Comprehensive Gene Optimization

Increasing recombinant protein yield through coding and non-coding DNA sequence optimization

Technology Overview

MNDL Bio's gene expression optimization platform is built on advanced computational models that integrate multiple factors to maximize recombinant protein yield. Unlike traditional codon optimization methods, which rely solely on frequency tables and change only the coding region, our AI-driven approach accounts for complex biological interactions that impact all gene expression phases and co-optimizes coding and non-coding regions.
Our platform, based on 15 years of research, incorporates, among others:

AI Codon Driven Optimization
Leveraging language models and deep learning to identify optimal codon usage patterns beyond simplistic frequency-based methods, improving translational efficiency [1].

Preservation of Hidden Genetic Information
Using machine learning to construct gene sequences that retain essential hidden information in the host genome, preventing loss of regulatory signals [2,3].

Vector Stability Modeling
Balancing recombinant protein expression with cellular fitness to enhance plasmid stability and maintain high yields [4,6].

Improving DNA Synthesis Success Rate
Removing difficult to synthesize, error-prone sequences while maintaining coding and key genetic elements.

Translation Dynamics Optimization
Modeling non-uniform ribosomal translation rates to prevent misfolding and enhance protein expression [5].

Construct Design and Optimization
Using machine learning and biophysical models to co-optimize promoters and UTRs with the coding sequence of interest, including optimization of protein tags and signal peptides.

Greater Genetic Stability
Integrating a novel, computational-driven optimization approach we design constructs that have greater evolutionary stability. The model indicates epigenetic and mutational hotspots with unstable tendencies. The ESO harnesses constructs that are evolutionary stable without significantly compromising on the levels of expression by predicting trade off between stability and gene expression[4,6].
MNDL Bio's platform app.mndl.bio/signup features an easy-to-use interface. Simply choose your target host and input the coding region of your gene, as well as the flanking 5′ and 3′ sequences.
Other features include:

- Expression temperature input
- Restriction enzymes to avoid
- Sequences to preserve

Case Studies

1. Increased protein expression in *E. coli*.
Objective: Increase expression of Glucose-6-phosphate dehydrogenase (G6PD) in *Escherichia coli* BL21(DE3).
Baseline situation and solution: hG6PD is a challenging enzyme to express in *E. coli* with a truncated product and low yield. We tested several of our algorithms separately and in combinations and found that they can improve the yield of the full-length enzyme.
Results: Many of MNDL Bio's engineered variants showed increased yield of the full-length enzyme (Figure 1).

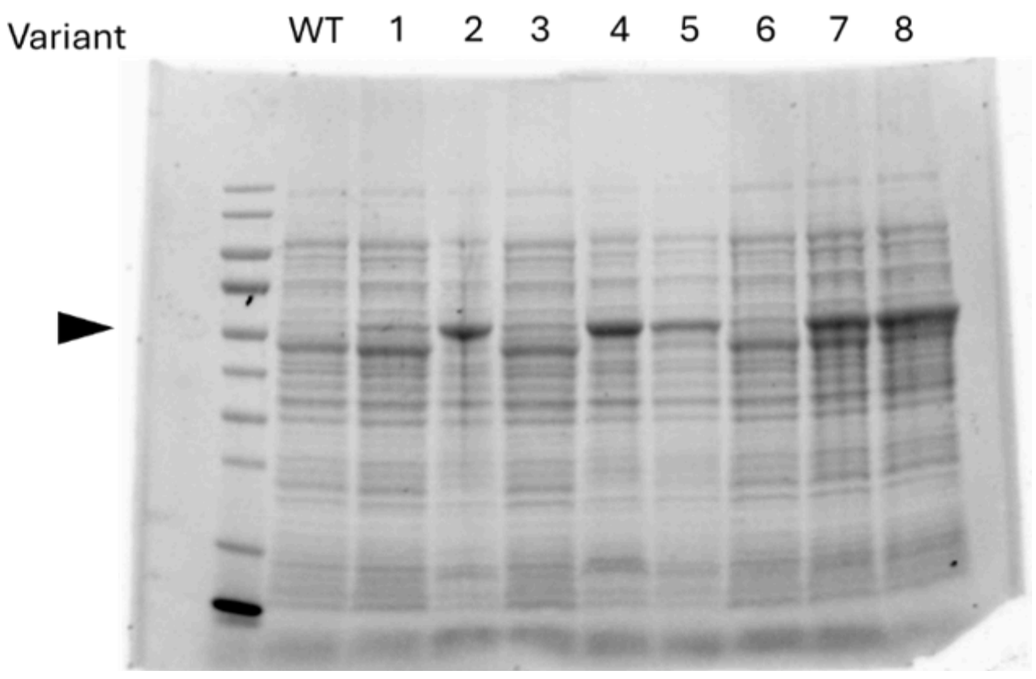
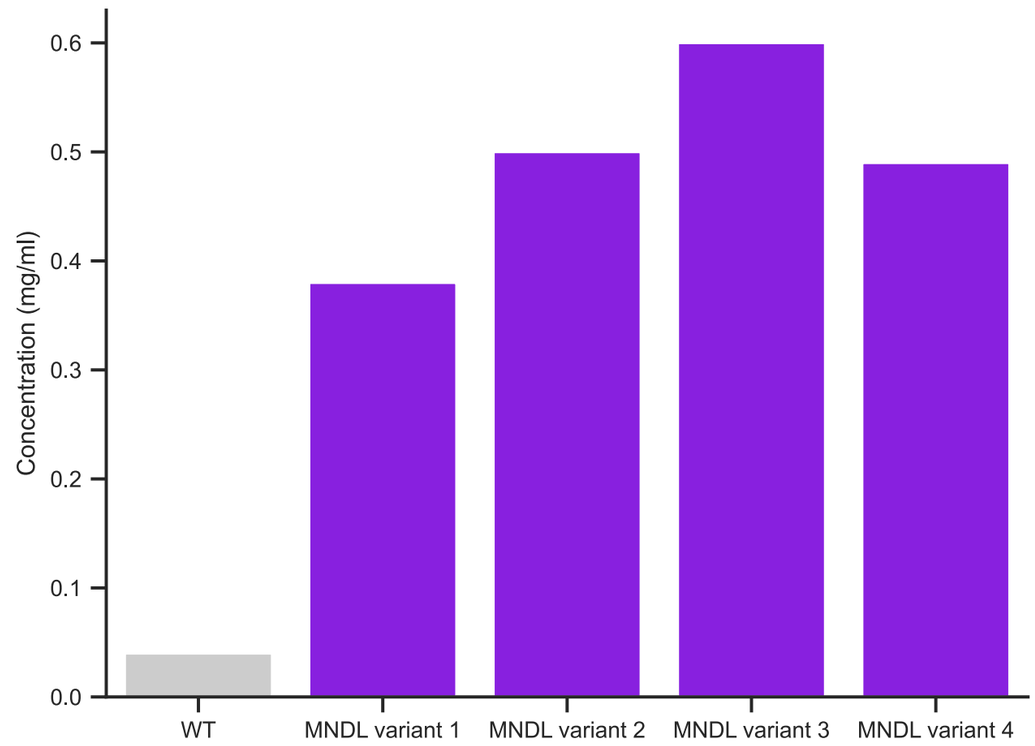


Figure 1. SDS-PAGE analysis of protein expression variants.
The arrow on the left indicates the expected molecular weight position for Human G6PD. The WT sample shows faint expression slightly below the expected band location, suggesting truncation or incomplete translation. In contrast, several optimized variants exhibit more intense bands at the expected size, indicating improved expression of full-length enzyme following sequence optimization.

2. Increased single-domain antibody expression in a cell-free system and live E.coli cultures.

Objective: Increase expression of a ~15KDa single-domain antibody (VHH) in an *E. coli* based cell-free system.
Baseline situation and solution: The WT sequence had very low expression and was not usable for scale-up. To rapidly identify productive designs, we first tested sequence variants in an E. coli cell-free expression system. After identifying the best-expressing variants, we validated a subset in E. coli cultures to ensure improved yields in a true cellular environment.
Results: In vitro (cell-free) expression: OurML, deep-learning, and biophysical-designed variants increased expression by ~9.5 to 15-fold (Figure 2a).
In vivo (E. coli) expression: The same high-performing variants maintained strong expression in live E. coli cultures, with top variants showing up to ~2.3-fold higher yields than WT (Figure 2b).

a.



b.

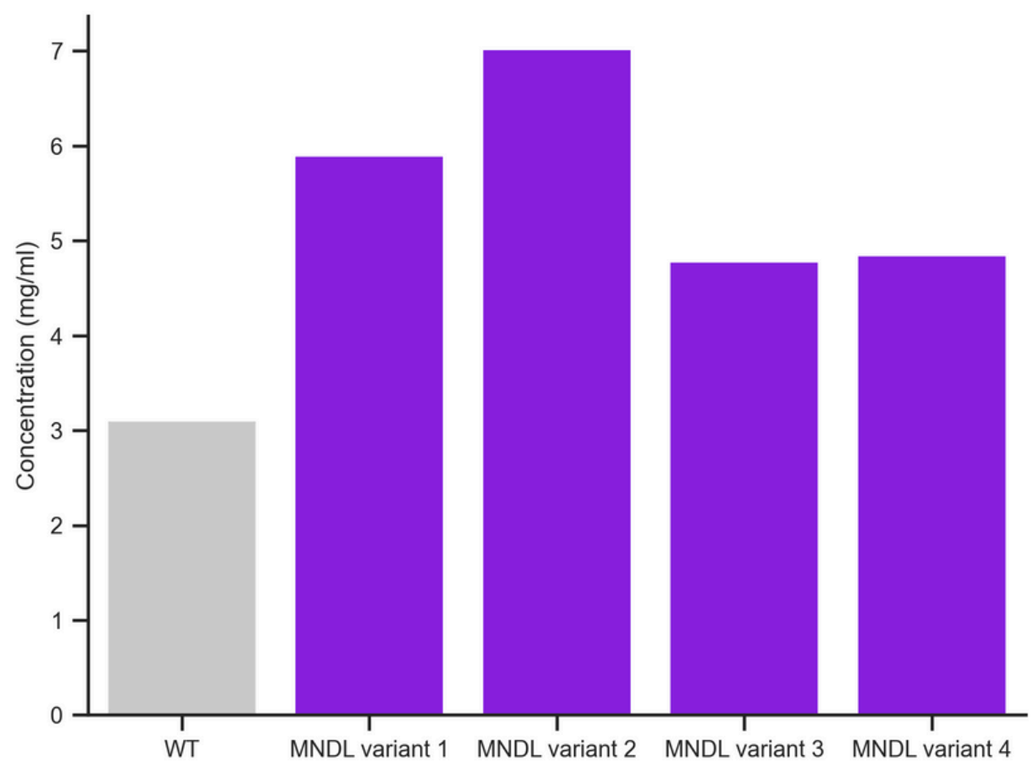


Figure 2. Increased expression of a 15 kDa protein using MNDL designed variants.
(*a*) In vitro expression in an *E. coli*-based cell-free system. Optimized variants (1–4) showed protein concentrations up to 0.6 mg/mL, representing ~9.5–15-fold improvements over WT.
(*b*) In vivo expression in an *E. coli* culture system. Optimized variants (1–4) achieved yields up to ~7.0 mg/mL compared to ~3.1 mg/mL for WT, an improvement of ~2.3-fold.

3. Increased enzyme expression in *E.coli*

Objective: Increase expression of an engineered enzyme in *E. coli*.
Baseline situation and solution: The engineered enzyme yield did not meet expression goals. MNDL
Bio designed 10 variants to be screened.
Results: Variant showed a ~2.5-fold increase in expression (Figure 3).

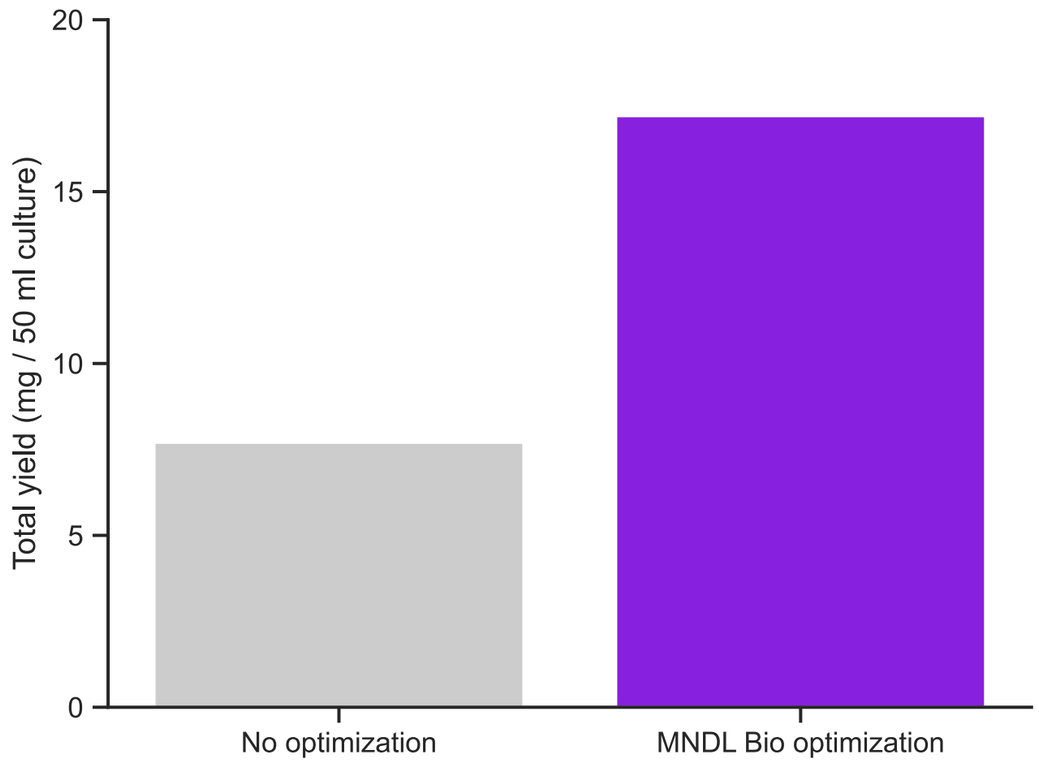


Figure 3. Expression of an engineered enzyme in *E. coli* before and after MNDL Bio optimization. The unoptimized sequence yielded approximately 7 mg of enzyme per 50 mL culture. Following MNDL Bio's optimization, expression increased to ~17.5 mg, representing a ~2.5-fold improvement in total enzyme yield.

4. Increased antibody production in CHO cells

Objective: Increase expression of a commerically available monoclonal antibody in CHO cells.
Baseline situation and solution: The user required high levels of expression to complete proof-of-concept studies on a monoclonal antibody. MNDL Bio designed an optimized variant of the monoclonal antibody to increase it's expression efficiency.
Results: The MNDL optimized variant showed a ~2 -fold increase in the relative level of expression (Figure 4).

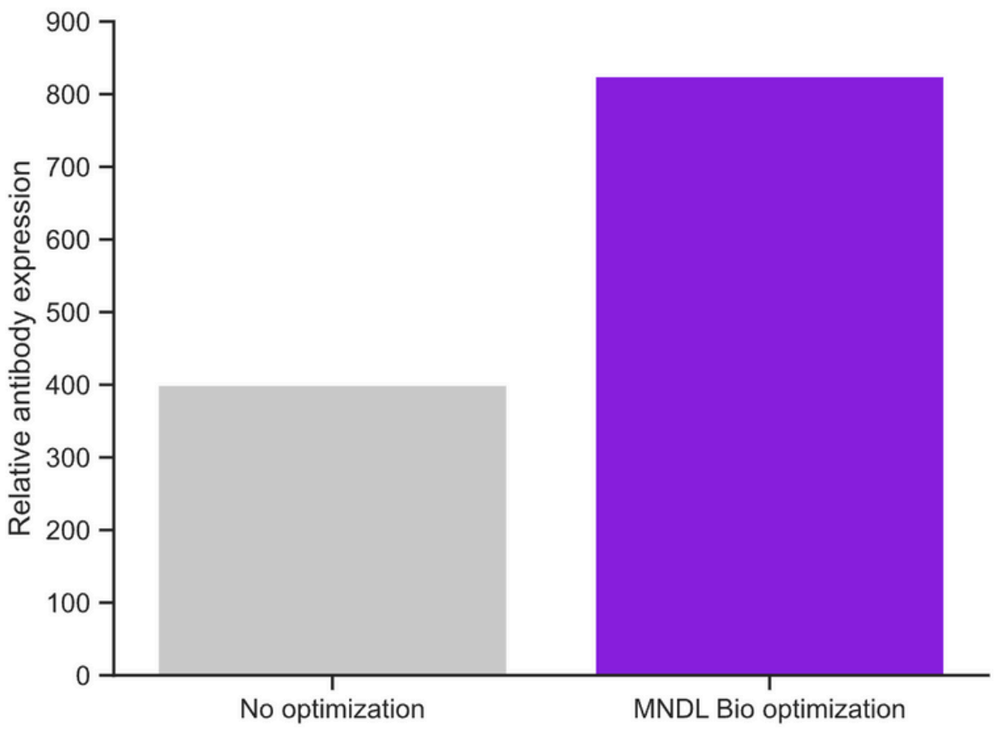


Figure 4. Expression of a monoclonal antibody in CHO before and after MNDL Bio optimization. The unoptimized sequence expressed at a relative level of 400. Following MNDL Bio's optimization, relative expression increased to ~ 825, providing more than a 2-fold improvement in total enzyme yield.

5. Increased growth factor expression in *Pichia Pastoris*

Objective: Increase the expression of a growth factor in *P.pastoris*.
Baseline situation and solution: MNDL Bio engineered 12 variants, each variant was optimized for translation kinetics and mRNA stability. 100 independent transformants were screened per variant to isolate the impact of stochastic genomic events. The median level of expression was calculated for each variant and the variant that showed the highest expression levels was selected.
Results: The graph displays the median expression of the lead variant as 5-fold greater than the median expression of wildtype growth factor gene (Figure 5).

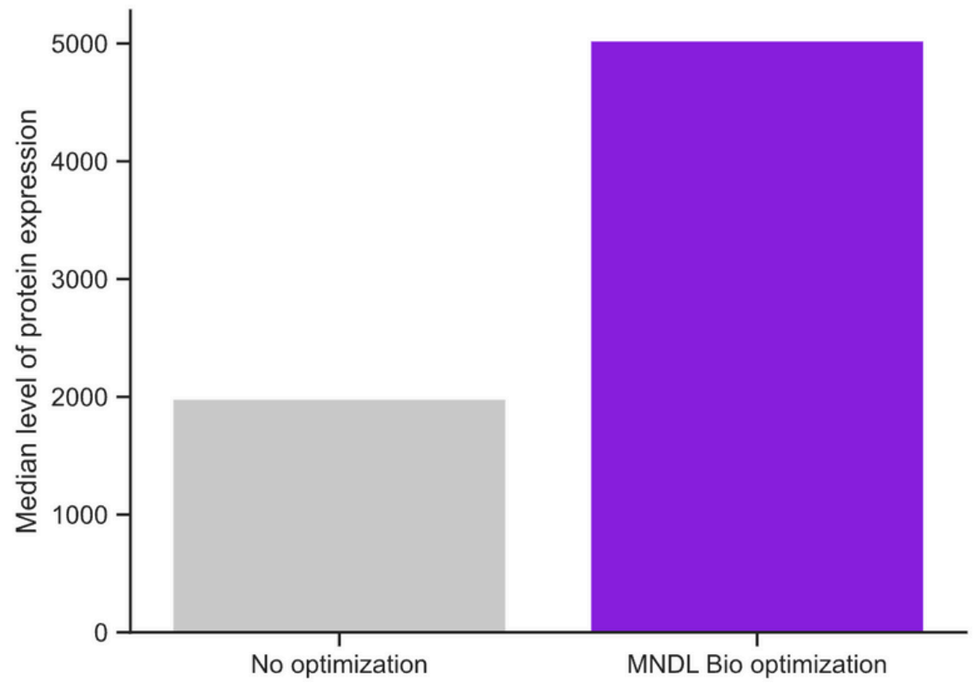


Figure 5. Expression of a growth factor in P.pastoris before and after MNDL optimization. The median expression levels of the growth factor of 100 transformants of the unoptimized sequence was ~ 2000. After MDNL Bio optimization, the lead variant provided a ~2.5-fold improvement in the median level of protein expression, with a value of ~ 5000.

Custom Projects

Some of MNDL Bio's more advanced algorithms are only available for custom projects and not through the online platform. Some of these algorithms have been published, while others are proprietary. They include:

Long-term Expression Stability
By coupling the expression of a target gene to a fitness-increasing gene, MNDL Bio improves the expression stability for plasmid-based and genomically integrated recombinant genes [6].

Copy Number Control
Design of plasmid origin of replication (Ori) for fine tuning of recombinant protein expression.

Antibody Design
An antibody-specific model for enhanced expression.

Bespoke Algorithms for Specific Projects
These include custom algorithms for non-conventional systems or unique design requirements such as novel inducible promoters or conditional expression.

Frequently Asked Questions

What host organisms are currently supported by MNDL Bio's platform?
Our platform currently supports optimization in 10 host organisms, with more hosts onboard soon. The current hosts are: *Aspergillus niger*, *Bacillus subtilis*, *Crisetulus griseus* (CHO), *E. coli*, *Homo sapiens*, *Nicotina tabacum*, *Pichia pastoris* (*Komagataella phaffii*), *Saccharomyces cerevisiae*, *Spodopetra frugipedra*, and *Trichoderma reesei*.

I'm only interested in 1 variant, the best one for my case. Why do I need more?
MNDL Bio uses a variety of algorithms and models to design variants. Some of the algorithms target specific stages of the gene expression process, while others employ deep learning methods that are agnostic to specific knowledge about gene expression. Moreover, we often use combinations of approaches. Each target gene and host combination presents a unique case and there is currently no way of knowing what combination will work best. Instead, MNDL Bio provides its users with variants that are likely to increase yield but cannot currently be ranked from best to worst before testing.

How many sequence variants can I get and how long does it take?
You can have as many as 10 variants per gene, and it usually takes up to an hour. You will be notified by email when your variants are ready for downloading.

Does the platform assist in construct design?
Yes, we currently support construct design for *B. subtilis*, *E. coli*, and *P. pastoris*, with more hosts added shortly. This includes a choice of a signal peptide, and N- and C-termini fusions.

References

[1] Sidi et. al. (2024) Predicting gene sequences with AI to study codon usage patterns. PNAS
[2] Zur and Tuller (2015) Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. Bioinformatics
[3] Diamant et. al. (2018) ChimeraUGEM: unsupervised gene expression modeling in any given organism. Bioinformatics
[4] Menuhin-Gruman et. al. (2022) Evolutionary Stability Optimizer (ESO): A Novel Approach to Identify and Avoid Mutational Hotspots in DNA Sequences While Maintaining High Expression Levels. ACS Synthetic Biology
[5] Neumann and Tuller (2022) Modeling the ribosomal small subunit dynamic in Saccharomyces cerevisiae based on TCP-seq data. Nucleic Acids Research
[6] Menuhin-Gruman et. al. (2025). AI-directed gene fusing prolongs the evolutionary half-life of synthetic gene circuits. bioRxiv