AI-Driven Gene Optimization:

Increasing Recombinant Protein Yield through Construct-Level Engineering

Introduction

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, our AI-driven approach accounts for complex biological interactions that impact gene expression and co-optimizes coding and non-coding regions.

Our platform, based on 15 years of research, incorporates:

- **AI-Driven codon optimization**: Leveraging 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 modelling**: Balancing recombinant protein expression with cellular fitness to enhance plasmid stability and maintain high yields [4]
- **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

- **Improving DNA synthesis success rate:** Smart recoding or alteration of problematic sequences in coding and non-coding regions, respectively

Online Platform Features

MNDL Bio's platform (https://app.mndl.bio) 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 (50 nucleotides minimum). Other features include:

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

For a video demo of the platform, see the following (https://mndl.bio/intro-video)

Optimization usually takes a few minutes, following which the optimized sequence variants are ready to download along with a report that describes how the variants were designed.

Case Studies

Case Study 1

- **Objective:**

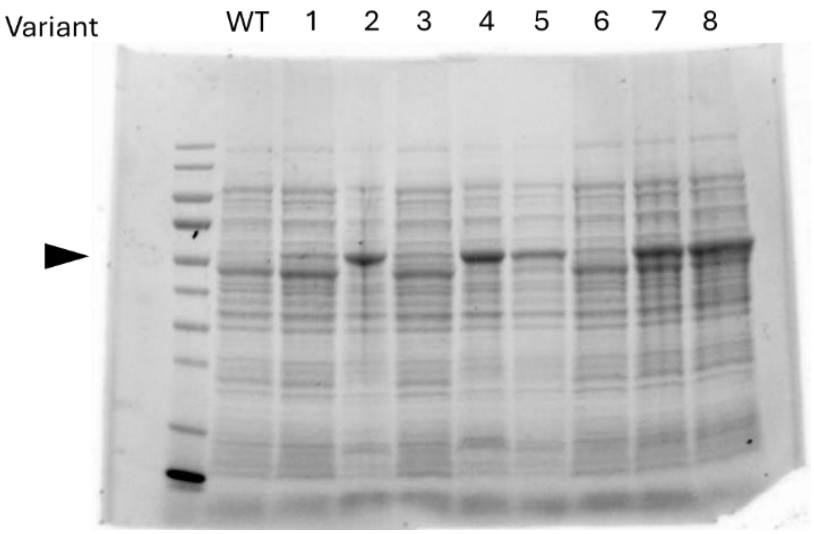
Increase expression of Glucose-6-phosphate dehydrogenase (G6PD) in *Escherichia coli* BL21(DE3).

- **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:**

Most of MNDL Bio's engineered variant showed increased yield of the full-length enzyme.



Case study 2

- **Objective**:

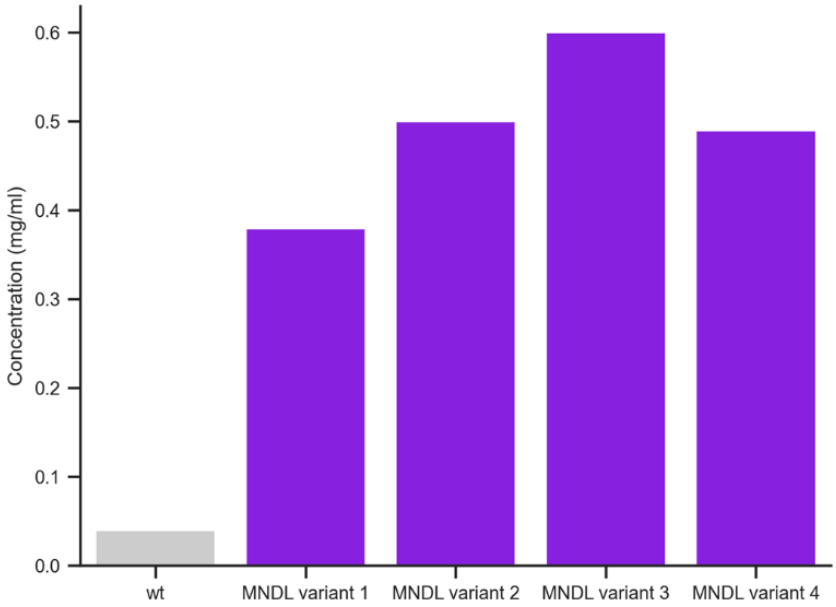
  Increase expression of a ~15KDa protein in an *E. coli*- based cell-free system.

- **Baseline situation and solution:**

The WT sequence had very low expression and was not usable for scale-up. We used a combination of our algorithms and models to design sequence variants and tested their expression in an *E. coli*- based cell-free system.

- **Results:**

Our ML, deep-learning, and biophysical-designed variants increased expression by ~9.5 to 15-fold.



Case Study 3

- **Objective**:

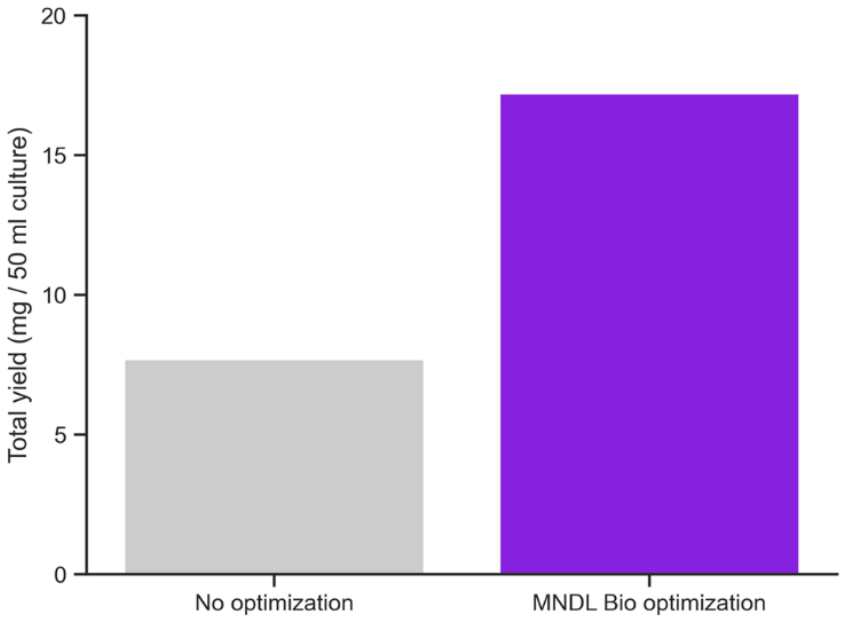
Increase expression of an engineered enzyme in *E. coli*.

- **Baseline situation and solution**:

The engineered enzyme yield did not meet expression goals. MNDL Bio designed 10 variants to be screened.

- **Results:**

One of the variants showed a ~2.5-fold increase in expression.



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].
- **Antibody design:** An antibody-specific model for enhanced expression.
- **Copy number control:** Desing of plasmid origin of replication (Ori) for fine tuning of recombinant protein 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. Contact us for more information at info@mndl.bio.

FAQ

Q: What host organisms are currently supported by MNDL Bio's platform?

A: 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), Escherichia coli, Homo sapiens, Nicotina tabacum, Pichia pastoris (Komagataella phaffii), Saccharomyces cerevisiae, Spodopetra frugipedra,* and *Trichoderma reesei*.

Q: Does the platform assist in construct design?

A: Yes, we currently support construct design for *Bacillus subtilis, Escherichia coli,* and *Pichia pastoris,* with more hosts added shortly. This includes a choice of a signal peptide, and N- and C-termini fusions.

Q: How many sequence variants can I get and how long does it take?

A: You can have as many as 20 variants per gene, and it usually takes several minutes. You will be notified by email when your variants are ready for downloading.

Q: I'm only interested in 1 variant, the best one for my case. Why do I need more?

A: 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.

Refences

[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

Licensing & Citations