De novo generation of functional terpene synthases using TpsGPT

Anonymous Author(s)
Affiliation
Address
email

Abstract

Terpene synthases (TPS) are a key family of enzymes responsible for generating the diverse terpene scaffolds that underpin many natural products, including front-line anticancer drugs such as Taxol. However, *de novo* TPS design through directed evolution is costly and slow. We introduce TpsGPT, a generative model for scalable TPS protein design, built by fine-tuning the protein language model ProtGPT2 on 79k TPS sequences mined from UniProt. TpsGPT generated *de novo* enzyme candidates *in silico* and we evaluated them using multiple validation metrics, including EnzymeExplorer classification, ESMFold structural confidence (pLDDT), sequence diversity, CLEAN classification, InterPro domain detection, and Foldseek structure alignment. From an initial pool of 28k generated sequences, we identified seven putative TPS enzymes that satisfied all validation criteria. Experimental validation confirmed TPS enzymatic activity in at least two of these sequences. Our results show that fine-tuning of a protein language model on a carefully curated, enzyme-class-specific dataset, combined with rigorous filtering, can enable the *de novo* generation of functional, evolutionarily distant enzymes.

1 Introduction

Terpene synthases (TPS) are a specialized family of enzymes that generate hydrocarbon scaffolds for terpenes—the largest and most diverse class of natural products, encompassing widely used flavors, fragrances, and frontline medicines [Samusevich et al., 2025]. Terpenes exhibit diverse bioactivities, including analgesic, anticonvulsant, and anti-inflammatory properties [Del Prado-Audelo et al., 2021]. More than 76,000 terpenes have been characterized to date [Rudolf and Chang, 2019]. Among them, Taxol, a diterpene, remains a first-line anticancer drug with multi-billion-dollar annual sales [Weaver, 2014].

Despite their importance, terpenes are notoriously difficult to synthesize industrially due to their structural complexity [Del Moral et al., 2019]. Conventional chemical synthesis requires numerous steps and incurs high energy and resource costs, making it unsustainable at scale. In contrast, synthetic biology offers a more efficient route by leveraging TPS enzymes to catalyze key reactions [Zhang and Hong, 2020].

Here we present **TpsGPT**[1], a terpene synthase sequence generation model fine-tuned on a distilled protein language model — ProtGPT2 Tiny [Ferruz et al., 2022, protgpt2 tiny, 2022]. TpsGPT is trained on a carefully curated 79k homologous TPS dataset mined from large scale repositories like UniProt. The mining process initially used a very small **1125** experimentally characterized actual TPS enzymes from published sources as a seed to identify TPS patterns based on which the mining process produced 79k homologous TPS sequences. TpsGPT generated evolutionary distant sequences while conserving key TPS structural features. The resulting *de novo* sequences exhibit high predicted

[1] `https://anonymous.4open.science/r/TpsGPT-C55/`

Submitted to the AI for Science workshop (NeurIPS 2025).

structural stability and low sequence identity relative to the training set. Our results demonstrate that fine-tuning protein language models on a carefully curated, enzyme-class-specific dataset, can effectively explore the vast protein sequence space, producing valid enzyme candidates even for underrepresented protein families like terpene synthases.

2 Related Work

Protein engineering: The design of novel TPS enzymes for terpene biosynthesis remains a complex and time-consuming task. There are two main approaches for protein engineering: rational design and directed evolution [Vidal et al., 2023]. Rational design involves performing chosen point mutations, insertions or deletions in the coding sequence. Directed evolution, on the other hand, bypasses the need to determine specific mutations a priority by mimicking the process of natural evolution in the laboratory. While promising, these methods have a major disadvantage — the sequences they generate often remain highly similar to naturally occurring proteins, leaving vast regions of the protein sequence space unexplored [Yang et al., 2024]. Moreover, robotics-accelerated high-throughput directed evolution techniques like Phage-Assisted Continuous Evolution are prohibitively expensive, with costs reaching hundreds of thousands of dollars [Aoudjane et al., 2024]

Computational design of terpene synthases: Machine learning-assisted annotation methods predict and label likely TPS enzymes in large protein databases like UniProt and UniRef [Samusevich et al., 2025, Bateman, 2018, Suzek et al., 2014] but such methods only uncover existing proteins in nature. *De novo* enzyme design approaches such as RFdiffusion use diffusion-based deep learning architectures to generate novel protein backbones [Watson et al., 2023]. Although promising, RFDiffusion is a structure-based method and requires a comprehensive understanding of a catalytic site and its activity to generate functional enzymes [Lauko et al., 2025]. To the best of our knowledge, little work has been done to explore the generation of valid *de novo* TPS enzymes that differ substantially from natural variants. We address this limitation with our work.

Protein Language Models (PLMs): PLMs are based on large language models (LLMs) like GPT2, which leverage the Transformer architecture to model sequential data [Vaswani et al., 2017]. Prior work has shown that fine-tuning PLMs can generate *de novo* proteins within specific families [Winnifrith et al., 2024]. However, existing PLM fine-tuning methods to generate sequences rely on extensive family-specific datasets and often require additional inputs such as control tags for model conditioning. Additionally, the fine-tuning is typically done on large models such as ProGEN with 280 million parameters [Madani et al., 2023]. ProtGPT2 is a state-of-the-art autoregressive Transformer-based PLM with 738 million parameters [Ferruz et al., 2022] and enables high-throughput protein generation in seconds. Additionally, ProtGPT2 offers a tiny model [protgpt2 tiny, 2022] with 38.9 million parameters with comparable performance as the original bigger model. Motivated by these properties, we fine-tune ProtGPT2 tiny to generate *de novo* terpene synthase sequences starting from a small dataset of 1125 TPS sequences.

3 Materials and Methods

We developed **TpsGPT**, a scalable *in silico* framework for *de novo* TPS enzyme design (Figure 1). The approach combines protein language model fine-tuning with principled sequence generation and multi-stage validation to produce viable, evolutionarily distant TPS candidates.

3.1 Dataset Preparation

As a starting point, we used a dataset of 1125 experimentally validated TPS sequences, which was later extended to 79k computationally-mined TPS sequences [Čalounová and Pluskal, 2024]. The mining process identified TPS enzyme patterns based on sequence length, protein embedding similarity, conserved motifs, taxonomy, and domain architectures across UniProt and other large-scale repositories, resulting in the dataset of 79k putative TPS sequences.

To avoid data leakage and ensure generalization, the sequences were clustered using SpanSeq [Florensa et al., 2024] into six partitions at 30% sequence identity between partitions. We combined five partitions ($\sim$63k sequences) for training while the remaining partition ($\sim$16k sequences) was reserved for validation, resulting in an 80/20 split.

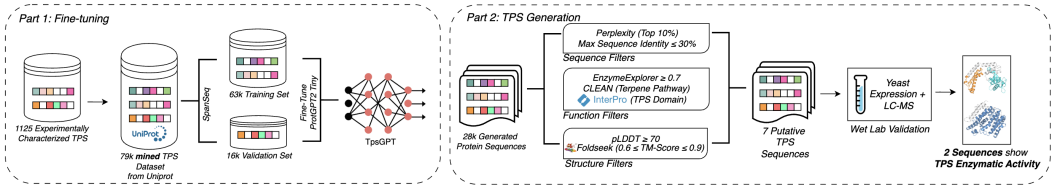

Figure 1: **Overview of our approach. Part 1:** We collected 1125 experimentally characterized TPS enzymes from all published sources to mine the 79k TPS dataset from UniProt [Čalounová and Pluskal, 2024]. We created an 80/20 split using SpanSeq [Florensa et al., 2024] with at most 30% sequence identity between the splits and fine-tuned the distilled ProtGPT2 tiny [protgpt2 tiny, 2022] model to create TpsGPT. **Part 2:** We generated 28k sequences using TpsGPT and filtered them using seven validation metrics: **Sequence filters:** Perplexity and max sequence identity to training set. **Function filters:** EnzymeExplorer TPS score [Samusevich et al., 2025], CLEAN enzyme classification [Yu et al., 2023], and InterPro domain prediction [Blum et al., 2024] **Structure filters:** pLDDT using ESMFold [esmfold, 2025] and max Foldseek TM-score to training set [Van Kempen et al., 2023]. Above filters reduced the 28k sequences to **seven** putative TPS sequences. Wet-lab validation using yeast expression followed by liquid chromatography coupled with mass spectrometry (LC-MS) showed TPS enzymatic activity in two sequences [Pitt, 2009].

3.2 Model Fine-Tuning

The original ProtGPT2 model contains 738 million parameters, making full fine-tuning computationally expensive [protgpt2, 2022]. Hence, we fine-tuned the distilled ProtGPT2 tiny model with 38.9 million parameters. The distilled tiny model retains comparable perplexities to the original large model while offering up to six times faster inference, enabling high-throughput sequence generation [protgpt2 tiny, 2022]. Fine-tuning was performed using Lightning AI on a single NVIDIA L4 tensor core [lightning, 2025].

3.3 TPS Sequence Generation and Filtering

After fine-tuning, we generated 28k protein sequences. A multi-stage filtering pipeline was applied to identify putative TPS enzymes from the 28k sequences:

Sequence Filters: The 28k sequences were ranked by perplexity, and the top 10% (2,800 sequences) were retained. Maximum pairwise sequence identity (maxID) to the training set was computed, and only sequences with maxID $\leq 60\%$ were retained to encourage evolutionary distance.

Function Filters: We used EnzymeExplorer with a TPS detection threshold of 0.7 (range 0–1) to select sequences likely to possess TPS activity [Samusevich et al., 2025]. CLEAN (Contrastive Learning Enabled Enzyme ANnotation) is a ML model that assigns EC (Enzyme Commission) number to protein sequences [Yu et al., 2023]. We used CLEAN to predict EC numbers and selected only those with a terpenoid/terpene biosynthetic pathway in BRENDA [brenda, 2025]. InterPro is another model that predicts domains given a sequence [Blum et al., 2024]. We selected only those sequences when the InterPro predicted domain was a terpene synthase specific domain or a domain with an overlapping superfamily containing a TPS domain.

Structure Filters: We computed Predicted Local Distance Difference Test (pLDDT) scores from ESMFold [esmfold, 2025], retaining only sequences with pLDDT ≥ 70, indicative of accurate backbone modeling and valid 3D structures. To ensure conservation of TPS structure, we used Foldseek to do 3D structural comparison of the generated sequences relative to their respective top structural matches in the training set and retained those with TM-scores between 0.6 and 0.9 [Van Kempen et al., 2023].

This pipeline produced candidates that are structurally feasible, TPS-like, and evolutionarily distant, representing potential *de novo* TPS enzymes suitable for downstream experimental validation.

4 Results

We fine-tuned the distilled ProtGPT2 tiny model on 79k TPS sequences mined from UniProt and generated 28k TPS sequence candidates. After picking the top 10% (2800) sequences by perplexity score, we applied the following filters: pLDDT score, EnzymeExplorer TPS detection score, max

sequence identity to training set, Foldseek alignment (TM-Score), CLEAN classified EC number, and InterPro domain to identify putative *de novo* TPS sequences.

4.1 TpsGPT Generates Valid TPS Sequences

Among the top 2,800 sequences ranked by perplexity, 40% achieved pLDDT scores ≥ 70 (Figure 2). From this set, **77 sequences** passed the EnzymeExplorer TPS detection threshold ($>= 0.7$). A detection score above 0.7 indicates the potential to catalyze terpenes.

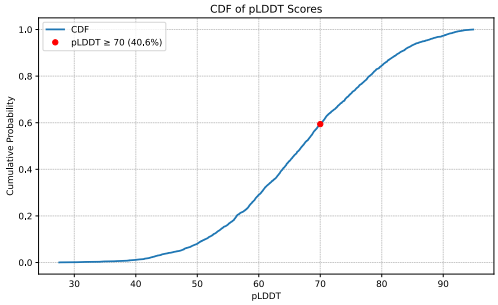

Figure 2: CDF of pLDDT scores of the top 2800 generated sequences. More than 40% had pLDDT ≥ 70 indicating stable structures.

4.2 Evolutionarily Distant Sequences with Conserved TPS Structures

From the 77 candidates, we filtered down to seven with $\leq 60\%$ sequence identity to the training set, representing potential *de novo* TPS enzymes (Table 1). 3D structural comparison of the generated sequences relative to their respective top structural matches in the training set using Foldseek confirmed TM-scores between 0.6 and 0.9 (Table 1), consistent with belonging to the same structural family [Van Kempen et al., 2023]. Moreover, CLEAN assigned all seven sequences to TPS EC classes, providing robust computational support [Yu et al., 2023] (Table 1). InterPro analysis detected at least one relevant TPS specific domain in each sequence as shown in Figure 3 [Blum et al., 2024]. Together, the above results show that TpsGPT generates evolutionary distant yet structurally conserved *de novo* TPS enzymes.

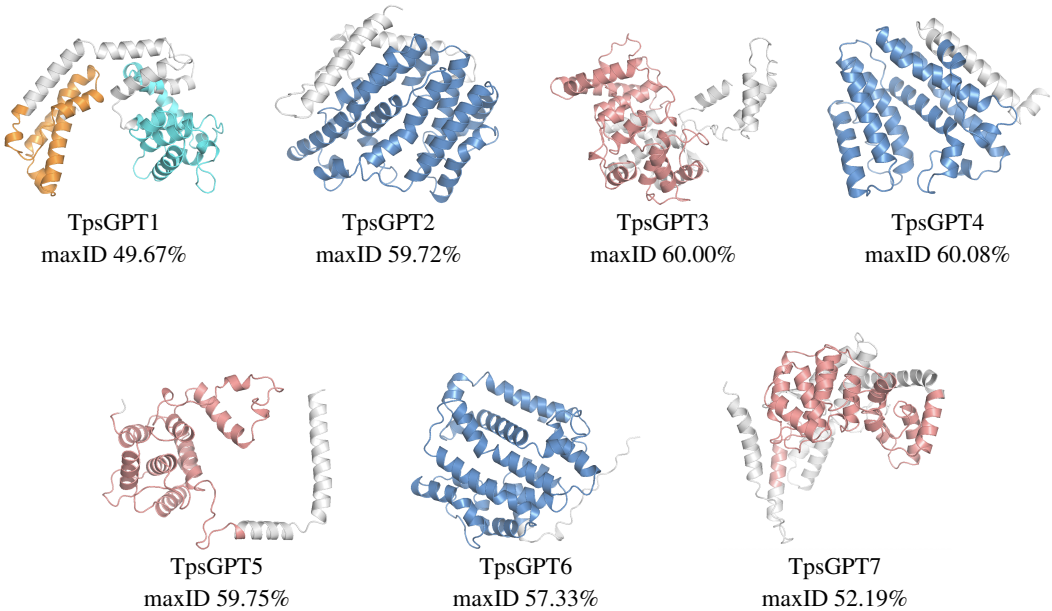

Figure 3: ColabFold-generated 3D structures of the seven *de novo* putative TPS enzymes, with TPS domains annotated by InterPro [Mirdita et al., 2022, Blum et al., 2024]. maxID denotes the maximum sequence identity of each sequence to those in the training set. The figure shows that TpsGPT can generate evolutionarily distant TPS sequences while conserving TPS domains.

Table 1: Properties of the seven putative *de novo* TPS sequences. Each sequence is distinguished by a unique Sequence ID. EnzymeExplorer TPS score measures TPS-like characteristics, pLDDT score indicates the stability of 3D folding, Max Foldseek TM-score indicates structural alignment with TPS sequences in training set, Max seq. identity to training set denotes the uniqueness of the TPS sequence, CLEAN-predicted EC number provides the enzyme classification, and InterPro predicts the domains in the sequence. TpsGPT1 and TpsGPT2 are shown in bold, indicating experimentally-validated enzymatic activity in both sequences.

Sequence ID	EnzymeExplorer TPS Score	pLDDT Score	Max Foldseek TM-Score to training set	Max seq. identity to training set	CLEAN Predicted EC number	InterPro Predicted Domain
TpsGPT1	0.75	78	0.73	49.67%	Germacrene D Synthase (4.2.3.75)	Terpene synthase, N-terminal domain and Terpene synthase, metal-binding domain
TpsGPT2	0.72	74	0.79	59.72%	Squalene Synthase (2.5.1.21)	Trans-isoprenyl diphosphate synthases, head-to-head domain
TpsGPT3	0.73	74	0.84	60.00%	Cucurbitadienol Synthase (5.4.99.33)	Squalene-hopene cyclase, N-terminal domain
TpsGPT4	0.73	70	0.65	60.08%	Squalene Synthase (2.5.1.21)	Trans-isoprenyl diphosphate synthases, head-to-head domain
TpsGPT5	0.78	80	0.72	59.75%	Beta-amyrin Synthase (5.4.99.39)	Squalene-hopene cyclase, N-terminal domain
TpsGPT6	0.73	71	0.69	57.33%	Squalene Synthase (2.5.1.21)	Trans-isoprenyl diphosphate synthases, head-to-head domain
TpsGPT7	0.74	71	0.72	52.19%	Cycloartenol Synthase (5.4.99.8)	Squalene-hopene cyclase, N-terminal domain

4.3 Experimental Validation Confirms Enzymatic Activity

To functionally characterize the enzymes designed with TpsGPT, we heterologously expressed the corresponding genes in the budding yeast Saccharomyces cerevisiae strain JWY501. This strain has been engineered for elevated production of the diterpene substrate geranylgeranyl pyrophosphate. Using liquid chromatography, coupled with mass spectrometry (LC-MS) [Pitt, 2009] we confirmed the enzymatic activity in two sequences (**TpsGPT1** and **TpsGPT2**) (Figure 4). Ongoing experiments aim to characterize their catalytic mechanisms in more detail, as well as to validate other generated TPSs.

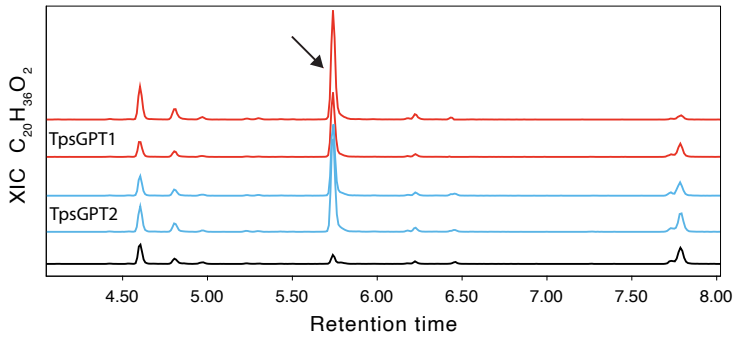

Figure 4: Chromatogram showing wet-lab validation of enzymatic activity for the generated **TpsGPT1** and **TpsGPT2** sequences. Extracted ion chromatograms (XIC) at the mass of C20H36O2 confirm the production of diterpene-like products (e.g., sclareol; CHEBI:9053) in yeast expressing TpsGPT1 (red; two replicates) and TpsGPT2 (blue; two replicates). Black trace represents the control.

5 Conclusion

In this work, we demonstrate the potential of fine-tuning protein language models, specifically ProtGPT2 Tiny, on a carefully curated TPS dataset to generate novel and valid terpene synthases. The seven sequences generated by **TpsGPT** exhibited high pLDDT scores, indicating stable 3D structures, and low perplexity scores, suggesting syntactically valid protein sequences. The seven sequences were also functionally validated by EnzymeExplorer, CLEAN and InterPro models. Furthermore, low pairwise sequence identity and favorable Foldseek TM-scores indicate the likelihood of discovering evolutionarily distant TPS enzymes not present in nature. Importantly, the entire pipeline was executed with less than **$200** in GPU cost, demonstrating the scalable and cost-efficient nature of this approach. Additionally, our approach can also be applied to underrepresented protein families with little characterized enzyme datasets. These results validate our hypothesis that ProtGPT2 can be fine-tuned on a carefully curated TPS dataset to produce valid *de novo* TPS candidates.

Among the seven generated sequences, enzymatic activity was so far confirmed in only two, and the presence of oxygen in the product chemical formula suggests they cannot yet be confirmed as canonical TPS enzymes. We plan to conduct further wet-lab experiments to characterize their activity and refine our *in silico* pipeline. Future directions include conditioning TPS generation on terpene subclasses via curated datasets to generate specific terpenes. While this study focused on the TPS family, the methodology can be generalized to other protein families, such as, for example, lysozymes, to explore functional diversity.

References

Samir Aoudjane, Stefan Golas, Osaid Ather, Michael J. Hammerling, and Erika DeBenedictis. A practical guide to phage- and robotics-assisted near-continuous evolution. *Journal of Visualized Experiments*, (203), 1 2024. doi: 10.3791/65974. URL https://doi.org/10.3791/65974.

Alex Bateman. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1): D506–D515, 10 2018. doi: 10.1093/nar/gky1049. URL https://doi.org/10.1093/nar/gky1049.

Matthias Blum, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, Ailsa Orr, Typhaine Paysan-Lafosse, Irina Ponamareva, Gustavo A Salazar, Nicola Bordin, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunic, Felipe Llinares-López, Aron Marchler-Bauer, Laetitia Meng-Papaxanthos, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Damiano Piovesan, Catherine Rivoire, Christian J A Sigrist, Narmada Thanki, Françoise Thibaud-Nissen, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456, 11 2024. doi: 10.1093/nar/gkae1082. URL https://doi.org/10.1093/nar/gkae1082.

brenda. BRENDA Enzyme Database, 3 2025. URL https://www.brenda-enzymes.org/index.php.

clean. Clean, 2025. URL https://clean.platform.moleculemaker.org/configuration.

José Francisco Quílez Del Moral, Álvaro Pérez, and Alejandro F. Barrero. Chemical synthesis of terpenoids with participation of cyclizations plus rearrangements of carbocations: a current overview. *Phytochemistry Reviews*, 19(3):559–576, 9 2019. doi: 10.1007/s11101-019-09646-8. URL https://doi.org/10.1007/s11101-019-09646-8.

María Luisa Del Prado-Audelo, Hernán Cortés, Isaac H. Caballero-Florán, Maykel González-Torres, Lidia Escutia-Guadarrama, Sergio A. Bernal-Chávez, David M. Giraldo-Gomez, Jonathan J. Magaña, and Gerardo Leyva-Gómez. Therapeutic applications of terpenes on inflammatory diseases. *Frontiers in Pharmacology*, 12, 8 2021. doi: 10.3389/fphar.2021.704197. URL https://doi.org/10.3389/fphar.2021.704197.

enzyme explorer. GitHub - pluskal-lab/EnzymeExplorer: Highly accurate discovery of terpene synthases powered by machine learning, 2025. URL https://github.com/pluskal-lab/EnzymeExplorer.

esmfold. GitHub - facebookresearch/esm: Evolutionary Scale Modeling (esm): Pretrained language models for proteins, 2025. URL https://github.com/facebookresearch/esm.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1), 7 2022. doi: 10.1038/s41467-022-32007-7. URL https://doi.org/10.1038/s41467-022-32007-7.

Alfred Ferrer Florensa, Jose Juan Almagro Armenteros, Henrik Nielsen, Frank Møller Aarestrup, and Philip Thomas Lanken Conradsen Clausen. SpanSeq: similarity-based sequence data splitting method for improved development and assessment of deep learning projects. *NAR Genomics and Bioinformatics*, 6(3), 7 2024. doi: 10.1093/nargab/lqae106. URL https://doi.org/10.1093/nargab/lqae106.

foldseek. GitHub - steineggerlab/foldseek: Foldseek enables fast and sensitive comparisons of large structure sets., 2025. URL https://github.com/steineggerlab/foldseek.

interpro. InterPro, 2025. URL https://www.ebi.ac.uk/interpro/search/sequence/.

Anna Lauko, Samuel J. Pellock, Kiera H. Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F. Shida, Andrew Hunt, Indrek Kalvet, Christoffer Norn, Ian R. Humphreys, Cooper Jamieson, Rohith Krishna, Yakov Kipnis, Alex Kang, Evans Brackenbrough, Asim K. Bera, Banumathi Sankaran, K. N. Houk, and David Baker. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025. doi: 10.1126/science.adu2454. URL https://www.science.org/doi/abs/10.1126/science.adu2454.

lightning. Lightning AI | Turn ideas into AI, Lightning fast, 2025. URL https://lightning.ai/.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 1 2023. doi: 10.1038/s41587-022-01618-2. URL https://doi.org/10.1038/s41587-022-01618-2.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6): 679–682, 5 2022. doi: 10.1038/s41592-022-01488-1. URL https://doi.org/10.1038/s41592-022-01488-1.

James J Pitt. Principles and applications of Liquid Chromatography-Mass Spectrometry in clinical biochemistry, 2 2009. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2643089/.

protgpt2. nferruz/ProtGPT2 · Hugging Face, 2022. URL https://huggingface.co/nferruz/ProtGPT2.

protgpt2 tiny. littleworth/protgpt2-distilled-tiny · Hugging Face, 7 2022. URL https://huggingface.co/littleworth/protgpt2-distilled-tiny.

Jeffrey D. Rudolf and Chin-Yuan Chang. Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. *Natural Product Reports*, 37(3):425–463, 10 2019. doi: 10.1039/c9np00051h. URL https://doi.org/10.1039/c9np00051h.

Raman Samusevich, Téo Hebra, Roman Bushuiev, Martin Engst, Jonáš Kulhánek, Anton Bushuiev, Joshua D. Smith, Tereza Čalounová, Helena Smrčková, Marina Molineris, Renana Schwartz, Adéla Tajovská, Milana Perković, Ratthachat Chatpatanasiri, Sotirios C. Kampranis, Dan Thomas Major, Josef Sivic, and Tomáš Pluskal. Structure-enabled enzyme function prediction unveils elusive terpenoid biosynthesis in archaea. *bioRxiv*, 2025. doi: 10.1101/2024.01.29.577750. URL https://www.biorxiv.org/content/early/2025/04/29/2024.01.29.577750.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 11 2014. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 5 2023. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 6 2017. URL https://arxiv.org/abs/1706.03762.

Lara Sellés Vidal, Mark Isalan, John T. Heap, and Rodrigo Ledesma-Amaro. A primer to directed evolution: current methodologies and future directions. *RSC Chemical Biology*, 4(4):271–291, 1 2023. doi: 10.1039/d2cb00231k. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10074555/#cit4.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 7 2023. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8.

Beth A. Weaver. How Taxol/paclitaxel kills cancer cells. *Molecular Biology of the Cell*, 25(18): 2677–2681, 9 2014. doi: 10.1091/mbc.e14-04-0916. URL https://doi.org/10.1091/mbc.e14-04-0916.

Adam Winnifrith, Carlos Outeiral, and Brian L Hie. Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 86:102794, 4 2024. doi: 10.1016/j.sbi.2024.102794. URL https://doi.org/10.1016/j.sbi.2024.102794.

Jason Yang, Francesca-Zhoufan Li, and Frances H. Arnold. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Central Science*, 10(2):229–229, 2 2024. doi: 10.1021/acscentsci.3c01275. URL https://doi.org/10.1021/acscentsci.3c01275.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 3 2023. doi: 10.1126/science.adf2465. URL https://doi.org/10.1126/science.adf2465.

Caizhe Zhang and Kui Hong. Production of terpenoids by synthetic biology approaches. *Frontiers in Bioengineering and Biotechnology*, 8, 4 2020. doi: 10.3389/fbioe.2020.00347. URL https://doi.org/10.3389/fbioe.2020.00347.

Tereza Čalounová and Tomáš Pluskal. Mining novel terpene synthases from large-scale repositories, 6 2024. URL https://dspace.cuni.cz/handle/20.500.11956/190195.

6 Appendix

6.1 Supplementary Methods

6.1.1 Hyperparameter Optimization

To obtain a well-generalized model, we optimized key hyperparameters of the ProtGPT2 fine-tuning process using the *run_clm.py* script from HuggingFace. The following hyperparameters were considered:

1. **Learning rate:** Learning rate controls how quickly the model updates its weights based on the training sequences. After experimentation, we selected a learning rate of 1e-4, at which point the validation loss converged. Higher rates (e.g., 1e-3) resulted in continued training loss reduction but increased overfitting, as shown in Table A1 and Figure A1.
2. **Block size:** We used a block size of 512 tokens, consistent with the original ProtGPT2 paper. Each block represents the maximum sequence length fed into the model during training.
3. **Batch size and Gradient accumulation steps:** Regular batch size was set to 64. To simulate a larger effective batch size of 512 on a single GPU, we set **gradient accumulation steps** to 8, summing the gradients over eight steps during backpropagation.
4. **Max steps:** The max steps parameter controls the total number of optimization steps (analogous to training epochs). We empirically determined that 4,000 steps were optimal, achieving convergence in both training and validation loss (Table A2 and Figure A2).

6.1.2 Sequence Validation Methods

1. **EnzymeExplorer**: We applied the EnzymeExplorer command-line tools [enzyme explorer, 2025] with a detection threshold of 0.7 to identify putative TPS sequences.

2. **CLEAN**: We used the CLEAN web server [clean, 2025] to predict the EC numbers for the seven generated TPS sequences.
3. **InterPro**: We used the InterProScan web server [interpro, 2025] to predict the protein domains for the seven generated TPS sequences.
4. **Foldseek**: We employed the Foldseek command-line tools [foldseek, 2025] to construct a target database from the training set proteins (63k). Using the *easy-search* command, we identified the top structural match in this database for each of the seven generated TPS sequences and recorded the corresponding TM-score (Figure A4).

6.2 Appendix Tables and Figures

Figure A1: Training and evaluation loss as a function of learning rate.

Table A1: Training and evaluation loss for different learning rates.

Learning rate	Training Loss	Evaluation Loss
1e-6	8.4	8.0
1e-5	7.5	7.8
1e-4	**6.1**	**7.5**
1e-3	4.2	7.4

Figure A2: Training and evaluation loss as a function of max steps.

Table A2: Training and evaluation loss as a function of max steps.

Max steps	Training Loss	Evaluation Loss
1200	6.07	7.49
1875	5.66	7.41
3000	5.21	7.34
4000	**4.94**	**7.32**

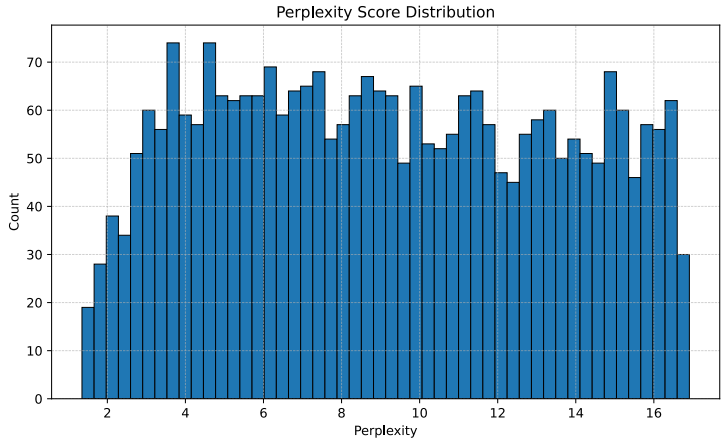

Figure A3: Distribution of perplexity scores for the top 2800 sequences.

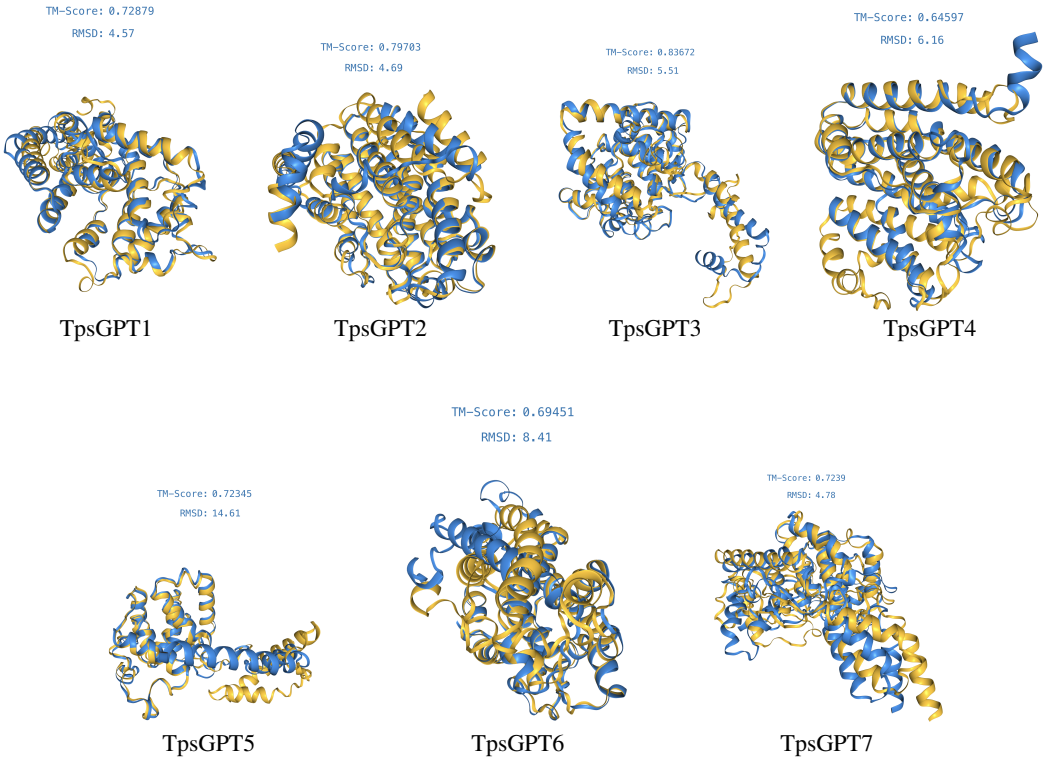

Figure A4: Foldseek structural alignment for the seven TPS sequences with their respective top matches in the training set. Foldseek TM-scores were between 0.6 and 0.9 consistent with belonging to the same TPS family. Blue represents the generated TPS sequences and yellow is the target top structural match in the training set.