PROJECT *UNMASK* · ERA AIxBIO FELLOWSHIP · WINTER 2026

Efficient Classification of DNA Sequences for Synthesis Order Screening

Researcher: **Ian D. Beatty**
University of North Carolina Greensboro
ian@ianbeatty.com

Mentor: **Gary Abel**
Fourth Eon Bio

THE PROBLEM

DNA synthesis screening is a critical bioterrorism safeguard, but expense inhibits widespread adoption. Human review is slow and costly, homology lookup is not robust against obscured, modified, or novel pathogens and toxins, and emerging functional prediction tools are too computationally expensive to apply indiscriminately.

Much natural and bioengineered DNA does not directly code for proteins, and tools suited for protein homology search or function prediction are wasted on non-protein sequences. Entirely abiological applications are a growing fraction of synthesis orders, and homology-based screening cannot efficiently distinguish them from potentially dangerous proteins.

Why not just use a foundation model?

Economics: DNA screening has a triage problem: providers process enormous order volumes. Screening methods must be fast and inexpensive per sequence.

Design mismatch: Protein language and genomic foundation models were designed for prediction and design, not categorization. They work on entire proteins and large genome sections, not short synthesis orders.

Robustness: These models were trained on natural sequences, not novel or deliberately obfuscated inputs.

OUR APPROACH

We propose a lightweight classification layer that uses amino acid (AA) sequence statistics to triage sequences before expensive screening, distinguishing between protein-coding and other DNA types.

Nucleic acid patterns and codon frequencies can be disguised by innocent codon optimization or adversarial obfuscation, but AA sequences are constrained by protein biophysics.

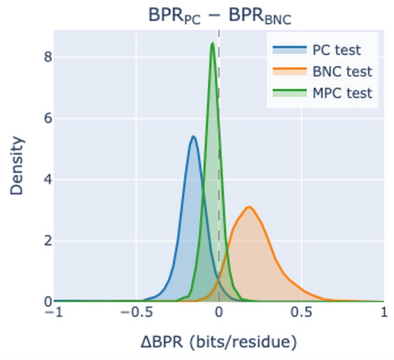
Our goal is routing, not clearance: protein-coding and non-coding sequences can be directed to appropriate category-specific screening, instead of sending all sequences through the same compute-intensive and/or human gauntlet.

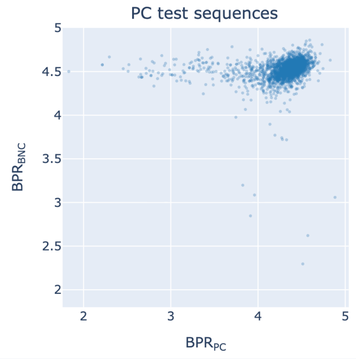
EXPLORATION 1: Can **compressibility** distinguish between coding and non-coding sequences? *Yes!* ✓

Method: Train three *prediction by partial matching* (PPM) compression models on protein-coding (PC), biological non-coding (BNC), and mis-translated protein-coding (MPC) amino acid sequences from human DNA. The model that best compresses a sequence (measured by *bits per residue*, BPR) indicates its category.

Key Findings:

- Binary coding/non-coding classification with depth-3 PPM-D models achieves an AUROC of 0.9802, with a false positive rate of **6% at 95% sensitivity** soaring to 60% at 99% sensitivity. *The last 5% of proteins are difficult to distinguish!*
- Only a subset of protein types (zinc fingers, keratins, etc.) have sufficient short-range structure to compress significantly with a PC-trained model, but…
- …classification succeeds primarily because most proteins don't compress particularly well under *any* model, while non-coding and mis-translated coding segments compress well under their own models. The classifier detects "doesn't look like a known kind of non-protein" rather than "looks like a protein."





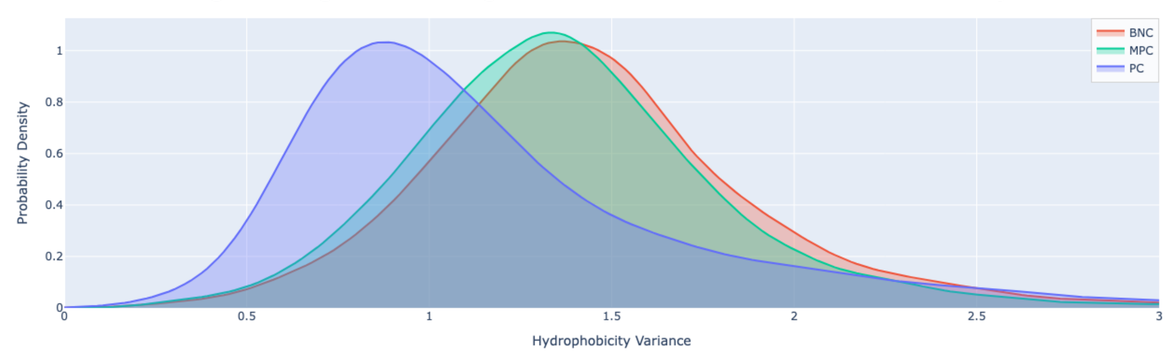
EXPLORATION 2: Do **biophysical metrics** improve classification beyond compressibility alone? *Yes!* ✓

Method: Define metrics targeting longer-range protein structure: amino acid distribution, charge clustering, hydrophobic moment, hydrophobicity variance, localized Fourier periodicity. Classify via differential compressibility *and* these metrics via machine learning methods.

Classifier	AUC	FP@95	FP@99
Gradient Boosting	0.9934	2.7%	14.0%
Random Forest	0.9928	3.4%	16.7%
Logistic Regression	0.9889	3.3%	24.3%
PPM only (depth-3 PPM-D)	0.9802	5.8%	60.8%
Gaussian naive Bayes	0.9790	8.5%	22.8%
linear SVM	0.9657	10.9%	57.0%

Key Findings:

- Gradient boosting with additional metrics **reduces false positive (FP) rate ~4x** (61% → 14%) at 99% sensitivity, relevant for reliable synthesis screening use.
- However, compressibility measures receive >99% of feature importance weight. Biophysical metrics contribute little… *Do better ones exist?*
- Hand-crafting, testing, and tuning metrics is difficult, slow, and arbitrary.



EXPLORATION 3: Can a **convolutional neural network** (CNN) learn to locate protein-coding intervals within sequences? In *any* reading frame? For a variety of natural and synthetic DNA sequence types? *It's looking promising!*

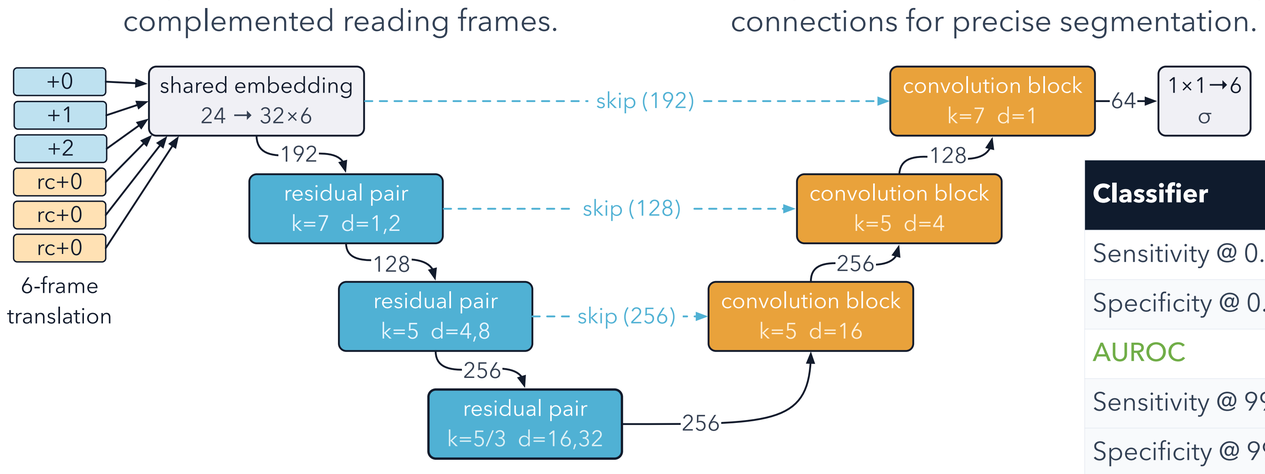
WORK IN PROGRESS

Training dataset:

- **Natural DNA** from protein-coding, fRNA-coding, gene-adjacent, and intergenic regions from 22 training species across eukaryotes, prokaryotes, and viruses. Five species held out for generalization testing.
- **Synthetic DNA** for engineered genes and backbones from *AddGene*, and BioBrick parts and composites from *iGEM*.
- **208,144 sequences** split into training (83%), validation (7.8%), and testing (9.2%) partitions.
- **Balanced** for clade and species, sequence length, and homo/heterogeneity (all protein-coding or non-coding, gene with flanks, dense multi-gene, synthetic composite).
- **Frames randomized** during training to ensure the network recognizes sequences in shifted and/or reverse-complemented reading frames.

Architecture:

- **Six-frame translation** of nucleotide sequence → amino acids (AAs) → vector embeddings, processed jointly.
- **1D dilated U-Net** encoder for wide receptive field (~400–500 AA), with skip connections for precise segmentation.
- **Six independent** sigmoid output channels for **per-frame, per-nucleotide protein-coding probability**, enabling precise segmentation and detection of overlapping or alternate-frame genes.
- **Compute:** ~1.5 hours to train 2.4M parameters, **~0.8 ms per sequence for classification** on one NVIDIA A40 GPU.



Classifier	validation partition	unseen species	test synth.	test compos.
Sensitivity @ 0.5 thresh.	0.9481	0.9495	0.9393	0.9562
Specificity @ 0.5 thresh.	0.8363	0.8417	0.8408	0.8681
AUROC	0.9642	0.9676	0.9619	0.9627
Sensitivity @ 99% spec.	0.4605	0.5988	0.3698	0.2941
Specificity @ 99% sens.	0.6579	0.7128	0.7648	0.6824

SO WHAT?

1. Protein-coding DNA sequences can be distinguished from non-coding sequences with high accuracy **based on amino acid sequence statistics alone**.
2. Short-range correlations are highly discriminatory.
3. **Biophysical property measures** improve classification accuracy.
4. A convolutional neural network can learn to recognize protein-coding intervals with **per-frame, per-nucleotide resolution**.
5. Assembling good datasets is much harder than building neural networks!

Further Work

- **Investigate CNN performance details:** segmentation precision, accuracy for short sequences or protein segments, and dominant failure modes.
- **Tune and improve:** vary hyperparameters, architecture, and training set.
- **Include abiological sequences:** DNA origami, data storage, and computation.
- **Evaluate adversarial robustness:** out-of-distribution cases and possible strategies to obfuscate coding sequences.
- **Extend to multi-category classification:** probabilities for natural protein, engineered protein, functional RNA, vector construct, or abiological sequence.