Machine Learning for Antimicrobial Peptide Design

Markus Orsi, Etienne Bonvin, Başak Olcay, Jean-Louis Reymond

Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, Bern, Switzerland
markus.orsi@unibe.ch

The emergence of antimicrobial resistance (AMR) is outpacing the development of new antibiotics.[1–3] Antimicrobial Peptides (AMPs) offer a promising solution to counteract resistance due to their broad effectiveness, slow resistance development, predictable metabolism, and relative ease of synthesis.[4] However, many AMPs are toxic to healthy human cells. Machine learning (ML) techniques can help identify **active and non-toxic** AMP candidates from a vast pool of potential sequences. Yet, most current ML strategies for peptide design focus solely on linear sequences of the 20 natural L-amino acids, which does not adequately represent the natural rich chemical diversity of AMPs.[5–11] Additionally, these methods depend on large curated datasets that might not always be of optimal quality. In this presentation, we delve into two ongoing projects where we leverage ML to design novel AMPs addressing these challenges:

1. Peptide Design Genetic Algorithm (PDGA): This algorithm systematically explores a wide chemical space, incorporating both amino acids and peptoid building blocks. By extending beyond linear sequences to include potential cyclizations, branching points, and N-terminal cappings, the PDGA investigates a large corpus of chemically diverse molecules.[12] We briefly demonstrate how we used the PDGA to generate novel bioactive analogs of polymyxin B2.

2. Leveraging In-House Data: Utilizing a semi-automated 48-well peptide synthesizer, we synthesized and tested 234 diastereomers of the antimicrobial peptide ln65.[13] We employed these in-house generated sets to train two neural networks, designed to predict antimicrobial activity against the five ESKAPE pathogens and hemolytic activity against human erythrocytes. Our results showcase the efficacy of ML with lower data volumes, emphasizing the potential of using of high-quality, in-house data for accurate predictions.

References

[1] C. J. L. Murray *et al*., *The Lancet* **2022**, *399*, 629.

[2] M. Magana, M. Pushpanathan, A. L. Santos, L. Leanse, M. Fernandez, A. Ioannidis, M. A. Giulianotti, Y. Apidianakis, S. Bradfute, A. L. Ferguson, A. Cherkasov, M. N. Seleem, C. Pinilla, C. De La Fuente-Nunez, T. Lazaridis, T. Dai, R. A. Houghten, R. E. W. Hancock, G. P. Tegos, *The Lancet Infectious Diseases* **2020**, *20*, e216.

[3] C. Årdal, M. Balasegaram, R. Laxminarayan, D. McAdams, K. Outterson, J. H. Rex, N. Sumpradit, *Nat. Rev. Microbiol.* **2020**, *18*, 267.

[4] M. Mahlapuu, C. Björn, J. Ekblom, *Critical Reviews in Biotechnology* **2020**, *40*, 978.

[5] A. Capecchi, X. Cai, H. Personne, T. Köhler, C. van Delden, J.-L. Reymond, *Chem. Sci.* **2021**, *12*, 9221.

[6] A. T. Müller, J. A. Hiss, G. Schneider, *J. Chem. Inf. Model.* **2018**, *58*, 472.

[7] F. Plisson, O. Ramírez-Sánchez, C. Martínez-Hernández, *Sci. Rep.* **2020**, *10*, 16581.

[8] P. B. Timmons, C. M. Hewage, *Sci. Rep.* **2020**, *10*, 10869.

[9] M. Ansari, A. D. White, *J. Chem. Inf. Model.* **2023**.

[10] M. M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, *Bioinformatics* **2020**, *36*, 3350.

[11] P. Szymczak, M. Możejko, T. Grzegorzek, R. Jurczak, M. Bauer, D. Neubauer, K. Sikora, M. Michalski, J. Sroka, P. Setny, W. Kamysz, E. Szczurek, *Nat. Commun.* **2023**, *14*, 1453.

[12] A. Capecchi, A. Zhang, J.-L. Reymond, *J. Chem. Inf. Model.* **2020**, *60*, 121.

[13] S. Baeriswyl, H. Personne, I. Di Bonaventura, T. Köhler, C. van Delden, A. Stocker, S. Javor, J.-L. Reymond, *RSC Chem. Biol.* **2021**, *2*, 1608.