# GPT-3 accurately predicts antimicrobial peptide activity and hemolysis

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) have gained significant attention in the field of drug discovery due to their potential therapeutic applications in the fight against antimicrobial resistance. Since rationally designing AMPs is notoriously difficult due to the vast number of possible peptide sequences and their complex structure-activity relationship landscape, this problem is ideally suited for machine-learning models, which can be trained from available data to predict new sequences with a desired activity profile. Here we investigated the performance of large language models (LLMs) fine-tuned with data from Database of Antimicrobial Activity and Structure of Peptides (DBAASP) to predict AMP antimicrobial activity and hemolysis from their amino acid sequence. We show that GPT-3 based models perform slightly better than previously reported recurrent neural networks (RNN) and related architectures on comparable datasets. Furthermore, GPT-3 based models perform remarkably well on low data regime. Advantages in terms of training time and costs are also discussed.

**Keywords**: large language models, LLM, GPT-3, 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 activity.[4,5]

To address this issue, several machine learning models have been developed for the *de novo* design of antimicrobial peptides.[6–17] Because property prediction from a peptide sequence can be framed as a natural language processing problem, many of these models use robust architectures specifically designed for language processing tasks.[18–20] Furthermore, the emergence of large language models (LLMs), such as GPT-3,[21] has opened new possibilities for leveraging powerful language processing capabilities in drug discovery applications. Recent attempts 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 the conventional methods, particularly in the low data regime.[22] 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.[23] However, to the best of our knowledge LLMs have not been implemented to predict the activity of peptides.

In this study we aim to compare GPT-3 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-3 models, especially in the low data regimes, we also seek to assess their overall usability and explore the advantages they offer in terms of time and cost effectiveness.

## *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),[24] as previously described.[13] The datasets used for the classification tasks contained 9,548 (7,160 training / 2,388 validation) sequences with annotated antimicrobial and 2,262 (1,723 training / 539 validation) sequences with annotated hemolytic activity. 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. Used datasets are further described in **Table 1**.

**Table 1**: Sizes and composition of the datasets used in the present study. All datasets are available at https://github.com/reymond-group/GPT3_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

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 accessible through the OpenAI API. 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. ROC AUC, accuracy, precision, recall and F1 scores were directly obtained from the OpenAI platform after fine-tuning was completed. These metrics were used to compare the performances of fine-tuned GPT-3 models

with the ones of Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Recurrent Neural Network (RNN) classifiers described in a previous project[13].

**Metrics**

All models were evaluated using five commonly accepted performance metrics: ROC AUC, Accuracy, Precision, Recall and F1.

*ROC AUC (Receiver Operating Characteristic Area Under the Curve*: 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*: 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*: 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*: 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*: 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*

### Datasets and models

To assess the performance of the fine-tuned GPT-3 models for classification of antimicrobial and hemolytic activities, we conducted a comprehensive comparison with Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) and a Recurrent Neural Network (RNN) models previously trained for the same tasks.[13] Since the RNN outperformed the other models in our previous work, we selected it as the benchmark model for the current study. Through this study, we sought to identify whether fine-tuned GPT-3 models could potentially outperform the existing models as well as elucidate benefits and limitations of repurposing state-of-the-art language models for drug discovery projects. We fine-tuned the Ada, Babbage, and Curie models only and made the deliberate decision not to train Davinci, the most powerful out of the four available models, due to its higher training costs. Additionally, we observed that the performances of the fine-tuned GPT-3 models were relatively comparable. As a result, we focused on these three models to ensure a cost-effective and efficient training process while maintaining comparable performance levels.

To maintain consistency and comparability of performance metrics across all tested models, we used the identical training and validation sets that we used to train and validate the NB, SVM, RF and RNN classifiers to fine-tune the GPT-3 models. To further explore the performance of fine-tuned models in scenarios with limited data, we created additional training sets. These sets were derived from the original antimicrobial activity training set and included subsets containing approximately 20% and 2% of the original data. Additionally, we prepared a training set for hemolysis prediction, which consisted of approximately 10% of the original training set. The purpose of creating these reduced training sets was to simulate low-data regimes, which frequently occur in real-life drug discovery projects.

## Training time and costs

Training times for all models were around 1 h and the associated costs found to be quite affordable. The highest training time and cost among all models were achieved by the Curie model, which took 01:15:05 h and $2.93 to train (detailed overview of training times and costs in **Table 2**). Furthermore, the API provided easy access to fine-tuning the GPT-3 models. This removed the need for complex setup and further contributed to the accessibility and usability of these models.

**Table 2.** Training times and costs of 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 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 |

## Antimicrobial activity classification

Our initial objective was to assess the performance of fine-tuned GPT-3 models to correctly predict antimicrobial activity of peptide sequences. In our previous work, we identified the RNN model to outperform the other models, achieving a ROC AUC of 0.84 and an accuracy of 0.76. Upon evaluating the fine-tuned GPT-3 models, we observed that all three tested models demonstrated comparable or even improved performance compared to the RNN. Among all tested models, Curie performed the best, achieving a ROC AUC of 0.86 and an accuracy of 0.79. It is also worth noting that the performances of all GPT-3 models were consistently similar between each other and that the Ada and Babbage models both performed better than the RNN as well (detailed performance metrics in **Table 3**).

As expected, there was a noticeable decrease in performance when we decreased the size of the training set. This decrease in performance can be attributed to the fact that machine learning models rely on large amounts of data to achieve optimal performance. Thus, it is expected that decreasing the training set size would have a negative impact on the model's performance. We found that all fine-tuned models performed poorly in the lowest data regime (training size: 140), indicating that the models struggled to generalize and make accurate predictions with such limited data. However, as the training set size was increased to the middle range (training size: 1400), the performance improved to an acceptable level.

**Table 3.** Performance metrics of all models tested on antimicrobial activity classification. The best value for each metric is highlighted in bold.

| Model | ROC AUC | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| NB | 0.55 | 0.55 | 0.59 | 0.32 | 0.42 |
| SVM | 0.75 | 0.68 | 0.68 | 0.68 | 0.68 |
| RF | 0.81 | 0.71 | 0.7 | 0.75 | 0.73 |
| RNN | 0.84 | 0.76 | 0.74 | 0.8 | 0.77 |
| RNN scrambled | 0.51 | 0.49 | 0.35 | 0.03 | 0.05 |
| GPT-3 Ada | 0.84 | 0.78 | 0.78 | 0.78 | 0.78 |
| GPT-3 Babbage | 0.85 | **0.79** | **0.79** | 0.78 | **0.79** |
| GPT-3 Curie | **0.86** | **0.79** | 0.78 | **0.81** | **0.79** |
| GPT-3 Ada 20% | 0.75 | 0.69 | 0.7 | 0.67 | 0.68 |
| GPT-3 Babbage 20% | 0.76 | 0.69 | 0.7 | 0.69 | 0.68 |
| GPT-3 Curie 20% | 0.76 | 0.7 | 0.71 | 0.71 | 0.71 |
| GPT-3 Ada 2% | 0.66 | 0.6 | 0.6 | 0.63 | 0.61 |
| GPT-3 Babbage 2% | 0.66 | 0.62 | 0.6 | 0.73 | 0.66 |
| GPT-3 Curie 2% | 0.65 | 0.6 | 0.6 | 0.63 | 0.61 |

## Hemolytic activity classification

The results we obtained for the classification of hemolytic activity were comparable to those obtained for antimicrobial activity classification. In our previous work the RNN model achieved the best performance, with a ROC AUC score of 0.87 and an accuracy of 0.76. As for the antimicrobial activity classification, we could observe that all three fine-tuned models performed better than the

RNN, with Curie being the best performing model with a ROC AUC score of 0.89 and an accuracy of 0.84. The performances of the three GPT-3 models were again consistently similar to each other, with Ada even performing slightly better than Curie on the ROC AUC score (detailed performance metrics in **Table 4**).

Generally, all GPT-3 models and the RNN performed better in hemolysis classification than antimicrobial activity classification, despite the dataset being approximately five times smaller. In accordance with our previous findings, reducing the training set size led to a decrease in performance for hemolytic activity classification as well. However, it is noteworthy that the performance in the lower data regime (training size: 170) was better for hemolytic activity compared to antimicrobial activity in the same data regime. This suggests that the models may have exhibited a certain level of generalization capability in the context of hemolytic activity prediction, even with limited data availability.

**Table 4.** Performance metrics of all models tested on antimicrobial activity classification. The best value for each metric is highlighted in bold.

| Model | ROC AUC | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| NB | 0.58 | 0.56 | 0.48 | 0.76 | 0.59 |
| SVM | 0.69 | 0.73 | 0.72 | 0.58 | 0.65 |
| RF | 0.8 | 0.77 | 0.81 | 0.6 | 0.69 |
| RNN | 0.87 | 0.76 | 0.7 | 0.76 | 0.73 |
| RNN scrambled | 0.45 | 0.61 | 0.41 | 0.05 | 0.1 |
| GPT-3 Ada | **0.9** | 0.82 | 0.8 | **0.79** | 0.79 |
| GPT-3 Babbage | 0.87 | 0.8 | 0.76 | 0.76 | 0.76 |
| GPT-3 Curie | 0.89 | **0.84** | **0.82** | **0.79** | **0.8** |
| GPT-3 Ada 10% | 0.72 | 0.68 | 0.63 | 0.58 | 0.6 |
| GPT-3 Babbage 10% | 0.72 | 0.7 | 0.65 | 0.6 | 0.62 |
| GPT-3 Curie 10% | 0.73 | 0.68 | 0.63 | 0.59 | 0.61 |

## *Conclusion*

In general, when comparing the performance of fine-tuned "out-of-the-box" GPT-3 models with specialized models, we observed that the GPT-3 models perform equally, if not better, than the specialized models. This finding is especially interesting, given that the GPT-3 models are not designed explicitly for the prediction of antimicrobial and hemolytic activity.

Furthermore, the training of GPT-3 models through fine-tuning is a relatively easy and fast process. Accessing the API directly eliminates the need for expensive GPUs to run the models. In our study, duration of the fine-tuning process was short, and the associated costs were low. This further highlights a significant advantage of GPT-3 models compared to other models, which typically require more work and optimization efforts. Overall, the good performance and ease of fine-tuning, along with the cost-effectiveness and accessibility of GPT-3 models, make them a promising option for various applications, including those in the field of drug discovery.

**Code availability**

The source codes and datasets used for this study are available at https://github.com/reymond-group/GPT3_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).

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

(16) 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.

(17) 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.

(18) 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.

(19) 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).

(20) 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).

(21) 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).

(22) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 All You Need for Low-Data Discovery in Chemistry?

(23) 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).

(24) 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.