabilities of GPT-3 for predicting properties of small molecules in various applications have shown that GPT-3 was able to perform comparably or even outperform conventional statistical models, particularly in the low data regime.[26] There also have been successful efforts into augmenting LLM capabilities to tackle tasks related to small molecule chemistry in the areas of organic synthesis, drug discovery, and materials design.[27–30] Hereby, the models mainly orchestrate a set of tools to solve chemistry tasks starting from a natural language prompt.[31–33] However, to the best of our knowledge LLMs have not been implemented to predict the bioactivity of peptides yet.

In this study, we aimed to compare GPT models fine-tuned on antimicrobial peptide sequence data with models that have been previously used to predict antimicrobial activity and hemolysis of peptide sequences.[13,14] Alongside evaluating the performance of the fine-tuned GPT models, we also seek to explore the advantages and disadvantages they offer in terms of time and cost effectiveness. Furthermore, we compare the performance of models trained on amino acid sequences to a support-vector machine (SVM) trained on the MAP4C fingerprint.[34]

Methods

Datasets

The datasets used in this study were peptide sequences with annotated antimicrobial and hemolytic activity collected from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP).[13,35] The dataset used for the classification tasks contained 9,548 (7,160 training / 2,388 validation) sequences with annotated antimicrobial activity, of which 2,262 (1,723 training / 539 validation) sequences had additional hemolytic activity annotations. To test models in low data regimes, we randomly selected subsets from the original training sets, representing approximately 20% and 2% of the original activity set, and approximately 10% of the original hemolysis set. All datasets are further described in **Table 1**. To ensure consistency, we maintained the same training and test split for all initial evaluations. For the detailed study, we used the same 5-fold cross-validation sets.

Table 1: Sizes and composition of the datasets used in the present study. Datasets are available at https://github.com/reymond-group/LLM_classifier.

Name	Size	# Actives / Not Hemolytic	# Inactives / Hemolytic
Activity Training	7,160	3,580	3,580
Activity Training 20%	1,400	701	699
Activity Training 2%	140	74	66
Activity Validation	2,388	1,194	1,194
Hemolysis Training	1,723	717	1,006
Hemolysis Training 10%	170	65	105
Hemolysis Validation	539	226	313

Models

As reference models, we used our previously reported Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Recurrent Neural Network (RNN) classifiers trained on the same data.[13] We furthermore trained two additional SVM models on alternative representations of peptide sequences: one utilizing the MAP4C fingerprint[34] with a custom Jaccard kernel, and another using predicted fraction of helical residues and hydrophobic moment with a linear kernel. Fraction of helical residues were predicted using SPIDER3.[36] Hydrophobic moment was computed using the method of Eisenberg *et al*.[37]

To explore the potential of GPT-3 models for antimicrobial and hemolytic activity classification, we performed fine-tuning of the Ada, Babbage, and Curie models which were accessible through the OpenAI API (v0.28.0, accessed between 25.05.2023 and 01.06.2023). The fine-tuning process involved training each model using the full, 20% and 2% sets for activity classification and the full and 10% set for the hemolysis classification. In the later evaluation with the more advanced LLM GPT-3.5 Turbo, fine-tuning was also performed via OpenAI's Python API (v1.11.1), following the provided guidelines, but we restricted ourselves to the full model. The utilized fine-tuning datasets contained a system role ("predicting antimicrobial activity/hemolysis from an amino acid sequence"), a user message (peptide sequence formatted as "SEQUENCE ->"), and a system message ("0" for negative labels and "1" for positive labels).

Metrics

All models were evaluated using five commonly accepted performance metrics: ROC AUC, Accuracy, Precision, Recall and F1. Metrics were either calculated using the scikit-learn (v1.4.0) Python (v3.12.1) package (reference models and GPT-3.5) or directly obtained from the OpenAI platform after fine-tuning was completed (for all GPT-3 models).

ROC AUC (Receiver Operating Characteristic Area Under the Curve: The ROC AUC measures the area under the Receiver Operating Characteristic curve, which plots the True Positive

Rate (Sensitivity) against the False Positive Rate. A higher ROC AUC value (ranging from 0 to 1) indicates better discrimination and predictive performance of the model.

Accuracy: Accuracy measures the overall correctness of the model's predictions, calculating the ratio of correctly classified instances to the total number of instances. It provides a general understanding of the model's performance but can be misleading in imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Precision: Precision measures the proportion of true positives out of all predicted positives. It focuses on the model's ability to avoid false positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall measures the proportion of true positives out of all actual positives. It represents the model's ability to identify positive instances accurately.

$$Recall = \frac{TP}{TP + FN}$$

F1 score: F1 is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Results and Discussion

Model screening

Starting from the DBAASP dataset of 9,548 peptide sequences annotated with antibacterial activity and 2,262 peptide sequences annotated with hemolysis effect, we had previously evaluated NB, RF, SVM and RNN models, and found the latter to perform best for predicting both activity and hemolysis from sequence data.[13,14] For additional reference, we trained an SVM on the fraction of helical residues and the hydrophobic moment, two properties commonly known to correlate with antimicrobial activity, as well as another SVM on MAP4C, a molecular fingerprint that can reliably encode large molecules such as natural products and peptides including their chirality,[34] a parameter which we considered important since our data listed sequences containing both L- and D-amino acids.

Aiming to test how LLMs perform in predicting antimicrobial activity and hemolysis, we first fine-tuned and evaluated GPT-3 Ada, Babbage, and Curie models. As discussed in our preprint, these models performed slightly better than the reference models, and even provided good performances when trained in low data regime (20% and 2% of full data). However, these models were later deprecated by OpenAI and their performance cannot be reproduced. We therefore discuss herein only the results obtained with the more recent GPT-3.5 model, in comparison with the reference models.

For both, prediction of antimicrobial activity and prediction of hemolysis, the top-performing models were the MAP4C SVM and the RNN model trained on sequence data, the latter being the best performer in our original work (**Table 2**).[13] The performances for both models were in a similar range, although the RNN displayed a notably higher ROC-AUC in both tasks. GPT-3.5 displayed the highest recall performance among the activity models, indicative of the model's tendency to overly favor positive predictions, potentially leading to increased false positive predictions. On the other hand, the features SVM trained only on helicity and hydrophobic moment did not perform significantly above background, and was later used as a negative control model.

Table 2. Performance metrics of all models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold.

Model	ROC AUC	Accuracy	Precision	Recall	F1
NB act.	0.55	0.55	0.59	0.32	0.42
RF act.	0.81	0.71	0.7	0.75	0.73
SVM act.	0.75	0.68	0.68	0.68	0.68
RNN act.	**0.84**	0.76	0.74	0.8	0.77
Features SVM act.	0.65	0.65	0.66	0.62	0.64
MAP4C SVM act.	0.8	**0.8**	**0.79**	0.83	**0.8**
GPT-3.5 Turbo act.	0.68	0.68	0.62	**0.93**	0.75
NB hem.	0.58	0.56	0.48	0.76	0.59
RF hem.	0.8	0.77	**0.81**	0.6	0.69
SVM hem.	0.69	0.73	0.72	0.58	0.65
RNN hem.	**0.87**	0.76	0.7	0.76	0.73
Features SVM hem.	0.62	0.63	0.57	0.5	0.54
MAP4C SVM hem.	0.83	**0.83**	0.76	**0.85**	**0.8**
GPT-3.5 Turbo hem.	0.65	0.69	0.72	0.43	0.54

Model comparison

Following the initial model screening, we aimed to validate our findings through a more robust approach: a 5-fold cross-validation involving GPT-3.5, the MAP4C SVM, the RNN, and finally the features SVM as negative control. For this purpose, we generated five data splits and conducted predictions anew.

The results, depicted in **Figure 1a** for antimicrobial activity prediction and **Figure 1b** for hemolysis prediction, confirmed our earlier observations (performances in **Table S2)**. Notably, the RNN performances were higher than those observed in the screening experiment, and were clearly above those of GTP-3.5. Furthermore, both the RNN and MAP4C SVM demonstrated comparable performances, indicating the validity of both approaches in predicting antimicrobial activity and hemolysis. The finding that simpler machine learning architectures, like SVM, can rival the performance of more complex RNNs in predicting antimicrobial activity and hemolysis is particularly interesting. A comparison with models trained on similar datasets, which achieve similar performances as reported in this study, further reinforces the consistency of our findings.[19–21]

This raises questions about the importance of model architecture versus foundational elements such as data quality and feature engineering. It suggests that a balanced approach, prioritizing optimization of these foundational components, could prove more beneficial than focusing solely on model complexity.

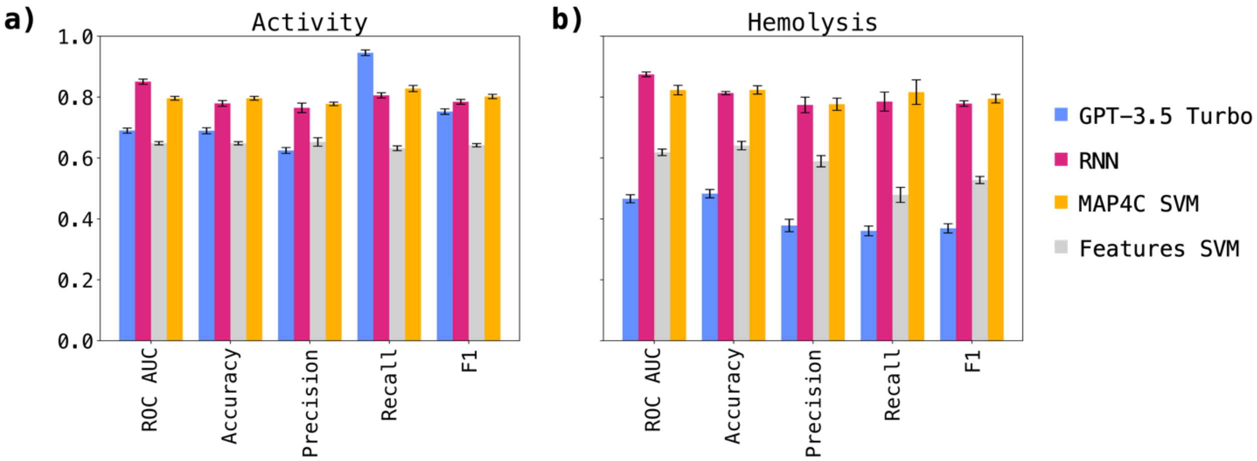

Figure 1: Results of the 5-fold cross-validation study aimed at validating MAP4C SVM, Features SVM, RNN, and GPT-3.5 turbo performance for **a)** antimicrobial activity and **b)** hemolysis predictions. The mean performance across the 5 cross-validations for each metric is shown as a bar, the standard deviation is displayed with an error bar. The results confirmed earlier observations but showed notably higher performances for the RNN compared to the one-shot screening experiment. Both the RNN and MAP4C SVM demonstrated comparable performances.

Data visualization

The high performance achieved by the SVM trained on the MAP4C fingerprint suggested that the nearest neighbor relationships in the MAP4C feature space could be sufficient to distinguish active from inactive and hemolytic from non-hemolytic peptide sequences. In our previous work, we observed that the MAP4 fingerprint[38] correctly clustered natural products, taken from the COCONUT database,[39] according to their organism of origin.[40,41] In analogy to our previous work, we were curious to see whether a spatial separation of actives/inactives and hemolytic/non-hemolytic sequences can be obtained from encoding with MAP4C, the chiral version of MAP4, possibly explaining the good performance of the MAP4C SVM model. For this, we reduced the 2048-dimensional feature space of MAP4C to 2D using the dimensionality reduction method TMAP,[42] and used the obtained visualization to display a set of molecular properties.

First, we wanted to confirm that the TMAP visualization aligns with intuitive distributions of structural features relevant for peptides. For that, we colored the data points based on their heavy atom count (HAC), an indicator of molecular size, and fraction of carbon atoms (fraction C), a simple proxy for the hydrophobicity of a peptide sequence. The TMAP revealed visible clusters for both, HAC (**Figure 2a**) and fraction C (**Figure 2b**), indicating that the reduced MAP4C features can reliably represent simple molecular descriptors in the underlying chemical space.

Following this first observation, we wanted to test if we can detect clusters within TMAP visualizations of more complex physicochemical properties, such as the predicted fraction of helical residues (**Figure 2c**) and the hydrophobic moment (**Figure 2d**). In both cases, we could not detect large homogenous clusters as was the case for HAC and fraction C. However, the data formed a large number of small local clusters, indicating that the nearest neighbor relationships in the MAP4C feature space can possibly be used to distinguish sequences with high helicity/hydrophobicity opposed to sequences with low helicity/hydrophobicity.

Finally, we analysed the distribution of active versus inactive (**Figure 2e**) and hemolytic versus non-hemolytic (**Figure 2f**) sequences in the MAP4C chemical space. Similarly to the visualizations of predicted fraction of helical residues and hydrophobic moment, active and inactive or hemolytic and non-hemolytic sequences are spatially separated in a large number of small, local clusters. This finding is particularly interesting as it suggests that nearest neighbor relationships in the MAP4C feature space are sufficient to separate peptide sequences based on their antimicrobial activity and hemolysis. It further provides an explanation to the good performance obtained with the

MAP4C SVM, which can leverage the nearest neighbor relationships stored in the MAP4C fingerprint feature space when provided with a custom Jaccard kernel function.

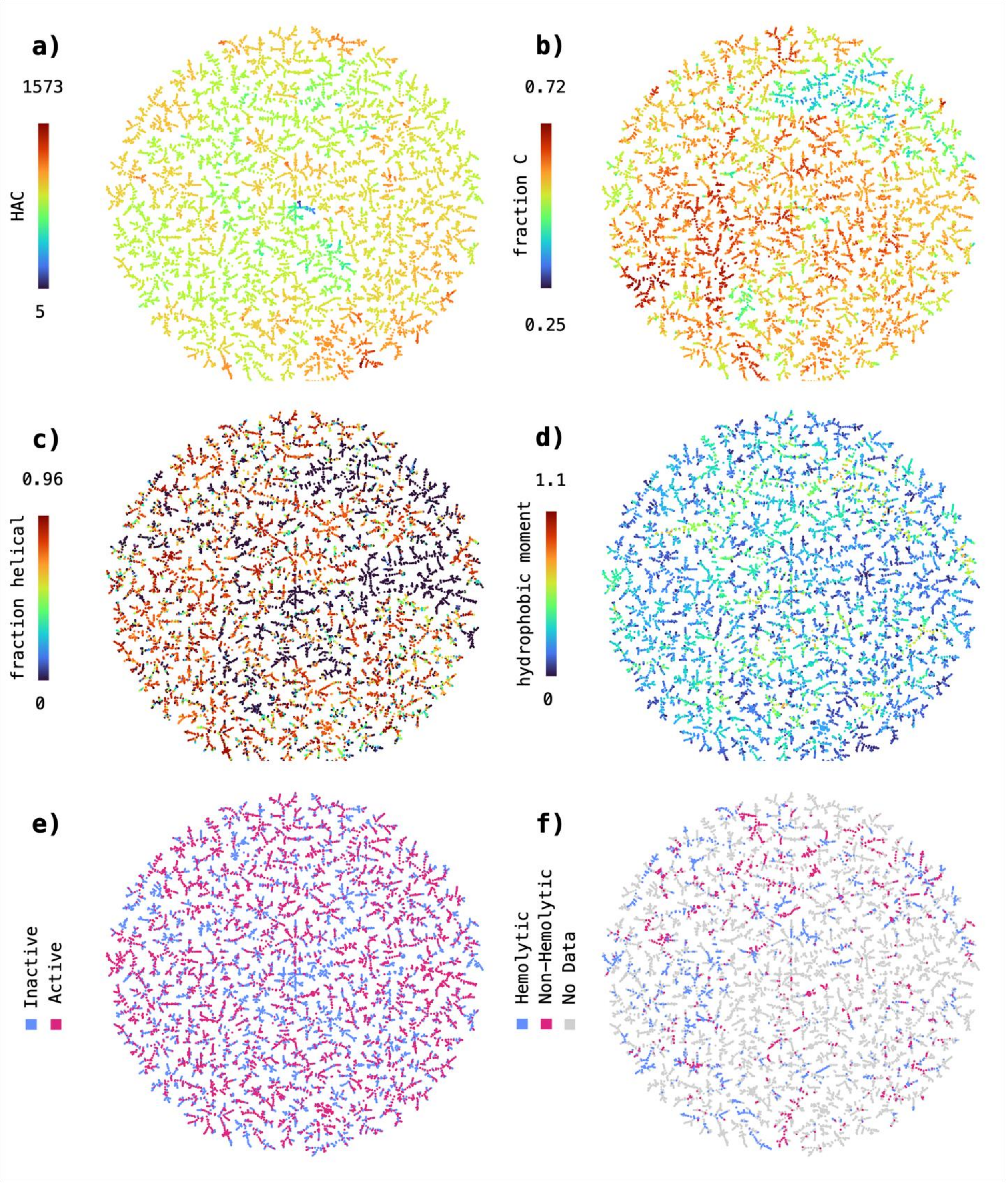

Figure 2: Chemical space covered by the 9,548 peptide sequences with annotated antimicrobial activity extracted from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP). The sequences are encoded using the MAP4C fingerprint and the resulting 2048-dimensional space reduced to 2D using TMAP. The sequences in the 2D TMAP were colored based on a) heavy atom count, b) fraction of carbon atoms, c) predicted fraction of helical residues, d) hydrophobic moment, e) annotated antimicrobial activity and f) annotated hemolysis.

Conclusion

In the present study we investigated the potential of LLMs as predictive tools for antimicrobial activity and hemolysis of peptide sequences. We assessed that fine-tuning GPT models in cloud is a relatively easy and fast process as access through the API eliminates the need to buy expensive hardware and requires little technical expertise. Duration of fine-tuning was short, and the associated costs were low (**Table S3**). In contrast to cloud-based fine-tuning, local model training involves setting up and maintaining hardware, which can be costly and require technical expertise. While less complex models like RNNs and SVMs have lower hardware requirements, training larger models such as LLMs locally can pose challenges in terms of scalability, as one can rapidly face limitations in terms of hardware capacity and maintenance costs.

However, the lack of control over the training environment in cloud-based approaches raises concerns regarding reproducibility of scientific results. In the course of this study, we had originally fine-tuned GPT-3 models Ada, Babbage and Curie. These models performed slightly better than the reference models, even achieving good performances in low data regimes. Unfortunately, these models were later deprecated by OpenAI and their performance cannot be reproduced. When fine-tuning a newer iteration of GPT-3 (GPT-3.5 Turbo), we observed a significant decrease in performance for the same task. We attribute the drop in performance to the increasing optimization of LLMs for conversational interactions, which may negatively impact their effectiveness in out-of-scope predictive tasks. These findings highlight the potential risk of how not controlling one's own models can compromise the reproducibility and reliability of scientific results.

The aforementioned findings suggest a diminishing suitability of chat oriented LLMs for classification tasks over time, a function beyond their intended design. This observation specifically applies to LLMs tailored for conversational or human interaction purposes, rather than specialized LLMs trained on domain-specific data. Unfortunately, the latter do not provide the ease of access and usability that GPT models do. Consequently, we expect that LLMs will increasingly be employed in human interaction settings, facilitating the integration of various chemical tools through natural language interfaces as is being pioneered by Bran[31] and Boiko *et al*.[32]

Finally, we could demonstrate in the present study that classical machine learning techniques, such as SVMs trained on MAP4C fingerprint encodings, can achieve state-of-the-art performance in the prediction of antimicrobial activity and hemolysis. This finding is especially interesting, as it showcases that good performance can be achieved by less complex models, putting the emphasis on data quality rather than model complexity.

Code availability

The source codes and datasets used for this study are available at https://github.com/reymond-group/LLM_classifier.

Author Contribution Statement

MO designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper. Both authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076). MO thanks Sacha Javor for the helpful discussion and comments.

References

(1) Lakemeyer, M.; Zhao, W.; Mandl, F. A.; Hammann, P.; Sieber, S. A. Thinking Outside the Box-Novel Antibacterials To Tackle the Resistance Crisis. *Angew. Chem. Int. Ed.* **2018**, *57* (44), 14440–14475. https://doi.org/10.1002/anie.201804971.

(2) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; Cherkasov, A.; Seleem, M. N.; Pinilla, C.; De La Fuente-Nunez, C.; Lazaridis, T.; Dai, T.; Houghten, R. A.; Hancock, R. E. W.; Tegos, G. P. The Value of Antimicrobial Peptides in the Age of Resistance. *Lancet Infect. Dis.* **2020**, *20* (9), e216–e230. https://doi.org/10.1016/S1473-3099(20)30327-3.

(3) Mookherjee, N.; Anderson, M. A.; Haagsman, H. P.; Davidson, D. J. Antimicrobial Host Defence Peptides: Functions and Clinical Potential. *Nat. Rev. Drug Discov.* **2020**, *19* (5), 311–332. https://doi.org/10.1038/s41573-019-0058-8.

(4) Torres, M. D. T.; Sothiselvam, S.; Lu, T. K.; De La Fuente-Nunez, C. Peptide Design Principles for Antimicrobial Applications. *J. Mol. Biol.* **2019**, *431* (18), 3547–3567. https://doi.org/10.1016/j.jmb.2018.12.015.

(5) Capecchi, A.; Reymond, J.-L. Peptides in Chemical Space. *Med. Drug Discov.* **2021**, *9*, 100081. https://doi.org/10.1016/j.medidd.2021.100081.

(6) Müller, A. T.; Hiss, J. A.; Schneider, G. Recurrent Neural Network Model for Constructive Peptide Design. *J. Chem. Inf. Model.* **2018**, *58* (2), 472–479. https://doi.org/10.1021/acs.jcim.7b00414.

(7) Veltri, D.; Kamath, U.; Shehu, A. Deep Learning Improves Antimicrobial Peptide Recognition. *Bioinformatics* **2018**, *34* (16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179.

(8) Liu, S. Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides. *Sci. Rep.* **2018**.

(9) Su, X.; Xu, J.; Yin, Y.; Quan, X.; Zhang, H. Antimicrobial Peptide Identification Using Multi-Scale Convolutional Network. *BMC Bioinformatics* **2019**, *20* (1), 730. https://doi.org/10.1186/s12859-019-3327-y.

(10) Vishnepolsky, B.; Zaalishvili, G.; Karapetian, M.; Nasrashvili, T.; Kuljanishvili, N.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M.; Grigolava, M.; Pirtskhalava, M. De Novo Design and In Vitro Testing of Antimicrobial Peptides against Gram-Negative Bacteria. **2019**.

(11) Plisson, F.; Ramírez-Sánchez, O.; Martínez-Hernández, C. Machine Learning-Guided Discovery and Design of Non-Hemolytic Peptides. *Sci. Rep.* **2020**, *10* (1), 16581. https://doi.org/10.1038/s41598-020-73644-6.

(12) Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H. K.; Wong, K. H.; Siu, S. W. I. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther. - Nucleic Acids* **2020**, *20*, 882–894. https://doi.org/10.1016/j.omtn.2020.05.006.

(13) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem. Sci.* **2021**, *12* (26), 9221–9232. https://doi.org/10.1039/D1SC01713F.

(14) Zakharova, E.; Orsi, M.; Capecchi, A.; Reymond, J. Machine Learning Guided Discovery of Non-Hemolytic Membrane Disruptive Anticancer Peptides. *ChemMedChem* **2022**. https://doi.org/10.1002/cmdc.202200291.

(15) Liu, G.; Catacutan, D. B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J. C.; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, J. M. Deep Learning-Guided Discovery of an Antibiotic Targeting Acinetobacter Baumannii. *Nat. Chem. Biol.* **2023**. https://doi.org/10.1038/s41589-023-01349-8.

(16) Wan, F.; De La Fuente-Nunez, C. Mining for Antimicrobial Peptides in Sequence Space. *Nat. Biomed. Eng.* **2023**. https://doi.org/10.1038/s41551-023-01027-z.

(17) Aguilera-Puga, M. D. C.; Plisson, F. *Structure-Aware Machine Learning Strategies for Antimicrobial Peptide Discovery*; preprint; In Review, 2024. https://doi.org/10.21203/rs.3.rs-3938402/v1.

(18) Wan, F.; Wong, F.; Collins, J. J.; De La Fuente-Nunez, C. Machine Learning for Antimicrobial Peptide Identification and Design. *Nat. Rev. Bioeng.* **2024**. https://doi.org/10.1038/s44222-024-00152-x.

(19) Timmons, P. B.; Hewage, C. M. HAPPENN Is a Novel Tool for Hemolytic Activity Prediction for Therapeutic Peptides Which Employs Neural Networks. *Sci. Rep.* **2020**, *10* (1), 10869. https://doi.org/10.1038/s41598-020-67701-3.

(20) Hasan, M. M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: Improved and Robust Prediction of Hemolytic Peptide and Its Activity by Fusing Multiple Feature Representation. *Bioinformatics* **2020**, *36* (11), 3350–3356. https://doi.org/10.1093/bioinformatics/btaa160.

(21) Ansari, M.; White, A. D. Serverless Prediction of Peptide Properties with Recurrent Neural Networks. *J Chem Inf Model* **2023**.

(22) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

(23) Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv October 7, 2014. http://arxiv.org/abs/1409.1259 (accessed 2023-05-31).

(24) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv December 5, 2017. http://arxiv.org/abs/1706.03762 (accessed 2023-05-31).

(25) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. arXiv July 22, 2020. http://arxiv.org/abs/2005.14165 (accessed 2023-05-31).

(26) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging Large Language Models for Predictive Chemistry. *Nat. Mach. Intell.* **2024**, *6* (2), 161–169. https://doi.org/10.1038/s42256-023-00788-1.

(27) Bran, A. M.; Schwaller, P. Transformers and Large Language Models for Chemistry and Drug Discovery. **2023**. https://doi.org/10.48550/ARXIV.2310.06083.

(28) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks.

(29) Castro Nascimento, C. M.; Pimentel, A. S. Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *J. Chem. Inf. Model.* **2023**, *63* (6), 1649–1655. https://doi.org/10.1021/acs.jcim.3c00285.

(30) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Peña Ccoa, W. J. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digit. Discov.* **2023**, *2* (2), 368–376. https://doi.org/10.1039/D2DD00087C.

(31) Bran, A. M.; Cox, S.; White, A. D.; Schwaller, P. ChemCrow: Augmenting Large-Language Models with Chemistry Tools. arXiv April 12, 2023. http://arxiv.org/abs/2304.05376 (accessed 2023-05-31).

(32) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.

(33) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; De Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digit. Discov.* **2023**, *2* (5), 1233–1250. https://doi.org/10.1039/D3DD00113J.

(34) Orsi, M.; Reymond, J.-L. *One Chiral Fingerprint to Find Them All*; preprint; Chemistry, 2023. https://doi.org/10.26434/chemrxiv-2023-33j02.

(35) Gogoladze, G.; Grigolava, M.; Vishnepolsky, B.; Chubinidze, M.; Duroux, P.; Lefranc, M.-P.; Pirtskhalava, M. DBAASP : Database of Antimicrobial Activity and Structure of Peptides. *FEMS Microbiol. Lett.* **2014**, *357* (1), 63–68. https://doi.org/10.1111/1574-6968.12489.

(36) Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-sequence-based Prediction of Protein Secondary Structures and Solvent Accessibility by Deep Whole-sequence Learning. *J. Comput. Chem.* **2018**, *39* (26), 2210–2216. https://doi.org/10.1002/jcc.25534.

(37) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The Helical Hydrophobic Moment: A Measure of the Amphiphilicity of a Helix. *Nature* **1982**, *299* (5881), 371–374. https://doi.org/10.1038/299371a0.

(38) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(39) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminformatics* **2021**, *13* (1), 2. https://doi.org/10.1186/s13321-020-00478-9.

(40) Capecchi, A.; Reymond, J.-L. Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning. *Biomolecules* **2020**, *10* (10), 1385. https://doi.org/10.3390/biom10101385.

(41) Capecchi, A.; Reymond, J.-L. Classifying Natural Products from Plants, Fungi or Bacteria Using the COCONUT Database and Machine Learning. *J. Cheminformatics* **2021**, *13* (1), 82. https://doi.org/10.1186/s13321-021-00559-3.

(42) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminformatics* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.

Supplementary Information for:

Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity?

Markus Orsi,[a] and Jean-Louis Reymond[a]*

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

e-mail: jean-louis.reymond@unibe.ch

Table of Contents

Table S1. Performance metrics of all models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold for activity and hemolysis separately. Results for reduced training sets are reported for 20% and 2% size of the original activity dataset and 10% of the original hemolysis set.

Model	ROC AUC	Accuracy	Precision	Recall	F1
GPT-3 Ada act.	0.84	0.78	0.78	0.78	0.78
GPT-3 Babbage act.	0.85	**0.79**	**0.79**	0.78	**0.79**
GPT-3 Curie act.	**0.86**	**0.79**	0.78	**0.81**	**0.79**
GPT-3 Ada 20% act.	0.75	0.69	0.7	0.67	0.68
GPT-3 Babbage 20% act.	0.76	0.69	0.7	0.69	0.68
GPT-3 Curie 20% act.	0.76	0.7	0.71	0.71	0.71
GPT-3 Ada 2% act.	0.66	0.6	0.6	0.63	0.61
GPT-3 Babbage 2% act.	0.66	0.62	0.6	0.73	0.66
GPT-3 Curie 2% act.	0.65	0.6	0.6	0.63	0.61
GPT-3 Ada hem.	**0.9**	0.82	0.8	**0.79**	0.79
GPT-3 Babbage hem.	0.87	0.8	0.76	0.76	0.76
GPT-3 Curie hem.	0.89	**0.84**	**0.82**	**0.79**	**0.8**
GPT-3 Ada 10% hem.	0.72	0.68	0.63	0.58	0.6
GPT-3 Babbage 10% hem.	0.72	0.7	0.65	0.6	0.62
GPT-3 Curie 10% hem.	0.73	0.68	0.63	0.59	0.61

Table S2. Mean and standard deviation of performance metrics of selected models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold.

Model	ROC AUC	Accuracy	Precision	Recall	F1
Features SVM act.	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.63 ± 0.01	0.64 ± 0.01
MAP4C SVM act.	0.8 ± 0.01	**0.8 ± 0.01**	**0.78 ± 0.01**	0.83 ± 0.01	**0.80 ± 0.01**
RNN act.	**0.85 ± 0.01**	0.78 ± 0.01	0.76 ± 0.02	0.81 ± 0.01	0.78 ± 0.01
GPT-3.5 Turbo act.	0.69 ± 0.01	0.69 ± 0.01	0.62 ± 0.01	**0.95 ± 0.01**	0.75 ± 0.01
Features SVM hem.	0.62 ± 0.01	0.64 ± 0.01	0.59 ± 0.02	0.48 ± 0.02	0.53 ± 0.01
MAP4C SVM hem.	0.82 ± 0.02	**0.82 ± 0.01**	**0.78 ± 0.02**	**0.82 ± 0.04**	**0.79 ± 0.01**
RNN hem.	**0.87 ± 0.01**	0.81 ± 0.01	0.77 ± 0.03	0.79 ± 0.03	0.78 ± 0.01
GPT-3.5 Turbo hem.	0.47 ± 0.01	0.48 ± 0.01	0.38 ± 0.02	0.36 ± 0.02	0.37 ± 0.02

Table S3. Training times and costs of GPT models on the full training sets.

Model	Time (h)	Costs ($)
GPT-3 Ada Activity	01:05:04	$0.39
GPT-3 Babbage Activity	01:09:38	$0.59
GPT-3 Curie Activity	01:15:05	$2.93
GPT-3.5 Turbo Activity	00:53:24	$7.00
GPT-3 Ada Hemolysis	00:55:37	$0.09
GPT-3 Babbage Hemolysis	00:57:19	$0.13
GPT-3 Curie Hemolysis	01:08:09	$0.67
GPT-3.5 Turbo Hemolysis	00:55:58	$1.66