# Gradient-based Optimization for mRNA Sequence Design The ID3 (Input Data Differentiable Designer) Framework

Li Hongmin, Goro Terai, Takumi Otagaki, Kiyoshi Asai
Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo

Paper

GitHub

## Introduction

### The mRNA Design Challenge

- Critical for COVID-19 vaccines (Pfizer/BioNTech, Moderna)
- Multiple objectives: accessibility, stability, translation efficiency
- Hard constraint: Preserve amino acid sequence

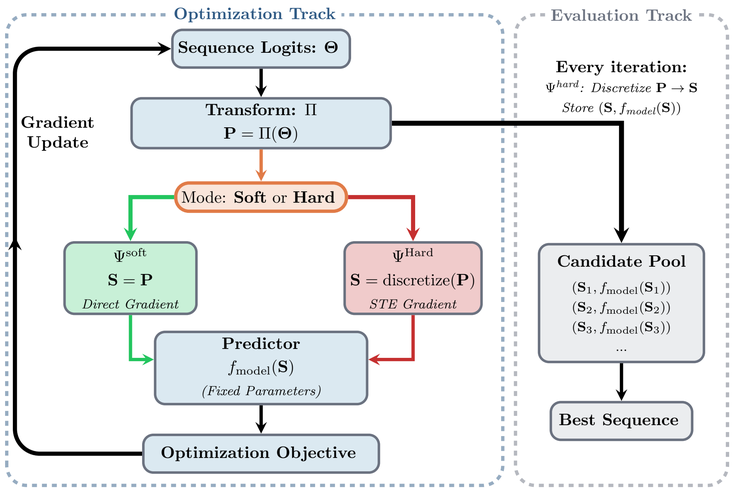
## The Problem

**The discrete-continuous gap prevents gradient-based optimization**

### Existing Approaches Fall Short:

1. **Discrete optimization** (GA, SA): Slow, no theoretical guarantees
2. **Straight-Through Estimator**: Lacks systematic constraint mechanisms
3. **No unified framework** for gradient optimization + biological constraints

## Core Innovation: Decoupled Architecture




**Key Insight:** Models accept probability distributions as input

### Core Objective:

$$\min_{\Theta} \mathcal{L} = \min_{\Theta} f_{\text{model}}(T(\Theta)), \quad T = \Psi \circ \Pi$$

### ID3 Pipeline:

$$\Theta \xrightarrow{\Pi} \mathbf{P} \xrightarrow{\Psi} \mathbf{S} \xrightarrow{f_{\text{model}}} \mathcal{L}, \quad \mathcal{L} \dashrightarrow \Theta \text{ via } \nabla$$

- $\Pi$: Parameter-to-Probability (Gumbel-Softmax)
- $\Psi$: Probability-to-Sequence (Soft/Hard modes)
- $\nabla$: Gradient-based updates

**Transformation Functions:**

*Parameter-to-Probability* ($\Pi$):

$$\mathbf{P}^{\text{det}} = \text{softmax}(\Theta/\sigma)$$
$$\mathbf{P}^{\text{sto}} = \text{softmax}((\Theta + g)/\sigma)$$

where $g$ is Gumbel noise, $\sigma$ is temperature parameter

*Probability-to-Sequence* ($\Psi$):

$$\psi^{\text{soft}}(\mathbf{P}) = \mathbf{P}$$
$$\psi^{\text{hard}}(\mathbf{P}) = \text{discretize}(\mathbf{P})$$

where discretize uses Straight-Through Estimator (STE)

**Four Operational Modes:**

| Mode | Parameter | Output |
|---|---|---|
| Det.Soft | Deterministic | Continuous |
| Det.Hard | Deterministic | Discrete |
| Sto.Soft | Stochastic | Continuous |
| Sto.Hard | Stochastic | Discrete |

## Three Constraint Mechanisms

**Challenge:** Optimize RNA while preserving amino acids

### 1. Codon Profile Constraint

*Parameter Domain:*

$$\Theta_{\text{codon}} = \{\theta_{j,c} : c \in \mathcal{C}(y_j)\}$$

*Probability Computation:*

$$\mathbf{P}_{\text{codon}} = \text{softmax}(\theta_{j,c}/\sigma)_{c \in \mathcal{C}(y_j)}$$

- Mathematically impossible to violate constraints
- 40-60% reduction: $\sum_{j=1}^{M} |\mathcal{C}(y_j)|$ vs $L \times 4$

### 2. Amino Matching Softmax

*Parameter Domain:*

$$\Theta \in \mathbb{R}^{L \times 4} \quad \text{(full RNA space)}$$

where $\Theta^{(j)} \in \mathbb{R}^{3 \times 4}$ for $j$-th amino acid's codon

*Probability Computation:*

$$\mathbf{P}_{\text{codon},c} = \text{softmax}(\langle \Theta^{(j)}, \mathcal{E}[c] \rangle / \sigma)_{c \in \mathcal{C}(y_j)}$$

- $\mathcal{E}[c] \in \{0,1\}^{3 \times 4}$: One-hot encoding of codon $c$
- $\langle \Theta^{(j)}, \mathcal{E}[c] \rangle$: Inner product similarity
- Only valid codons get non-zero probabilities

### 3. Lagrangian Multiplier

*Parameter Domain:*

$$\Theta \in \mathbb{R}^{L \times 4}, \quad \mathbf{P} \in [0,1]^{L \times 4}$$

*Total Loss with Penalty:*

$$\mathcal{L}_{\text{total}} = f_{\text{model}}(T(\Theta)) + \lambda \cdot \mathcal{C}(\mathbf{P})$$

*Constraint Violation Measure:*

$$\mathcal{C}(\mathbf{P}) = \frac{1}{M} \sum_{j=1}^{M} \min_{c \in \mathcal{C}(y_j)} \|\mathbf{P}^{(j)} - \mathcal{E}[c]\|^2$$

L2 distance to closest valid codon

*Adaptive Multiplier Update:*

$$\lambda^{(t+1)} = \max(0, \lambda^{(t)} + \eta_t \cdot g^{(t)}), \quad \eta_t = \eta_0/\sqrt{t+1}$$

Subgradient method with decaying step size

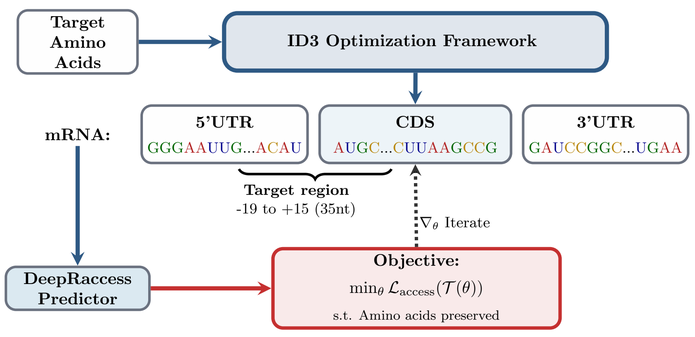
**4 Modes × 3 Mechanisms = 12 Variants**

## Applications

### 1. Accessibility Optimization with DeepRaccess[1]

*Loss Function:*

$$\mathcal{L} = f_{\text{access}}(T(\Theta))_{\text{ATG}-19:+15}$$
$$\text{s.t.} \quad \text{AA}(T^{\text{hard}}(\Theta)) = \mathbf{y}$$




*Goal: Minimize RNA accessibility in translation initiation region*

[1]DeepRaccess predictor: Hara et al., Frontiers in Bioinformatics, 3:1275787, 2023

### 2. CAI Co-Optimization

*CAI-aware Discretization:*

$$\Psi^{\text{hard}}_{\text{CAI}}(\mathbf{P}_{\text{codon}}) = \arg\max_{S} P(S|\mathbf{P}_{\text{codon}})$$
$$\text{s.t.} \quad \text{CAI}(S) \geq \text{threshold}$$

*Implementation:*

$$\mathbf{P}(\gamma) = \gamma \cdot \mathbf{w}_{\text{organism}} + (1-\gamma) \cdot \mathbf{P}_{\text{codon}}$$

Binary search finds optimal $\gamma \in [0,1]$ satisfying CAI threshold

*Alternative: Multi-objective Joint Loss*

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{access}} + \lambda_{\text{CAI}} \cdot \mathcal{L}_{\text{CAI}}$$

Gradient-based, flexible, works for both soft and hard modes

## Theoretical Guarantees

**First convergence proof for sequence optimization:** Gradient descent converges to stationary points under Lipschitz continuity, covering all 12 variants. See supplementary materials for detailed proofs.

## Experimental Setup

**Experiments:**

- 12 variants × 12 proteins
- 20 independent runs
- 1000 evaluations/run

**Baselines:**

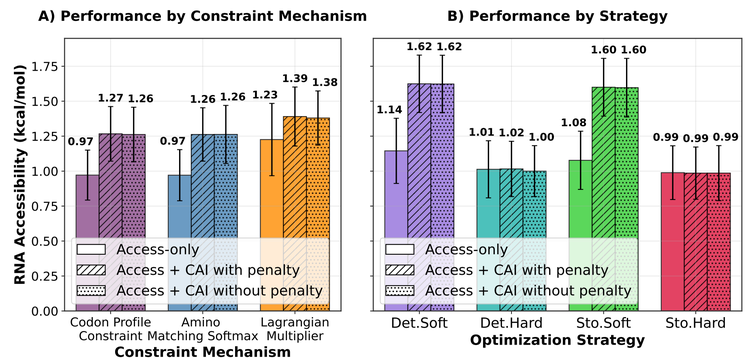
- Genetic Algorithm (GA)
- Simulated Annealing (SA)
- C10: Exhaustive (1.36M)

**Dataset (12 proteins, 64-1273 AA)**

| ID | Protein | AA |
|---|---|---|
| O15263 | Defensin beta 4A | 64 |
| P99999 | Cyt c (H.s.) | 105 |
| P00004 | Cyt c (E.c.) | 105 |
| P01308 | Insulin | 110 |
| P01825 | Immuno. HC | 116 |
| P31417 | FABP2 | 132 |
| P61626 | Lysozyme C | 148 |
| P42212 | GFP | 238 |
| P04637 | p53 | 393 |
| P0DTC9 | SARS-CoV-2 N | 419 |
| P0CG48 | Polyubiquitin-C | 685 |
| P0DTC2 | SARS-CoV-2 Spike | 1273 |

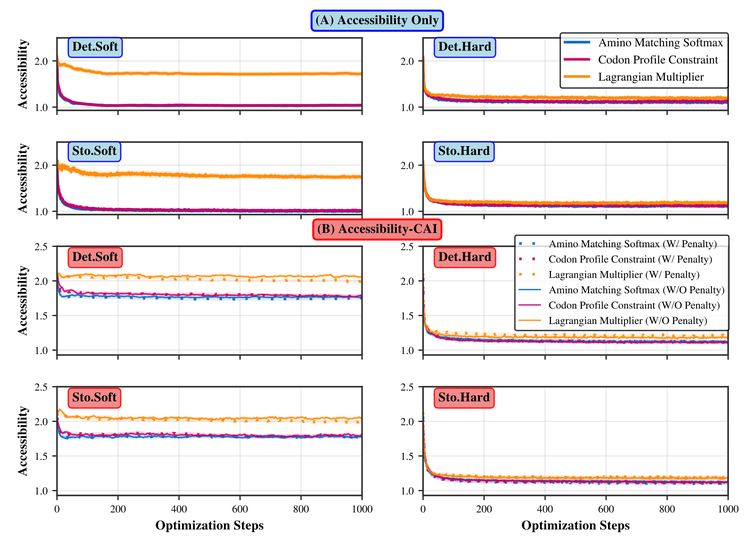
## Mechanism & Strategy Analysis

### Constraint Mechanism Comparison




Codon Profile and Amino Matching achieve similar performance

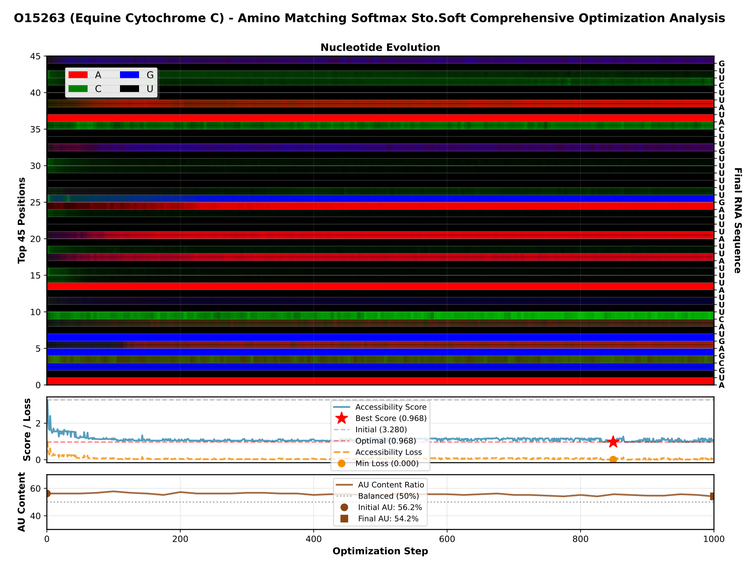
### Convergence by Strategy




Soft and hard comparable for single objective; hard superior for multi-objective

## Case Study: O15263 (Defensin beta 4A)

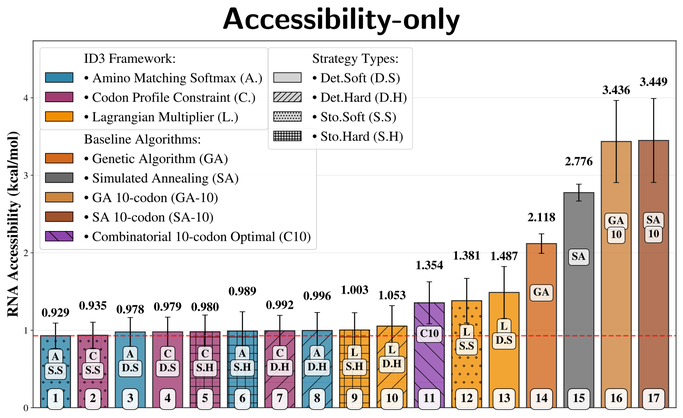
### Optimization Trajectory using Amino.Sto.Soft:




Rapid convergence: 3.28 → 0.97 kcal/mol with systematic AU balancing

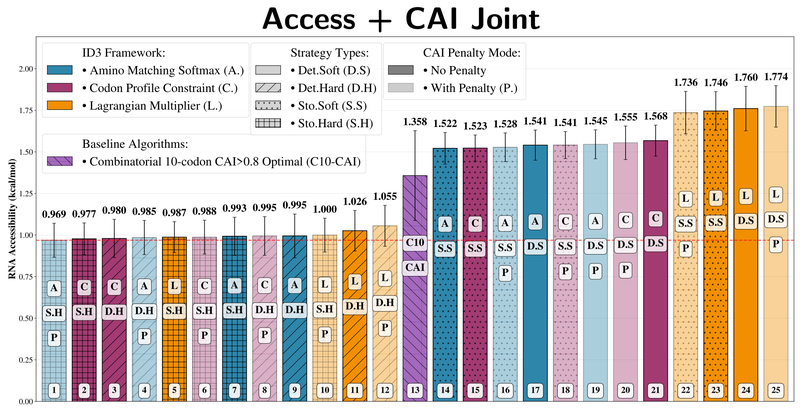
## Performance Comparison & Conclusion

### Accessibility-only




**10/12 variants outperform all baselines**

### Access + CAI Joint




**Hard constraints excel in multi-objective optimization**

### Conclusion

**Key Contributions:**

- Decoupled optimization-evaluation architecture
- Three systematic constraint mechanisms
- First convergence proof for sequence optimization
- 12 systematic variants covering optimization landscape
- Superior performance across diverse proteins

**Impact:**

- Principled foundation for gradient-based discrete sequence optimization
- Applications in mRNA vaccines and therapeutics
- Extensible to DNA and protein design

**Reference:** Li Hongmin, Goro Terai, Takumi Otagaki, Kiyoshi Asai. "Gradient-based Optimization for mRNA Sequence Design." *bioRxiv* 2025.10.22.683691

Li Hongmin | lihongmin@edu.k.u-tokyo.ac.jp | li-hongmin.github.io | GitHub: github.com/Li-Hongmin/ID3
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan