White Paper

Gene Pools are an AI-Ready Solution for the Synthesis of Bespoke 1.8kb Gene Libraries

Authors: Kai-Chun Chang, Esteban Toro, Siyuan Chen

April 2026

Abstract

Engineering proteins with optimized function requires exploring large and diverse sequence spaces, often through the construction and screening of thousands of long gene variants. However, assembling genes exceeding ~1 kb—particularly when variants are distributed across the sequence and generated in pooled formats—remains technically challenging. Conventional assembly methods can introduce errors, uneven representation, and chimeric constructs, limiting library quality and slowing experimental progress.

In this white paper, we describe an approach for generating accurate, uniform gene libraries up to 1.8 kb in length using Twist Bioscience's Gene Pools technology. By enabling high-fidelity assembly of bespoke sequence libraries while minimizing chimera formation, this method supports more efficient and scalable protein engineering workflows.

Introduction

Over the last half century, protein engineering has emerged as a promising solution to some of the world's most pressing challenges. Through the precise manipulation of amino acid sequences, enzymes may be engineered to degrade pollutants[1, 2]; crops rendered drought resistant[3, 4]; and natural biomolecules transformed into therapeutics.[5] Yet, engineering proteins to have such transformative abilities remains a significant technical challenge, in large part because of the vast scale of the protein sequence space.

Historically, engineers navigating protein sequence space have relied on directed evolution, wherein a wild-type protein sequence is iteratively mutated and tested. Synthesizing these variants was feasible because they only differ in limited regions, making it possible to restrain mutagenesis and the potential for unintended variants. However, the incremental nature of this approach means that forays into the sequence space were narrowly focused and, consequently, modest in gains.

In recent years, the development of advanced AI and machine learning (AI/ML) algorithms has transformed this process. AI/ML tools enable protein engineers to iteratively test, hone, and prioritize billions of variants in silico according to their predicted effects on protein function. And, unlike directed evolution, variant sequences can be extremely diverse, at times sharing less than 30% homology to known protein sequences while retaining core functionality.[6, 7] Achieving this level of functional diversity with traditional means would be time and resource intensive, if not impossible.[8,9] Instead, with AI/ML, each new variant output by the models can be treated as a completely new hypothesis, and much more of the protein sequence space can be explored without a wet lab.

At some point, however, strong AI-designed variants will need to be tested in the wet-lab. This presents a considerable challenge because, unlike in directed evolution, variation may be combinatorial and spread across the protein sequence. Therefore, each variant must be deliberately synthesized. Bringing such high precision to gene synthesis has proven to be a technical bottleneck. At the time of writing, roughly 41% of proteins in the UniProt database have coding sequences longer than a thousand bases in length.[10] Most platforms tasked with synthesizing and assembling such a long stretch of DNA will suffer from synthesis errors, non-uniformity, chimeras, and other artifacts that ultimately reduce the number of viable screening candidates (and experimental efficiency).

To relieve this bottleneck, we designed a platform for the scalable and accurate assembly of genes up to 1.8kb in length in a pooled format (long enough to encode roughly 85% of all proteins in UniProt at the time of writing).[10] Traditionally, such a platform would have been undermined by the frequent formation of chimeras from errant recombination events among similar sequences (**Figure 1**). Twist's Pooled Gene Assembly platform, however, is specifically designed to prevent chimera formation and reduce errors. Combining Twist's silicon-based long oligo synthesis platform with a precise Pooled Gene Assembly platform, individual oligos are precisely "stitched" and amplified into the full-length, high-fidelity Gene Pools.

By enabling the accurate synthesis of long sequences, Twist's platform also makes it possible for researchers to design variant barcodes into gene sequences. Rather than identify each variant through costly long-read sequencing, variant barcodes enable the use of short-read methods, greatly reducing experimental costs.

Here, we sought to validate the capabilities of this platform by creating and analyzing multiple Gene Pools, each consisting of thousands of sequences with pool-specific gene lengths. To challenge the platform's accuracy, we designed the pools to have a high-degree of similarity among sequences, making them particularly prone to chimera formation. We then assessed the Gene Pools for synthesis errors, uniformity, and chimera formation using PacBio single-molecule sequencing.

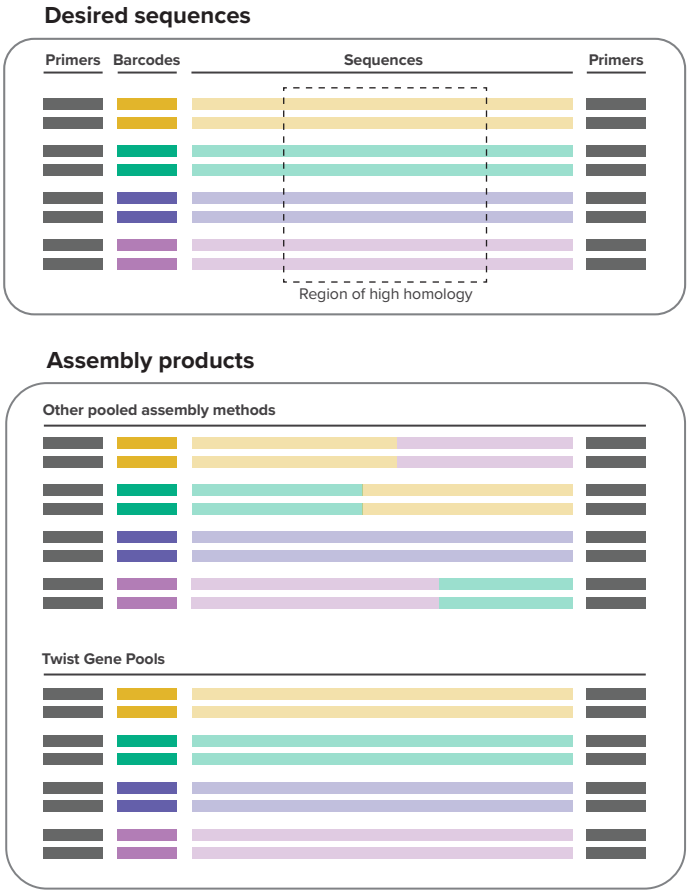

Figure 1: Vulnerability of pooled gene synthesis to chimera formation. Standard Pooled Gene Synthesis typically relies on homology-based gene assembly and PCR amplification of the pool to produce long genes. While valuable, this approach is vulnerable to PCR-mediated homologous recombination (also known as sequence chimeras), particularly among sequences with stretches of high homology.

Study Design

Two Gene Pools—each containing 3,800 sequences—were designed at 650bp and 1.8kb gene lengths to test the capabilities of Twist's platform for both accuracy and uniformity as well as challenge the platform's ability to avert chimeras.

Within each pool, sequences were designed to have up to 97% pairwise similarity, leaving them highly prone to chimera formation using traditional, PCR-based assembly methods.

Once Gene Pools were designed, we used Twist's proprietary design tools to automatically score and generate optimized oligo designs for each sequence. These oligos were synthesized and assembled. The quality of each pool was analyzed by capillary fragment analysis and sequenced using the PacBio long read sequencing platform to measure uniformity, drop-out rate, and any sequence discrepancies (Levenshtein distance) between the read and sequence designs.

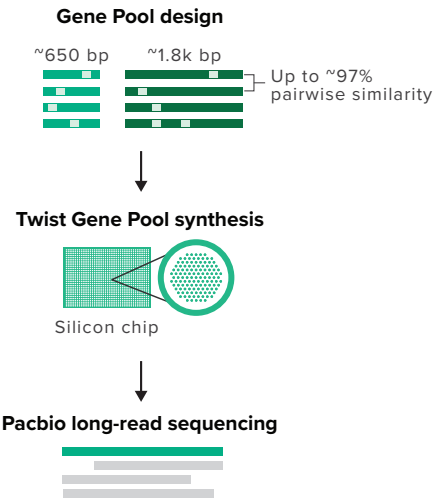

Figure 2. Gene Pool Characterization Workflow. Twist Gene Pools enable the production of a diverse array of sequences in a single pool. In this study, two pools were designed with gene lengths of 650bp or 1.8kb, each containing diverse variant combinations. Gene pools were then produced using Twist's silicon-based synthesis platform, the quality of which was assessed using Pacbio long-read sequencing.

Results

Gene pools display high uniformity and near 0% dropout rate

Efficiency in protein engineering is heavily influenced by the uniformity of screening libraries and, by extension, the Gene Pools that underlay them. A perfectly uniform Gene Pool is one in which every variant sequence is represented by an equal number of molecules. The uniformity of a Gene Pool is typically presented as a ratio between the 95th percentile read count and the 5th percentile readcount (representing the width of a Gaussian curve in which the read counts of each unique sequence is plotted). The closer to 1 this ratio gets, the more uniform the library is. Perfect uniformity is unattainable with modern technology.

DESIGN LENGTH	FULL-LENGTH[a]	PERFECT[b]	UNIFORMITY[c]	CHIMERA[d]	ERROR RATE
~650 bp	>90%	90%	1.28	None detected	1:3024 bp
~1.8 kb	>90%	72%	1.82	None detected	1:3396 bp

Table 1. Summary of gene pool quality metrics

a. Full-length is measured by Qiaxcel capillary electrophoresis.

b. Perfect molecules have reads that match exactly (edit distance = 0) to the mapped reference sequence.

c. Uniformity is defined as the ratio of the 95th to the 5th percentile of read counts, which quantifies the fold-difference in representation across the central 90% of sequences.

d. No chimera detected for pools with up to 97% similarities.

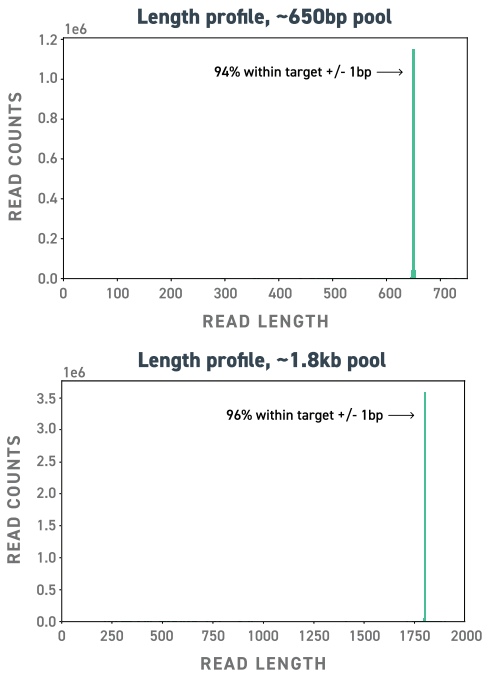

Figure 3. Read lengths form a single tight peak. Read length profiles displayed a single distinct peak for each of the Gene Pools, confirming the absence of large misassembly byproducts and short deletions.

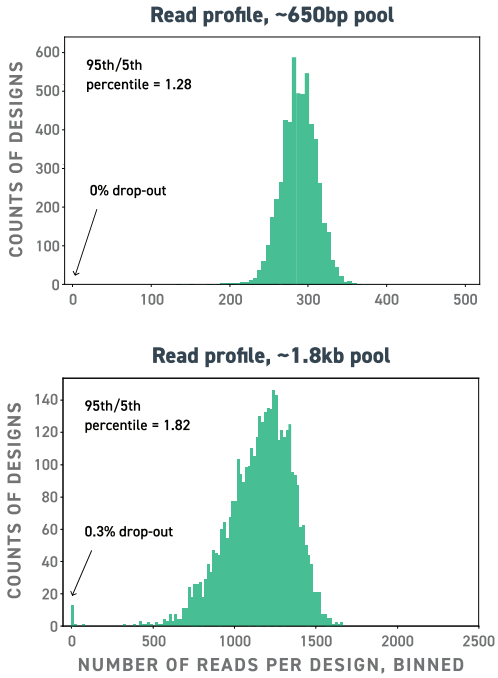

Figure 4. Uniform read distribution and low drop-out rate. Read profiling reports the number of reads per design. A uniform pool should display a tight Gaussian-like peak. Designs with peaks near zero represent drop-out sequences that would, consequently, be absent from subsequent screens.

Twist's Pooled Gene Assembly Platform was capable of producing highly uniform Gene Pools. For both pools, capillary electrophoresis confirmed that >90% of fragments match their expected lengths (data not shown) with PacBio sequencing confirming that both pools had more than 94% of fragments within 1 base pair of the expected length (**Figure 3**). In addition, PacBio sequencing (mean Q-score = 60) analysis revealed excellent uniformity, with >90% of sequences falling within <2.0x of the mean read count and read profiles exhibiting a tight, Gaussian-like peak. The 95th/5th percentile of these pools was 1.28 and 1.82 respectively for the 650bp and 1.8kb pools (**Figure 4**). Drop-out rates ranged from 0% in the ~650bp pool to 0.3% in the ~1.8kb pool.

Pooled gene assembly is highly accurate and chimera-free

Long-read sequencing data confirmed a low gene synthesis error rate of ~1:3000 bp, which is crucial to delivering a high percentage of perfect molecules per pool.

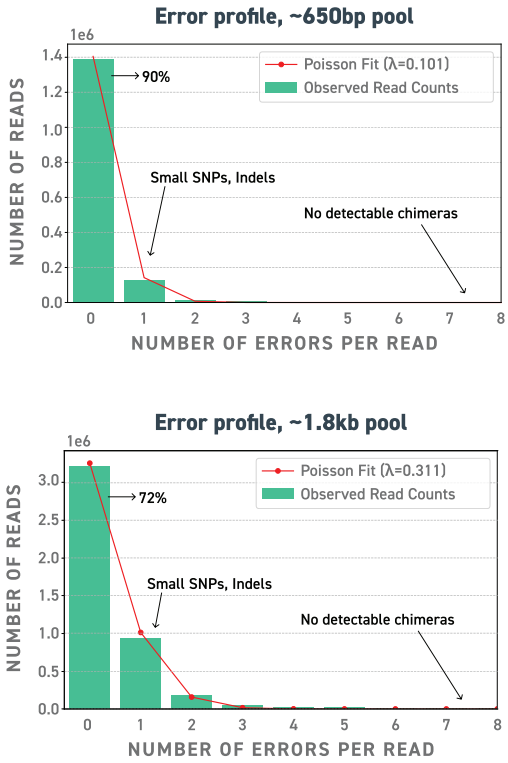

Figure 5. No structural errors. PacBio long-read sequencing was used to assess sequence errors. The graph shows the number of errors per read (X-axis) vs the number of reads (Y-axis). This data fits a Poisson distribution where zero errors indicate a sequence perfect molecule, one or two errors indicate SNPs or small Indels, and many errors (not seen) would indicate sequence chimeras.

We found the percentage of perfect molecules is well-fitted by a Poisson distribution, wherein the ratio of imperfect molecules (λ) only scales with mean fragment length, which can be approximated by λ ~ [fragment length] / [error factor], where error factor is measured and averaged over 3 separate pools to be 5876. The observed error profiles confirm that observed errors are dominated by rare, random, and small mismatches, rather than structural alterations such as chimeras. This yields a predictably high perfect-sequence rate that is a function of sequence length, achieving 90% and 72% perfect molecules for the ~650bp and 1.8kb pools respectively (**Figure 5**). For the molecules that were not perfect, the number of errors detected were minor with the vast majority equal to either 1 or 2 errors per read and therefore represent small SNPs or Indels. Because we found no more than a few errors per read, we can conclude that there are no detectable chimeras found in either the ~650bp and 1.8kb pools, as chimeras would result in a much larger number of errors detected per read.

This data confirms that Twist's Gene Pools platform is capable of synthesizing highly accurate, uniform, and chimera free genes up to 1.8kb in length.

Discussion

The protein engineering field has long faced the challenge of exploring a sequence space that stretches far beyond the bounds of modern technology—one that cannot be effectively navigated through random chance. Despite these limitations, engineered proteins have had far reaching effects, particularly in the field of medicine where, in 2025, the global protein therapeutics market size was valued at USD 388.9 Billion. Projections suggest the market could grow to USD 637.9 Billion by 2034, but to do so, the field will need to overcome its technical challenges.[11]

To engineer proteins with new and optimized functionality, researchers need the ability to methodically screen a diverse array of protein variants, iteratively building upon those candidates that show improvement. With recent developments in AI/ML, such iterative exploration has become possible in silico. However, engineers now face a new challenge: Accurately synthesizing bespoke, AI-designed sequences for wet-lab validation at scale.

In our study, we evaluated Twist's Gene Pools platform as a unique solution to this need. Specifically, we evaluated the platform's ability to produce pooled gene libraries that have uniform variant representation and are synthesized with near perfect accuracy, even with long gene lengths.

PacBio analysis revealed excellent uniformity, with >90% of sequences falling within <2.0x of the mean read count and drop-out rates ranging from 0% to 0.3% across pools. This data also confirmed a low synthesis error rate (~1 error per 3,000 bp), wherein errors were rare, small mismatches rather than recombination-driven misassemblies (such as chimeras). Notably, synthesis errors scaled predictably with gene length, such that the fraction of perfect molecules was 90% and 72% for ~650 bp and ~1.8 kb pools, respectively. Collectively, this data validates Gene Pools as a scalable "wet-lab companion" to AI-driven protein engineering.

As AI/ML approaches reduce the need for brute-force randomization and shift protein engineering toward targeted exploration of designed sequence neighborhoods, the limiting factor increasingly becomes the economical, accurate, and uniform synthesis of long, bespoke gene libraries. The performance observed here—high uniformity, near-zero dropout, predictable perfect-molecule rates with increasing length, and an absence of chimeras—supports Twist Gene Pools as a robust platform for building the high-fidelity libraries that AI-driven protein engineering workflows now require.

With this platform, engineers have the ability to faithfully produce screening libraries with few diversity restrictions—variants may differ by a single point mutation, scattered combinatorial mutations, or deliberate structural variants. In theory, the precision and scale provided by this platform could accelerate the pace and productivity of protein engineering considerably, with directed evolution projects that once took years to complete now being replaced by an AI and Gene Pool enabled workflow that takes weeks. Engineers will likely waste less time and resources on unintended constructs, opting instead to deliberately navigate the protein sequence space under the efficient guidance of AI/ML programs.

Moreover, the ability of Twist's platform to assemble genes up to 1.8 kb—long enough to encode most known proteins in UniProt—greatly expands the amount of the protein sequence space that could now be explored. Whether engineering a novel enzyme, building a more effective therapeutic, or designing resilient crops, having 1.8kb of sequence space to accurately generate entire protein domains—or entire proteins—places more control in the researchers hands. Together with advances in AI/ML, protein engineering is poised for a highly productive era.

References

1. Zhu, Baotong, et al. "Enzyme Discovery and Engineering for Sustainable Plastic Recycling." Trends in Biotechnology, vol. 40, no. 1, Mar. 2021, https://doi.org/10.1016/j.tibtech.2021.02.008.

2. Radley, Emily, et al. "Engineering Enzymes for Environmental Sustainability." Angewandte Chemie, vol. 135, no. 52, 5 Oct. 2023, https://doi.org/10.1002/ange.202309305.

3. Martignago, Damiano, et al. "Drought Resistance by Engineering Plant Tissue-Specific Responses." Frontiers in Plant Science, vol. 10, 22 Jan. 2020, https://doi.org/10.3389/fpls.2019.01676.

4. Park, Sang-Youl, et al. "Agrochemical Control of Plant Water Use Using Engineered Abscisic Acid Receptors." Nature, vol. 520, no. 7548, 23 Apr. 2015, pp. 545–548, https://doi.org/10.1038/nature14123.

5. Sato, Aaron, and Riffle, Stephen. "Synthetic Biology in Drug Development and Beyond." Bioprocessing, Bioengineering and Process Chemistry in the Biopharmaceutical Industry., Springer, Cham, 24 Nov. 2024, doi.org/10.1007/978-3-031-62007-2_2.

6. Watson, Joseph L., et al. "De Novo Design of Protein Structure and Function with RFdiffusion." Nature, vol. 620, 11 July 2023, pp. 1–3, https://doi.org/10.1038/s41586-023-06415-8

7. Madani, Ali, et al. "Large Language Models Generate Functional Protein Sequences across Diverse Families." Nature Biotechnology, 26 Jan. 2023, pp. 1–8, https://doi.org/10.1038/s41587-022-01618-2

8. Yang, Kevin K., et al. "Machine-Learning-Guided Directed Evolution for Protein Engineering." Nature Methods, vol. 16, no. 8, 15 July 2019, pp. 687–694, https://doi.org/10.1038/s41592-019-0496-6.

9. Pierce, Niles A., and Erik Winfree. "Protein Design Is NP-Hard." Protein Engineering, Design and Selection, vol. 15, no. 10, Oct. 2002, pp. 779–782, https://doi.org/10.1093/protein/15.10.779.

10. "UniProt." Www.uniprot.org, www.uniprot.org/uniprotkb/statistics#amino-acid-composition. Accessed 19 Feb. 2026.

11. Group, IMARC. "Protein Therapeutics Market Size, Share, Growth and Industry Report." Imarcgroup.com, 2024, www.imarcgroup.com/protein-therapeutics-market. Accessed 19 Feb. 2026.