ESCAPING PATENTS USING GENERATIVE MACHINE LEARNING

Mohamed Hassan Kane, Medium Biosciences
hassan@medium.bio
Adil Yusuf, Medium Biosciences

Key Words: Enzyme Engineering, Patents, Machine Learning, Protein Language Models, Generative AI

Protein engineering's intellectual property landscape is predominantly governed by patents, serving as a protective measure for biotech companies. A typical patent usually encompasses a handful of reference sequences with an accompanying sequence identity threshold (typically 10-15% edit distance). Historically, obtaining freedom to operate by circumventing patents has been arduous. The most effective techniques often involve discovering new enzymes through homolog search or metagenome mining as directed evolution methods tend to be less efficient for this use case.

This work showcases the application of generative machine learning for escaping patents. Through two case studies, we unravel the potential and constraints of generative protein language models in patent evasion.

Case Study 1: A reference reaction catalyzed by a wild-type enzyme composed of about 550 amino acids was presented, with patent-protected enzyme variants. Utilizing a 6B+ parameters autoregressive protein language model, trained on UniRef90, we seeded the model with sections of the reference enzyme sequence. This resulted in roughly 500 generated sequences. After a rigorous filtering process, emphasizing 15% edit distance and sequence diversity, two sequences were synthesized and screened using 7 replicates. Remarkably, despite an edit distance of 15-16%, equating to over 80 amino acid substitutions, these sequences retained activity levels at 0.9 and 0.85 compared to the wild type.

Case Study 2: Twelve sequences were synthesized and examined against natural and non-natural substrates for a commercially engineered enzyme involved in API synthesis (~250 amino acids). The generated sequences had edit distances between 8 to 12% (20-30 mutations). Out of the 12 synthesized sequences, seven showcased activity ranging from 0.2 to 0.41 times the activity of the commercial enzyme on natural substrates. No significant activity was observed when testing the generated sequences on non-natural substrates or non-native reaction mechanisms.

In conclusion, generative machine learning manifests promising capabilities in generating protein sequences that navigate around patent restrictions. However, the technology's current iteration seems best suited for natural reactions on natural substrates. Adaptability towards diverse substrates or pioneering reaction mechanisms remains a hurdle. As we continue to refine this approach, we anticipate broader applications that will reshape the protein engineering IP landscape.

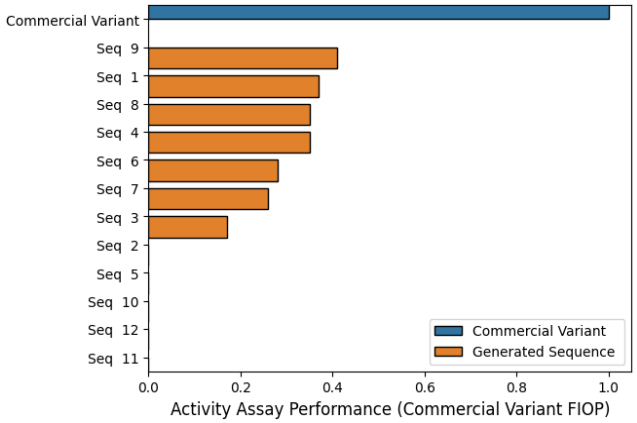

Figure 1: Generated sequence performance when benchmarked against commercial variant (Case Study #2)