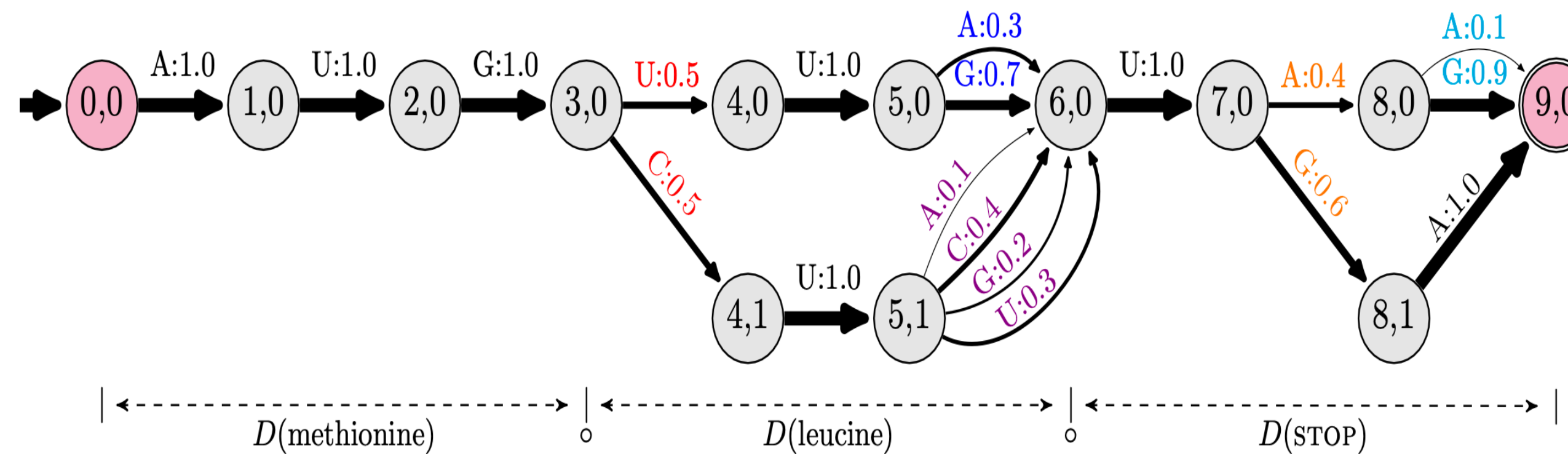
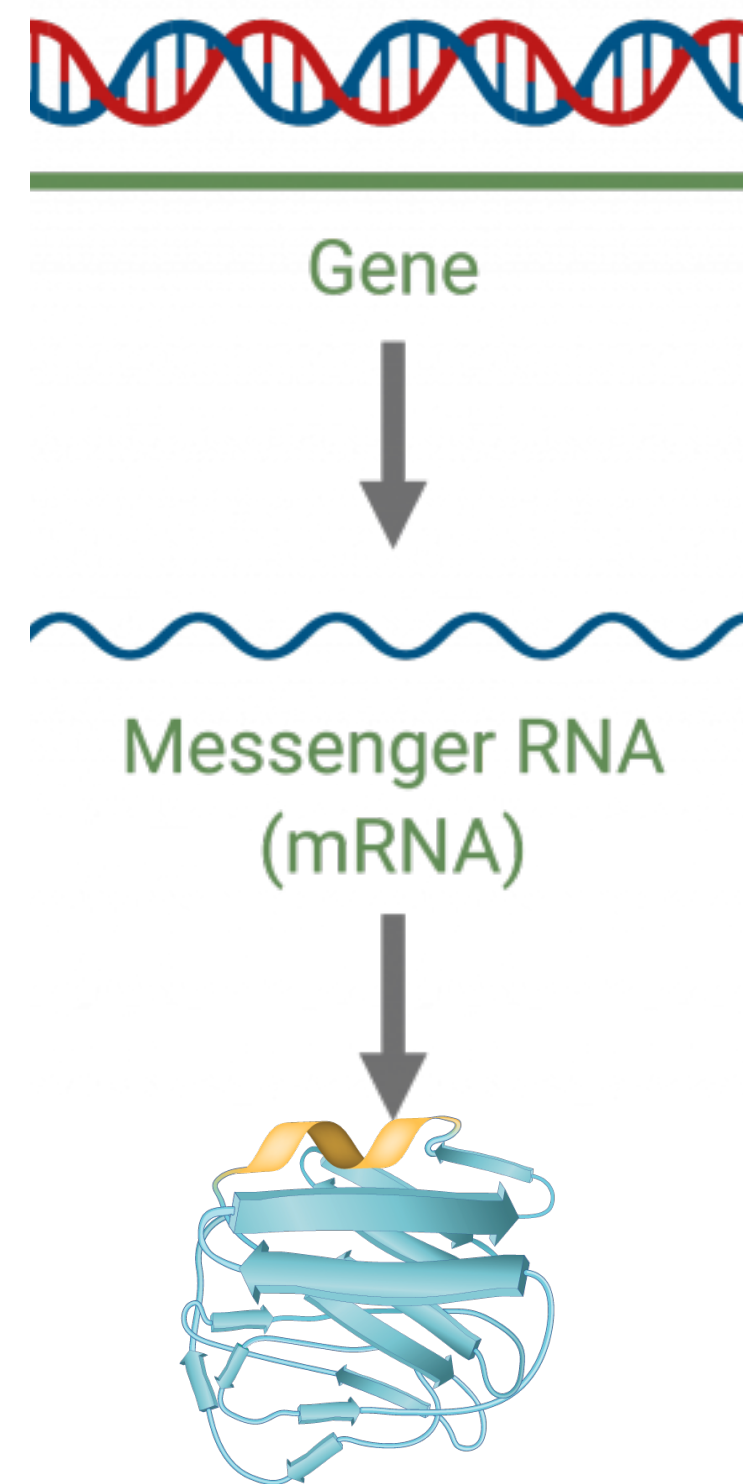
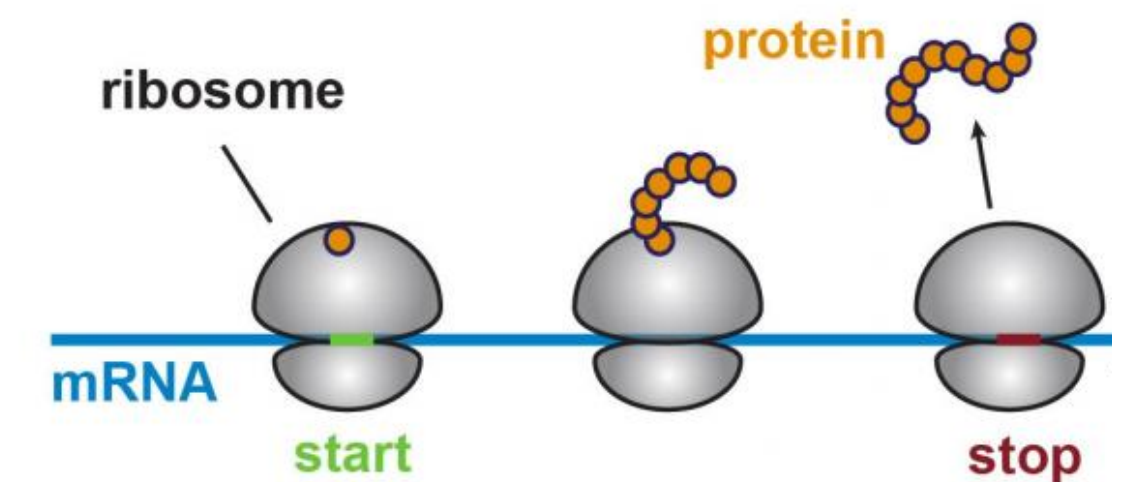
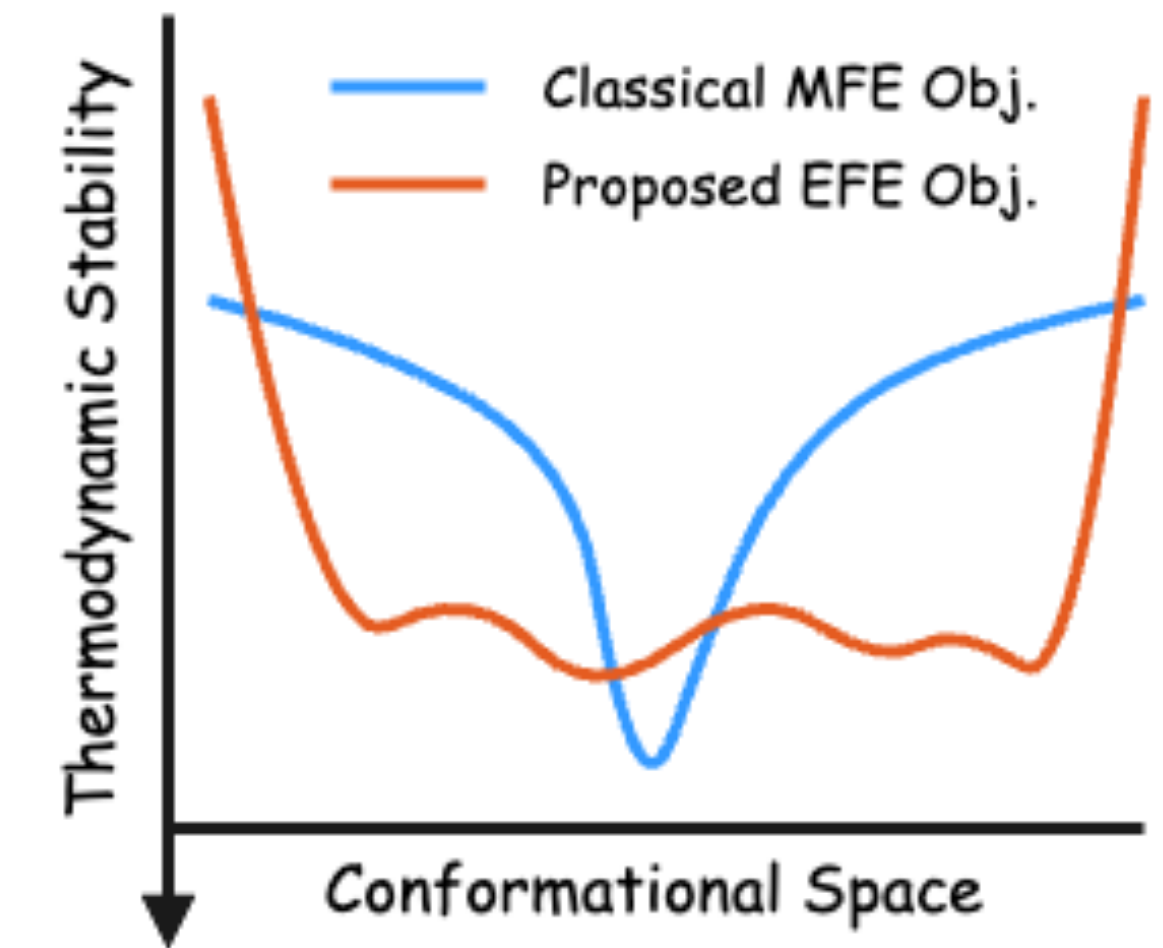


EnsembleDesign: mRNA Design Minimizing Ensemble Free Energy via Probabilistic Lattice Parsing



AUGUUUGA



Ning Dai, Tianshuo Zhou, Wei Yu Tang, David H. Mathews*, and Liang Huang



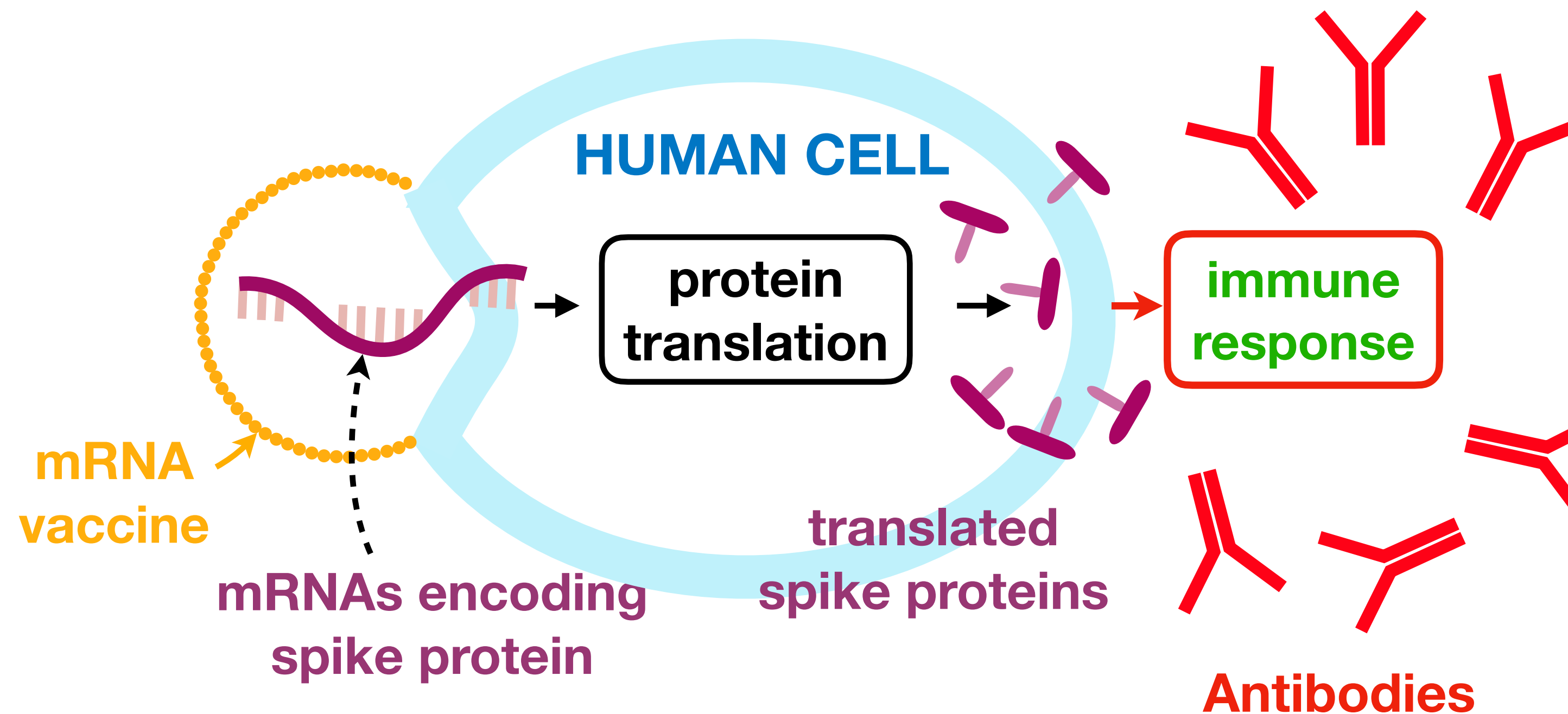
Proceedings Talk



Goal: Designing Stable mRNAs

- mRNA vaccine (and therapeutics) has many advantages, but suffers from **instability**
 - mRNA molecules degrade easily => lower protein expression => lower immunogenicity
- question: how to design more stable & efficient mRNA sequences computationally?

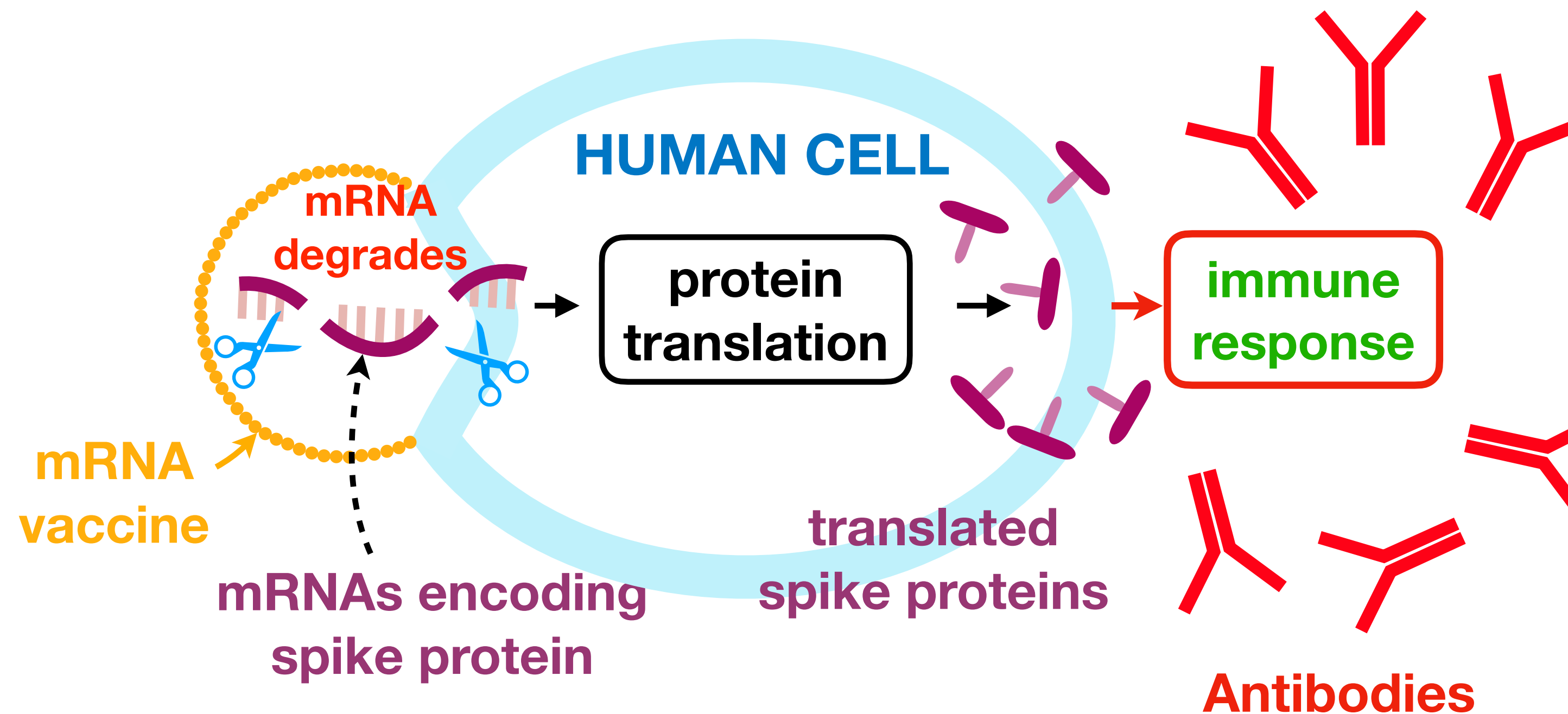
Translation  mRNA → Protein	
mRNA vaccine ("source code")	protein vaccine ("executables")
fast mass production	slow production
no risk of infection	small risk of infection
unstable	stable



Goal: Designing Stable mRNAs

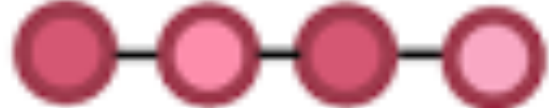
- mRNA vaccine (and therapeutics) has many advantages, but suffers from **instability**
 - mRNA molecules degrade easily => lower protein expression => lower immunogenicity
- question: how to design more stable & efficient mRNA sequences computationally?

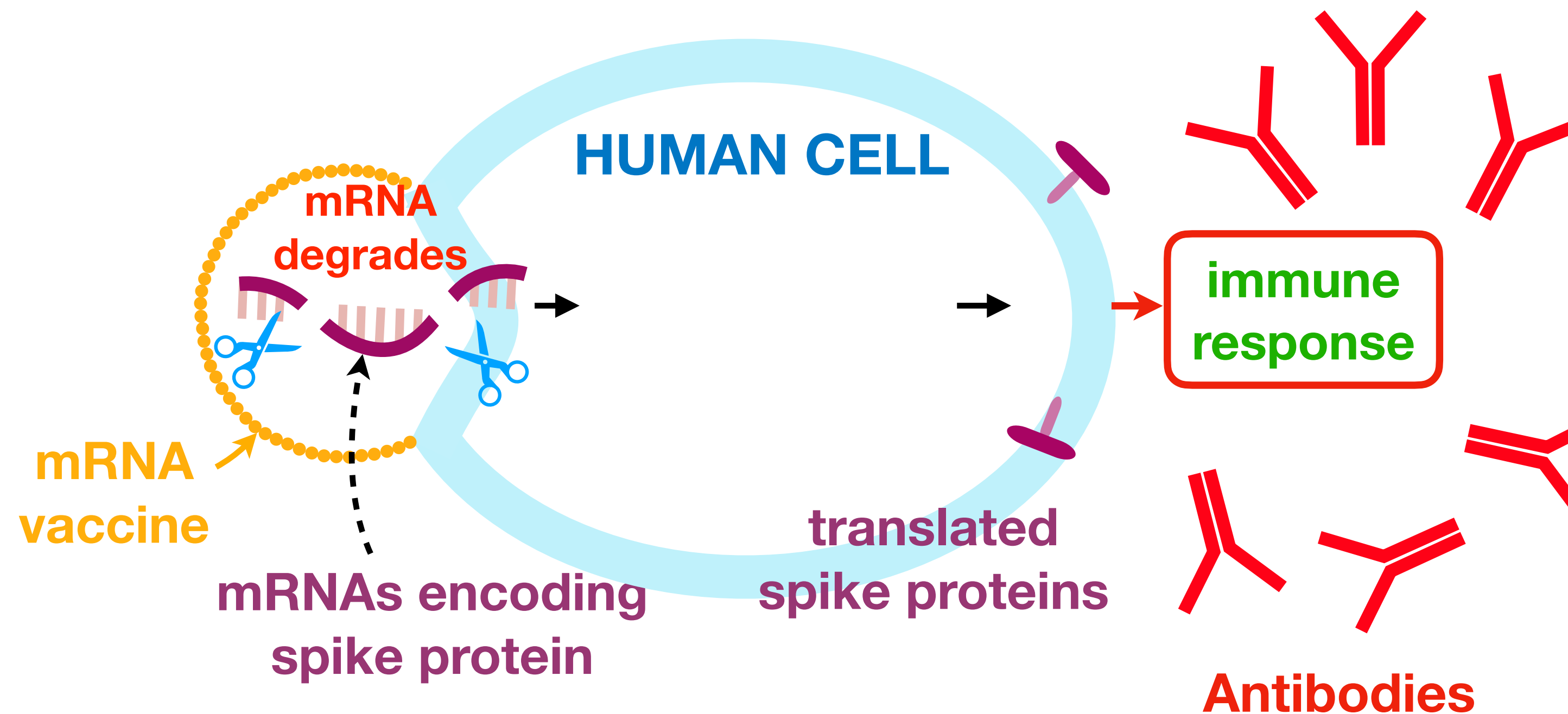
Translation  mRNA → Protein	
mRNA vaccine ("source code")	protein vaccine ("executables")
fast mass production	slow production
no risk of infection	small risk of infection
unstable	stable



Goal: Designing Stable mRNAs

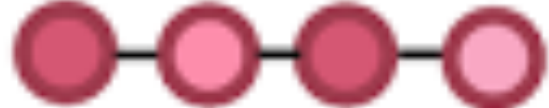
- mRNA vaccine (and therapeutics) has many advantages, but suffers from **instability**
 - mRNA molecules degrade easily => lower protein expression => lower immunogenicity
- question: how to design more stable & efficient mRNA sequences computationally?

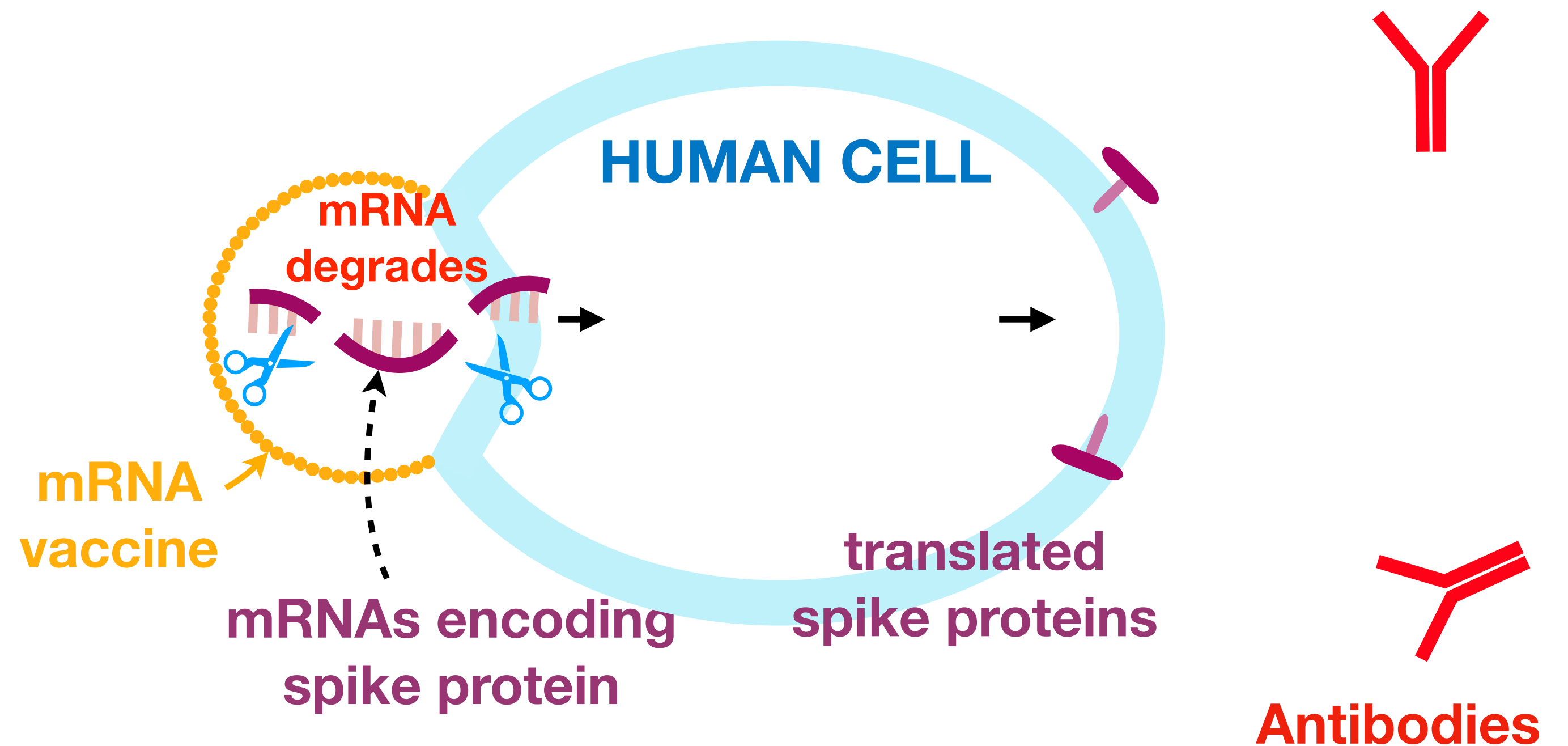
Translation	
	
mRNA	Protein
mRNA vaccine ("source code")	protein vaccine ("executables")
fast mass production	slow production
no risk of infection	small risk of infection
unstable	stable



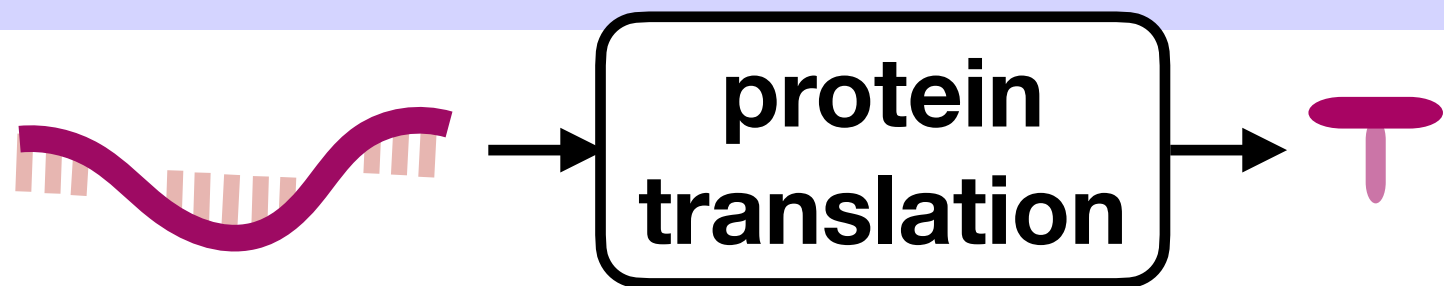
Goal: Designing Stable mRNAs

- mRNA vaccine (and therapeutics) has many advantages, but suffers from **instability**
 - mRNA molecules degrade easily => lower protein expression => lower immunogenicity
- question: how to design more stable & efficient mRNA sequences computationally?

Translation	
	
mRNA	Protein
mRNA vaccine ("source code")	protein vaccine ("executables")
fast mass production	slow production
no risk of infection	small risk of infection
unstable	stable



mRNA Design: the Mathematical Problem

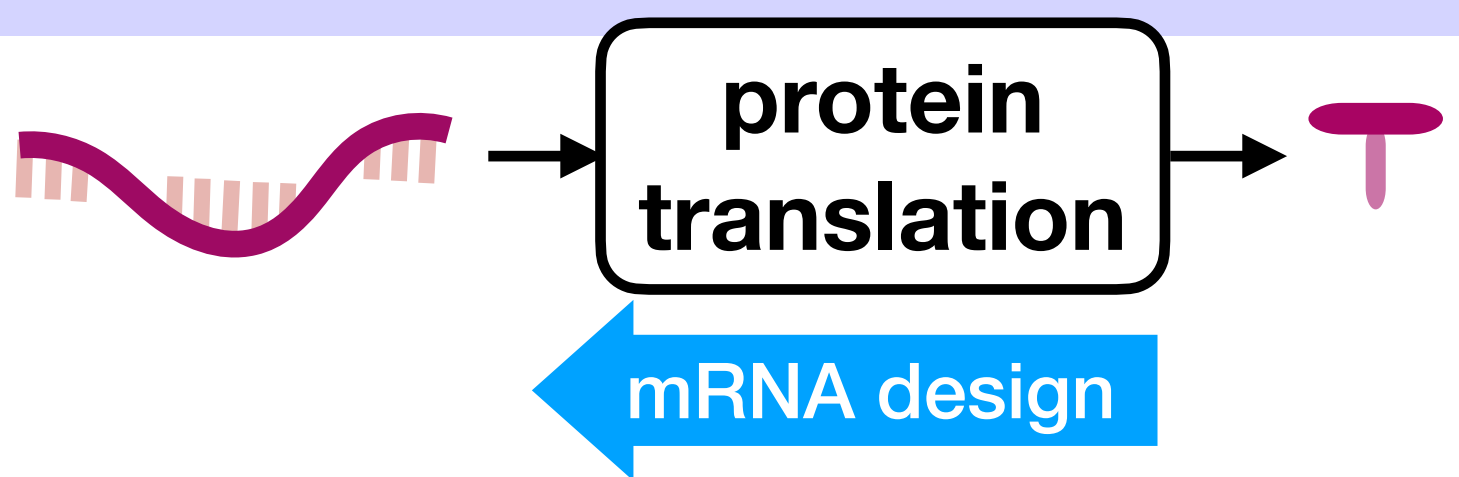


Spike mRNA A U G U U U G U U U U U C U U . . . A C A U A A

RNA codons						amino acid
GCU	GCC	GCA	GCG			Ala
CGU	CGC	CGA	CGG	AGA	AGG	Arg
AAU	AAC					Asn
GAU	GAC					Asp
UGU	UGC					Cys
CAA	CAG					Gln
GAA	GAG					Glu
GGU	GGC	GGA	GGG			Gly
CAU	CAC					His
AUU	AUC	AUA				Ile
CUU	CUC	CUA	CUG	UUA	UUG	Leu
AAA	AAG					Lys
AUG						Met
UUU	UUC					Phe
CCU	CCC	CCA	CCG			Pro
UCU	UCC	UCA	UCG	AGC	AGU	Ser
ACU	ACC	ACA	ACG			Thr
UAU	UAC					Tyr
UGG						Trp
GUU	GUC	GUA	GUG			Val
UAA	UAG	UGA				Stop

- mRNA design: inverse problem of protein translation
- exploit redundancy in the genetic code
 - each amino acid can be translated by multiple codons
- *exponentially* many synonymous mRNA sequences for a protein
- Moderna: stable mRNAs => high protein yield (Mauger et al, *PNAS* 2019)
- Q: how to find the most stable sequence in this huge space?

mRNA Design: the Mathematical Problem



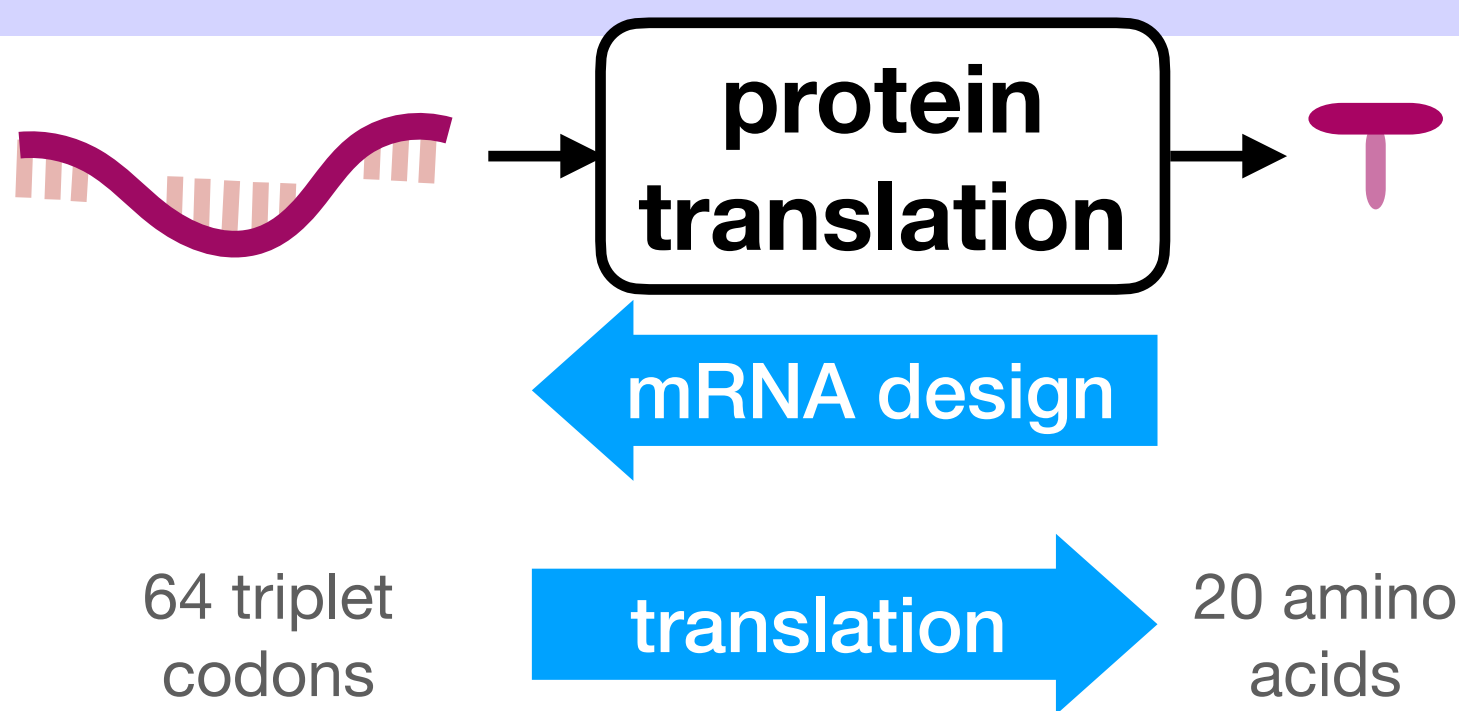
64 triplet codons → translation → 20 amino acids

RNA codons						amino acid
GCU	GCC	GCA	GCG			Ala
CGU	CGC	CGA	CGG	AGA	AGG	Arg
AAU	AAC					Asn
GAU	GAC					Asp
UGU	UGC					Cys
CAA	CAG					Gln
GAA	GAG					Glu
GGU	GGC	GGA	GGG			Gly
CAU	CAC					His
AUU	AUC	AUA				Ile
CUU	CUC	CUA	CUG	UUA	UUG	Leu
AAA	AAG					Lys
AUG						Met
UUU	UUC					Phe
CCU	CCC	CCA	CCG			Pro
UCU	UCC	UCA	UCG	AGC	AGU	Ser
ACU	ACC	ACA	ACG			Thr
UAU	UAC					Tyr
UGG						Trp
GUU	GUC	GUA	GUG			Val
UAA	UAG	UGA				Stop

Spike mRNA	A U G	U U U	G U U	U U U	C U U	. . .	A C A	U A A
translation	↓	↓	↓	↓	↓		↓	↓
Spike protein (1273 amino acids)	Met	Phe	Val	Phe	Leu	. . .	Thr	Stop

- mRNA design: inverse problem of protein translation
- exploit redundancy in the genetic code
 - each amino acid can be translated by multiple codons
- *exponentially* many synonymous mRNA sequences for a protein
- Moderna: stable mRNAs => high protein yield (Mauger et al, *PNAS* 2019)
- Q: how to find the most stable sequence in this huge space?

mRNA Design: the Mathematical Problem



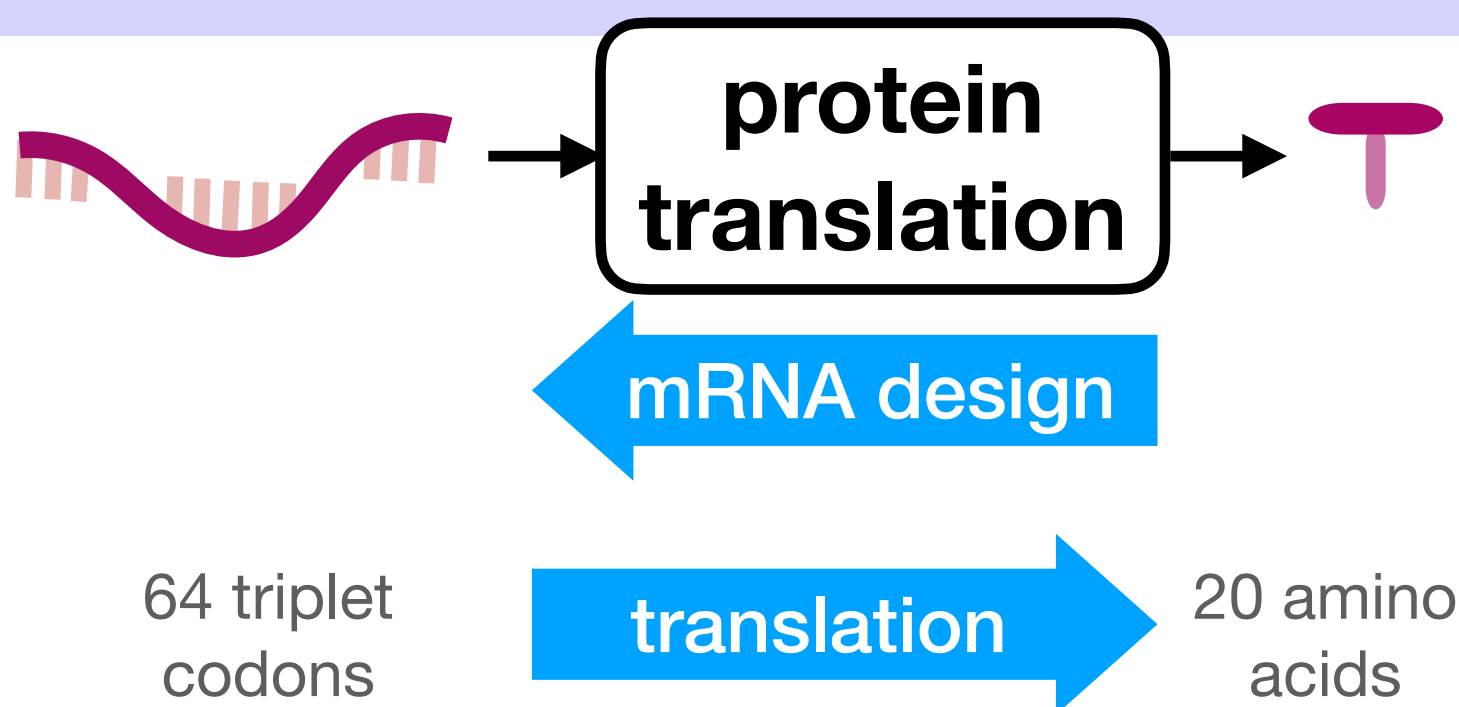
RNA codons						amino acid
GCU	GCC	GCA	GCG			Ala
CGU	CGC	CGA	CGG	AGA	AGG	Arg
AAU	AAC					Asn
GAU	GAC					Asp
UGU	UGC					Cys
CAA	CAG					Gln
GAA	GAG					Glu
GGU	GGC	GGA	GGG			Gly
CAU	CAC					His
AUU	AUC	AUA				Ile
CUU	CUC	CUA	CUG	UUA	UUG	Leu
AAA	AAG					Lys
AUG						Met
UUU	UUC					Phe
CCU	CCC	CCA	CCG			Pro
UCU	UCC	UCA	UCG	AGC	AGU	Ser
ACU	ACC	ACA	ACG			Thr
UAU	UAC					Tyr
UGG						Trp
GUU	GUC	GUA	GUG			Val
UAA	UAG	UGA				Stop

Spike mRNA translation	A U G	U U U	G U U	U U U	C U U	...	A C U	U A G
							A C C	U A G
							A C G	U G A
							A C A	U A A
Spike protein (1273 amino acids)	Met	Phe	Val	Phe	Leu	...	Thr	Stop

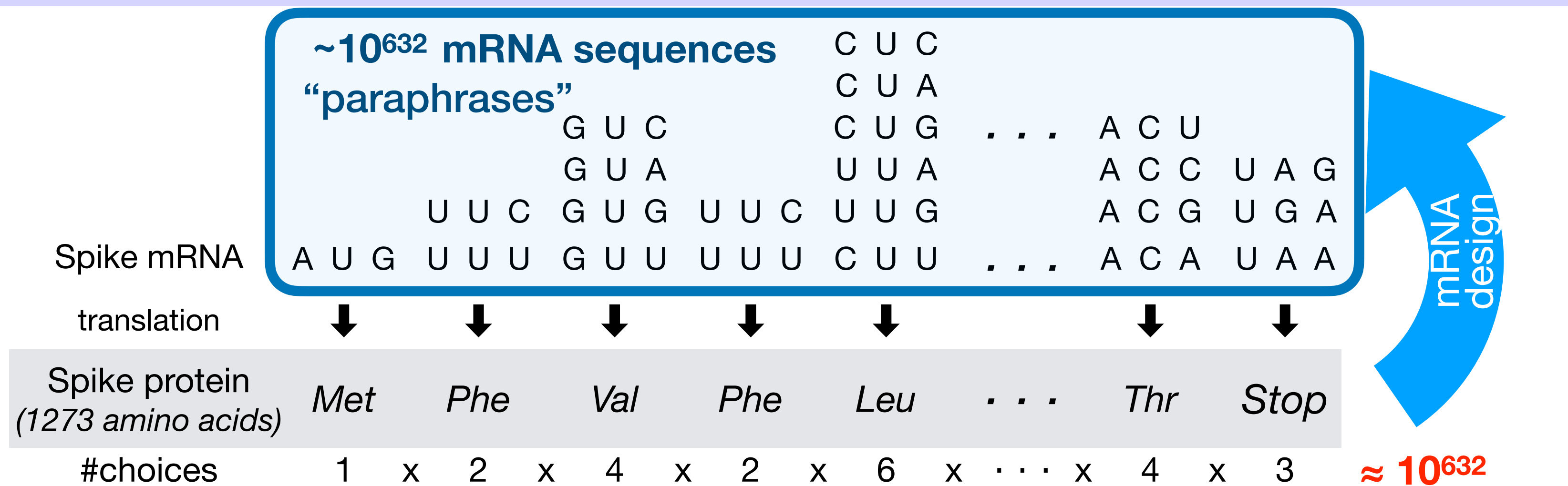


- mRNA design: inverse problem of protein translation
- exploit redundancy in the genetic code
 - each amino acid can be translated by multiple codons
- *exponentially* many synonymous mRNA sequences for a protein
- Moderna: stable mRNAs => high protein yield (Mauger et al, *PNAS* 2019)
- Q: how to find the most stable sequence in this huge space?

mRNA Design: the Mathematical Problem



RNA codons						amino acid
GCU	GCC	GCA	GCG			Ala
CGU	CGC	CGA	CGG	AGA	AGG	Arg
AAU	AAC					Asn
GAU	GAC					Asp
UGU	UGC					Cys
CAA	CAG					Gln
GAA	GAG					Glu
GGU	GGC	GGA	GGG			Gly
CAU	CAC					His
AUU	AUC	AUA				Ile
CUU	CUC	CUA	CUG	UUA	UUG	Leu
AAA	AAG					Lys
AUG						Met
UUU	UUC					Phe
CCU	CCC	CCA	CCG			Pro
UCU	UCC	UCA	UCG	AGC	AGU	Ser
ACU	ACC	ACA	ACG			Thr
UAU	UAC					Tyr
UGG						Trp
GUU	GUC	GUA	GUG			Val
UAA	UAG	UGA				Stop

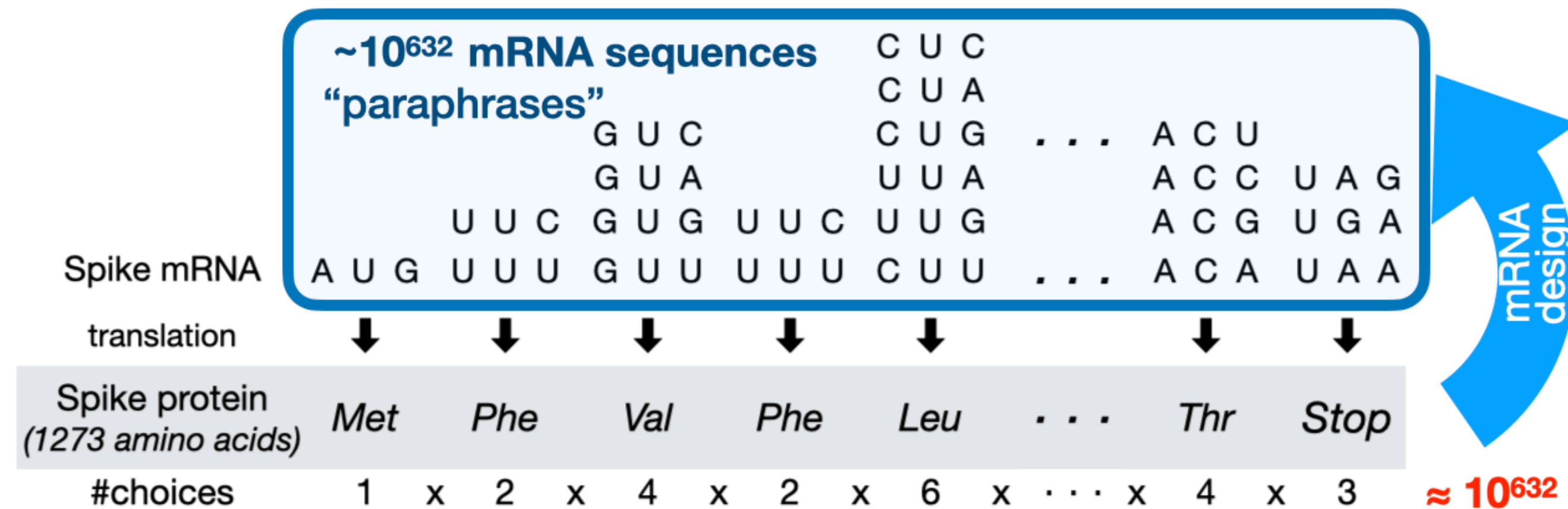


enumeration
takes ~10⁶¹⁶
billion years!

- mRNA design: inverse problem of protein translation
- exploit redundancy in the genetic code
 - each amino acid can be translated by multiple codons
- *exponentially* many synonymous mRNA sequences for a protein
- Moderna: stable mRNAs => high protein yield (Mauger et al, PNAS 2019)
- Q: how to find the most stable sequence in this huge space?

Conventional Objective (MFE) & LinearDesign

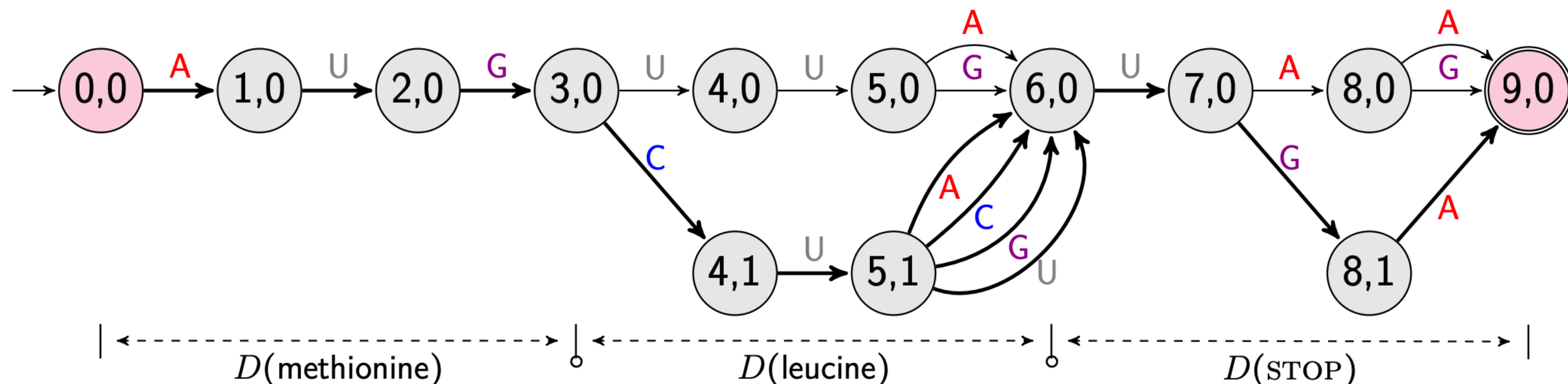
- stability metric: minimum free energy (MFE)
- among all mRNAs encoding the protein, find the one with the minimum MFE
- this is a minimization over a minimization



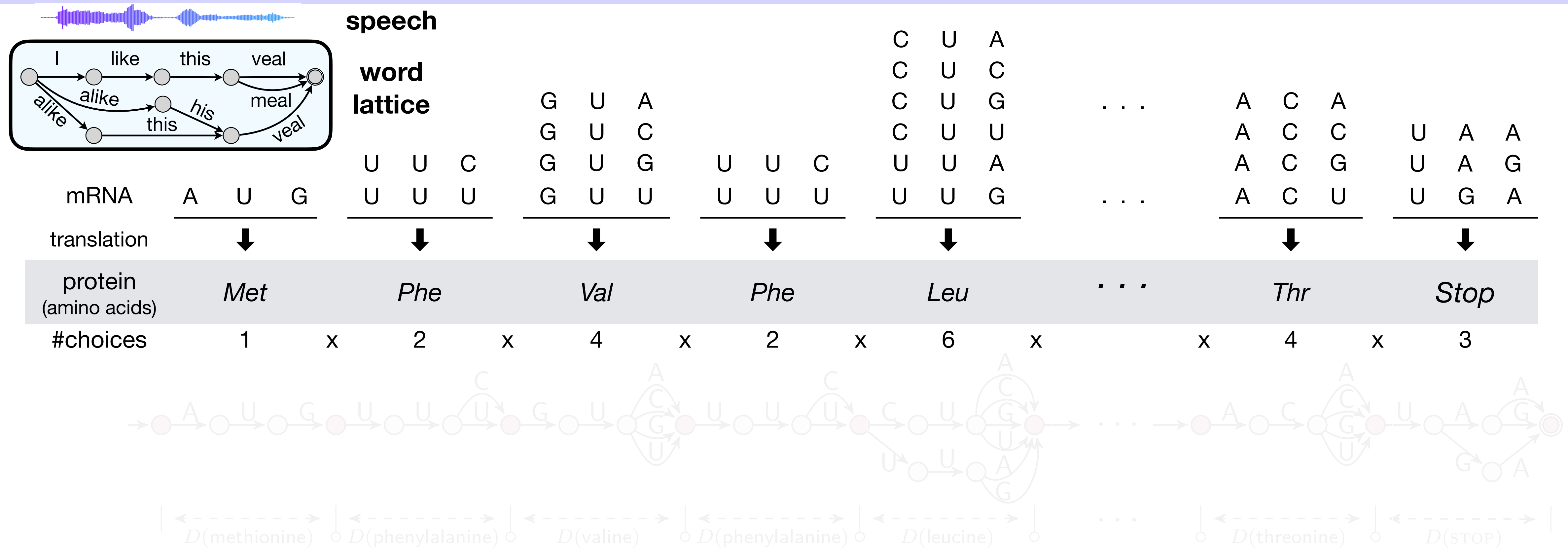
$$\min_{\mathbf{x} \in \mathcal{X}(\mathbf{p})} \text{MFE}(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{X}(\mathbf{p})} \min_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \Delta G^\circ(\mathbf{x}, \mathbf{y})$$

mRNA design space is exponentially large

- it can be solved by dynamic programming, e.g., via lattice parsing in LinearDesign (Nature, 2023)

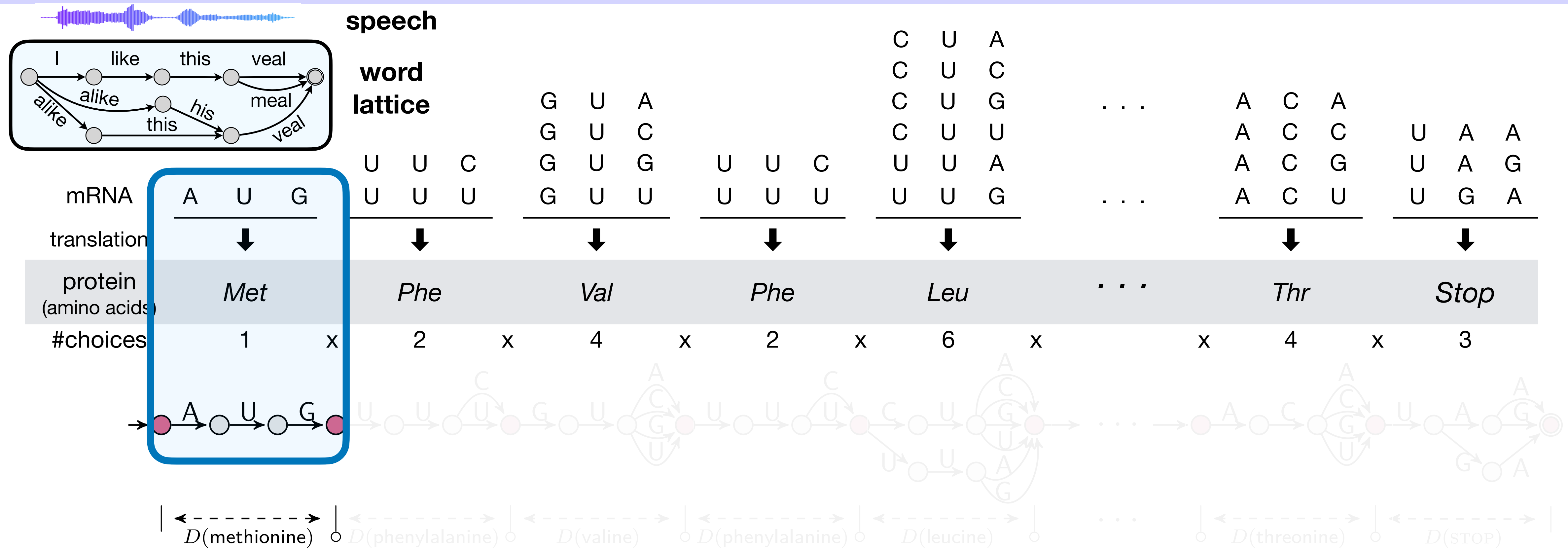


Linear Design Representation: Word Lattice



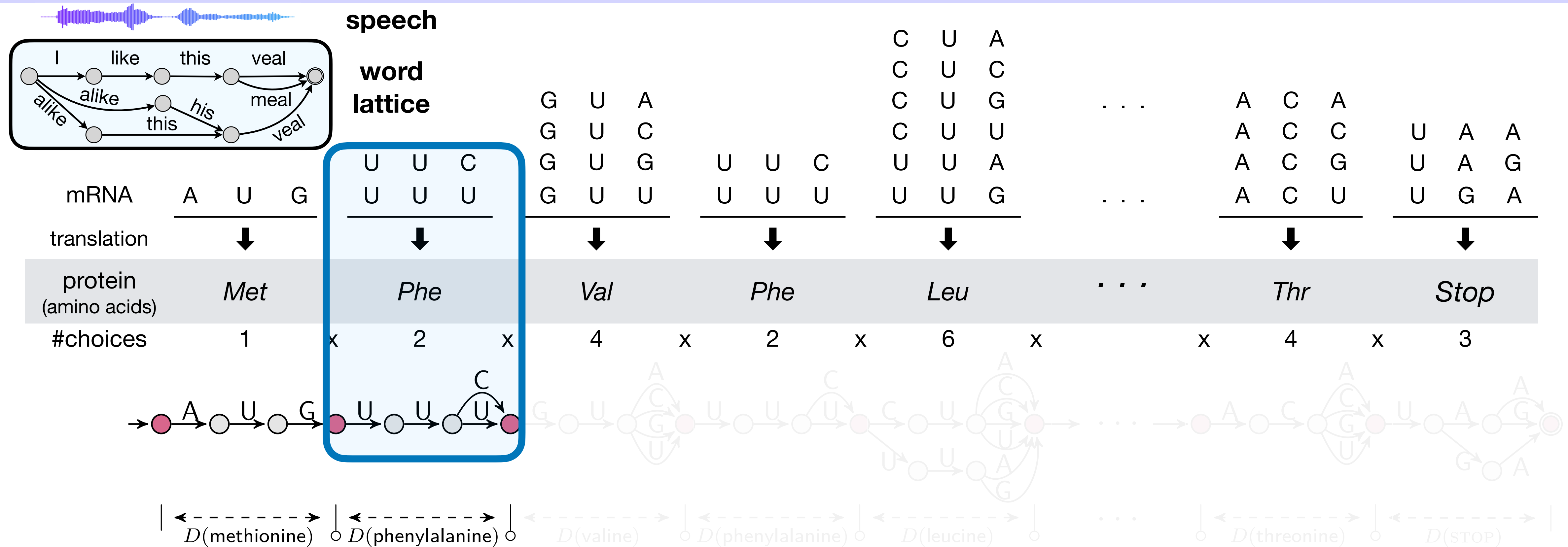
- represent mRNA design space as a lattice (finite-state automata)
- compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

Linear Design Representation: Word Lattice



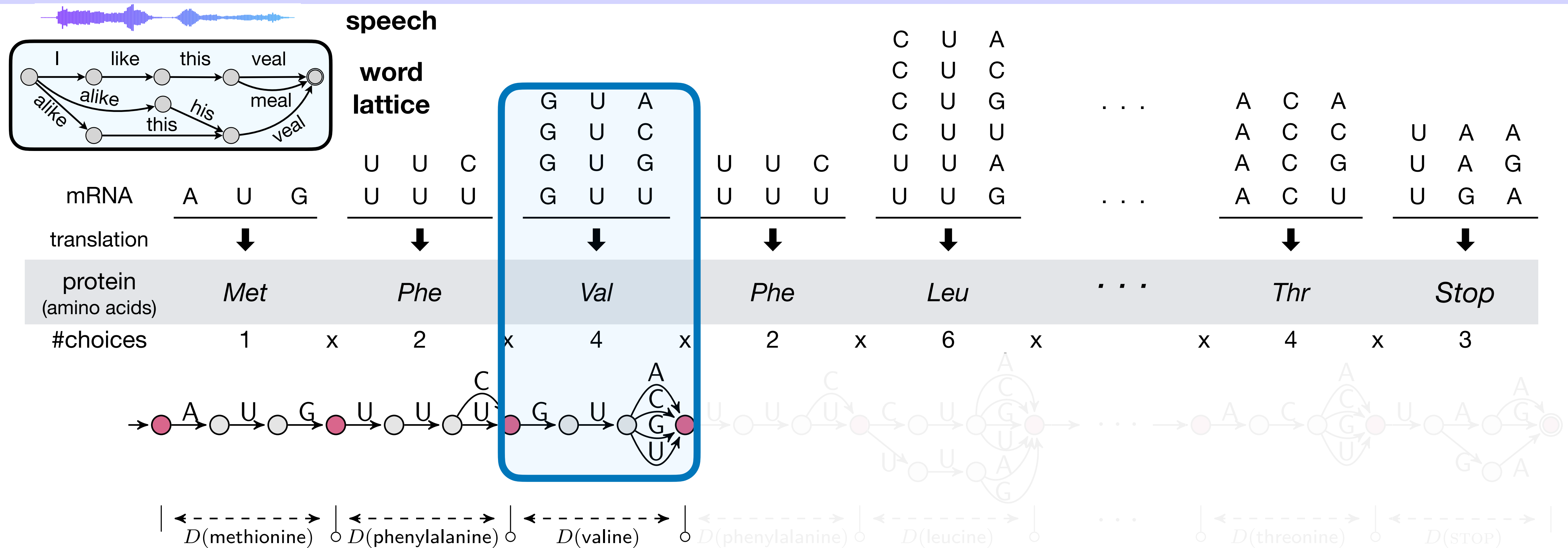
- represent mRNA design space as a lattice (finite-state automata)
- compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

Linear Design Representation: Word Lattice



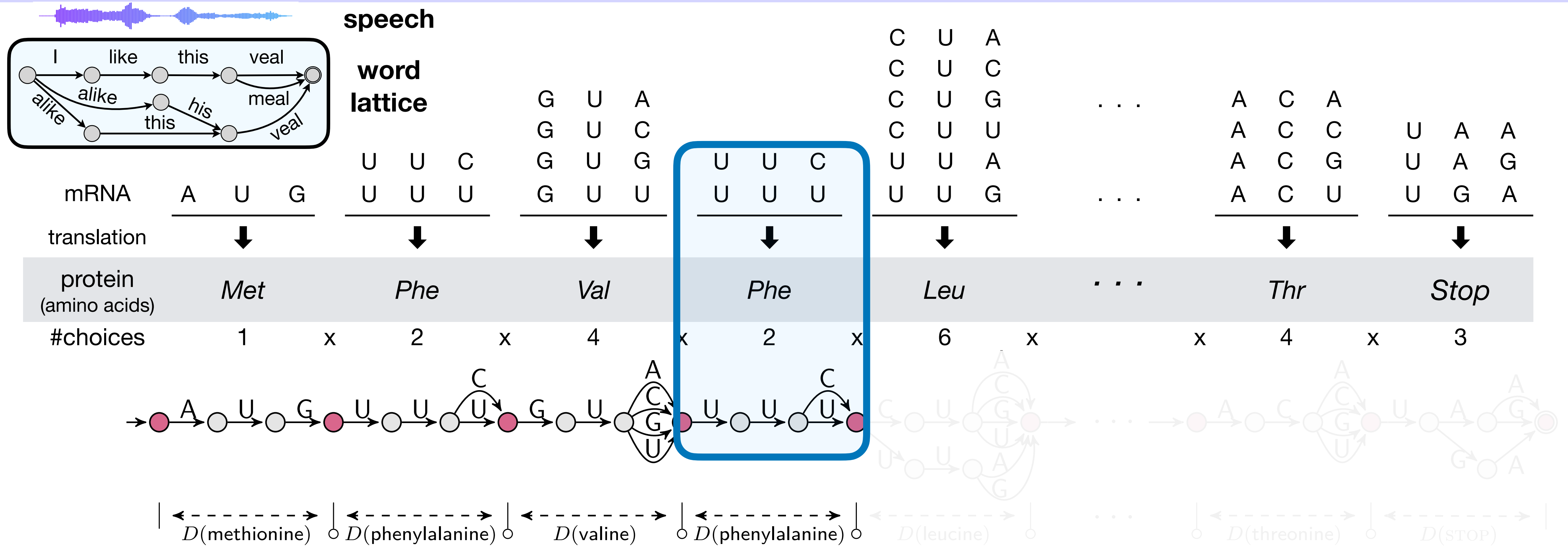
- represent mRNA design space as a lattice (finite-state automata)
- compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

Linear Design Representation: Word Lattice



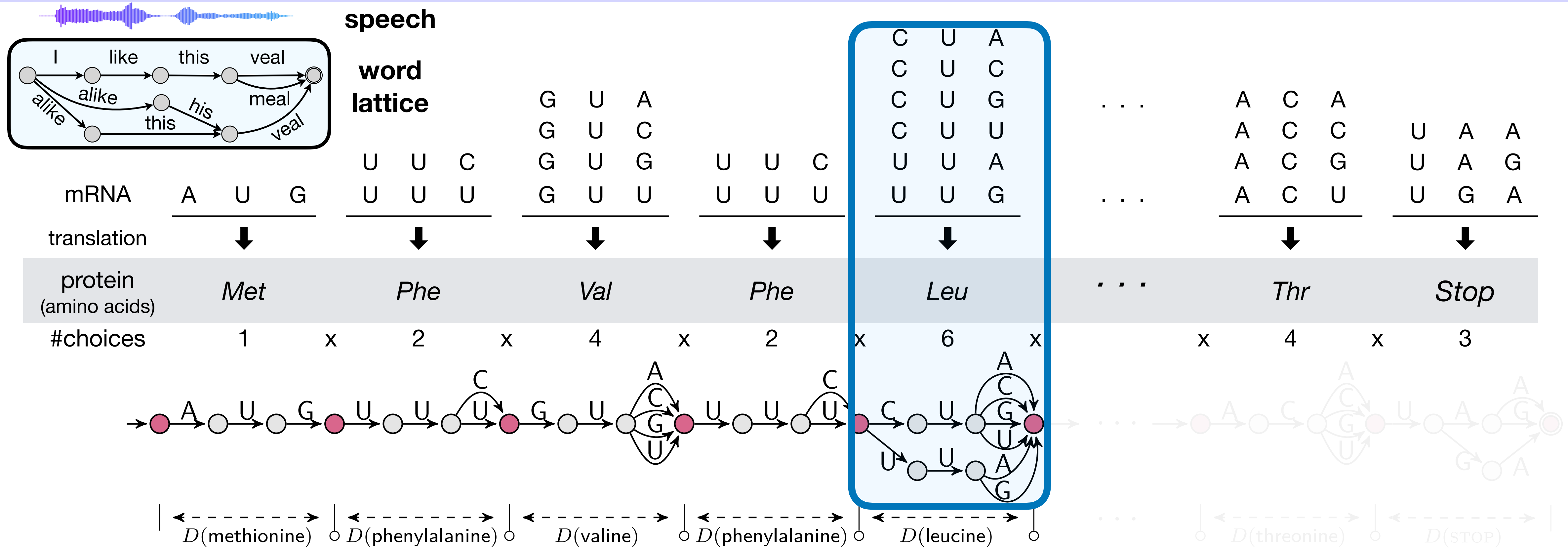
- represent mRNA design space as a lattice (finite-state automata)
- compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

LinearDesign Representation: Word Lattice



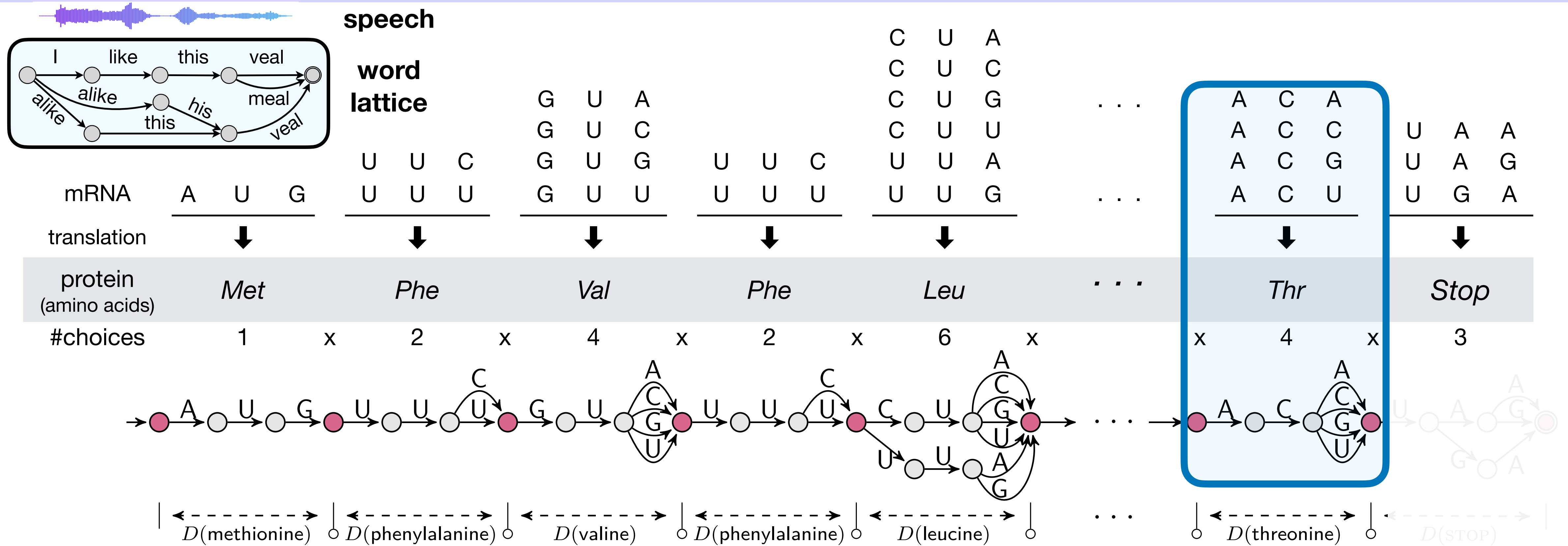
- represent mRNA design space as a lattice (finite-state automata)
 - compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

LinearDesign Representation: Word Lattice



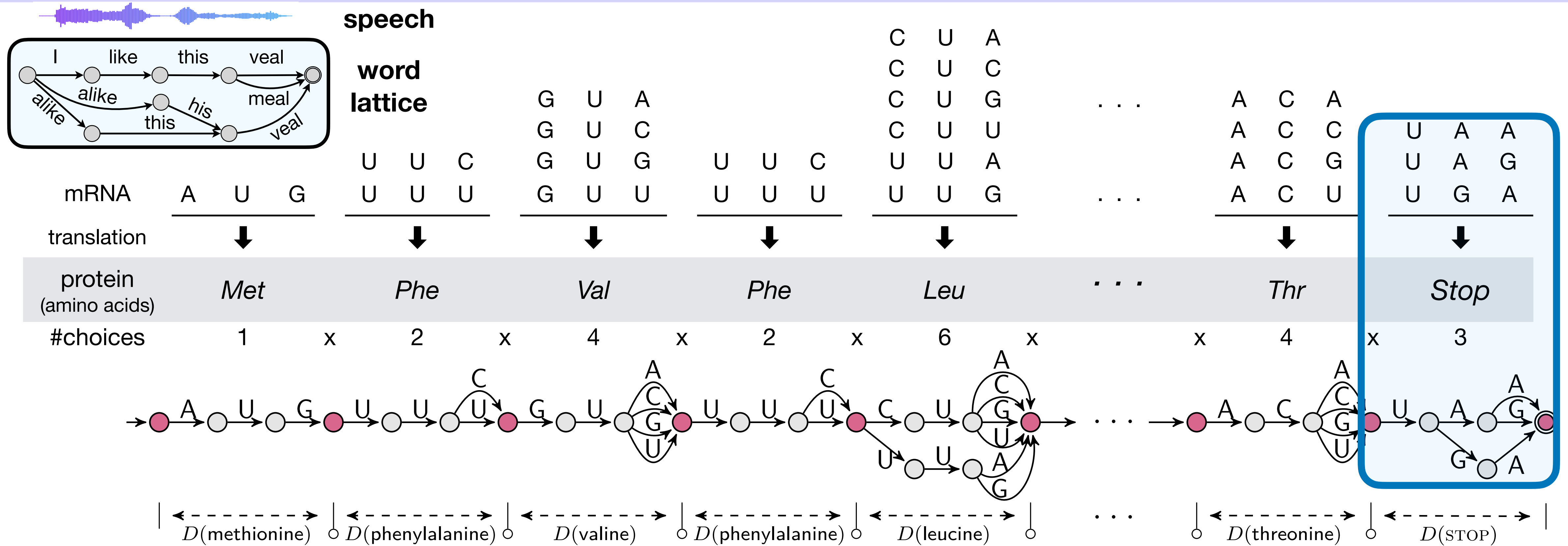
- represent mRNA design space as a lattice (finite-state automata)
 - compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

LinearDesign Representation: Word Lattice



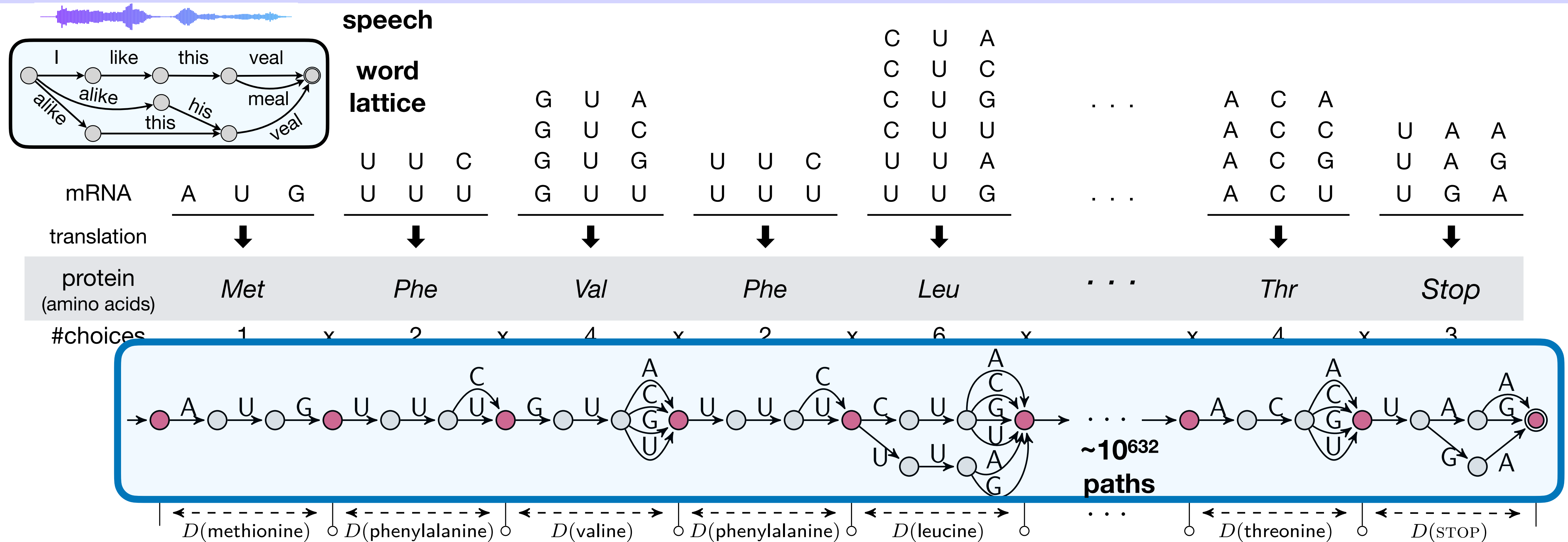
- represent mRNA design space as a lattice (finite-state automata)
 - compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

LinearDesign Representation: Word Lattice



- represent mRNA design space as a lattice (finite-state automata)
 - compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

Linear Design Representation: Word Lattice



- represent mRNA design space as a lattice (finite-state automata)
- compact (polynomial-sized) representation of exponentially many alternatives
- big question: how to find the most stable (lowest energy) path in this lattice?

LinearDesign Algorithm: Lattice Parsing

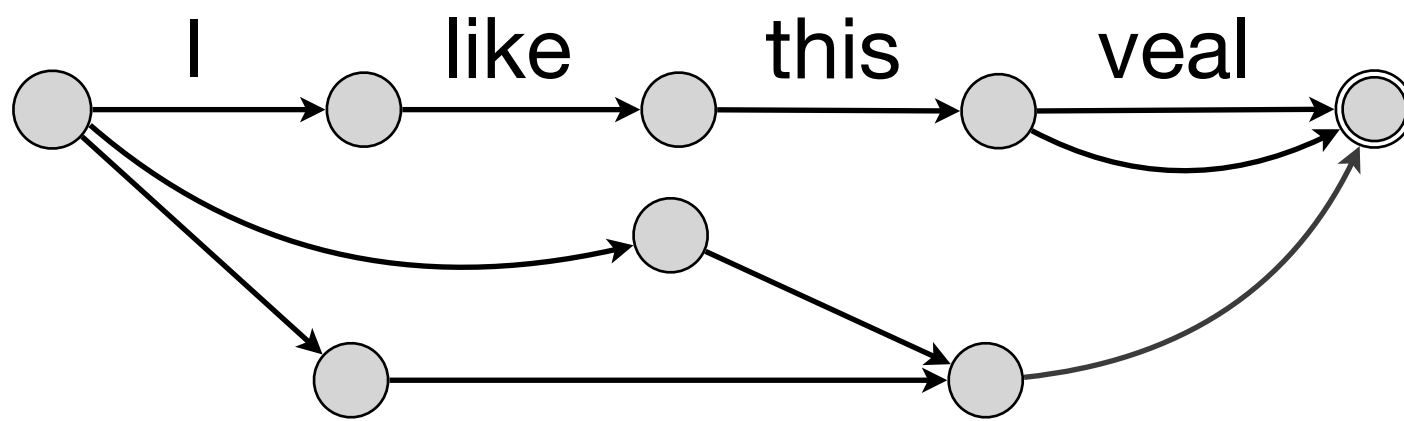
speech signal



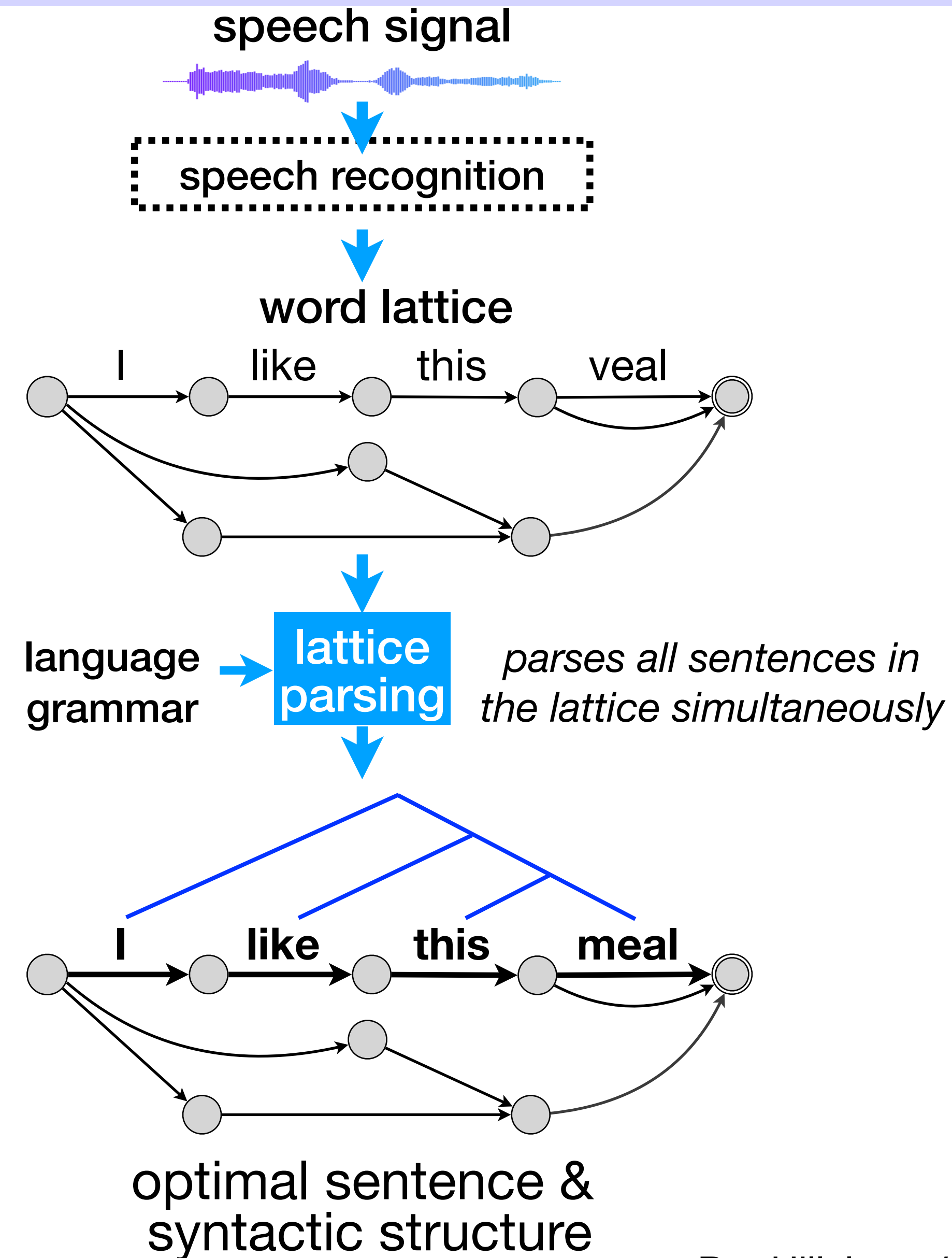
speech recognition



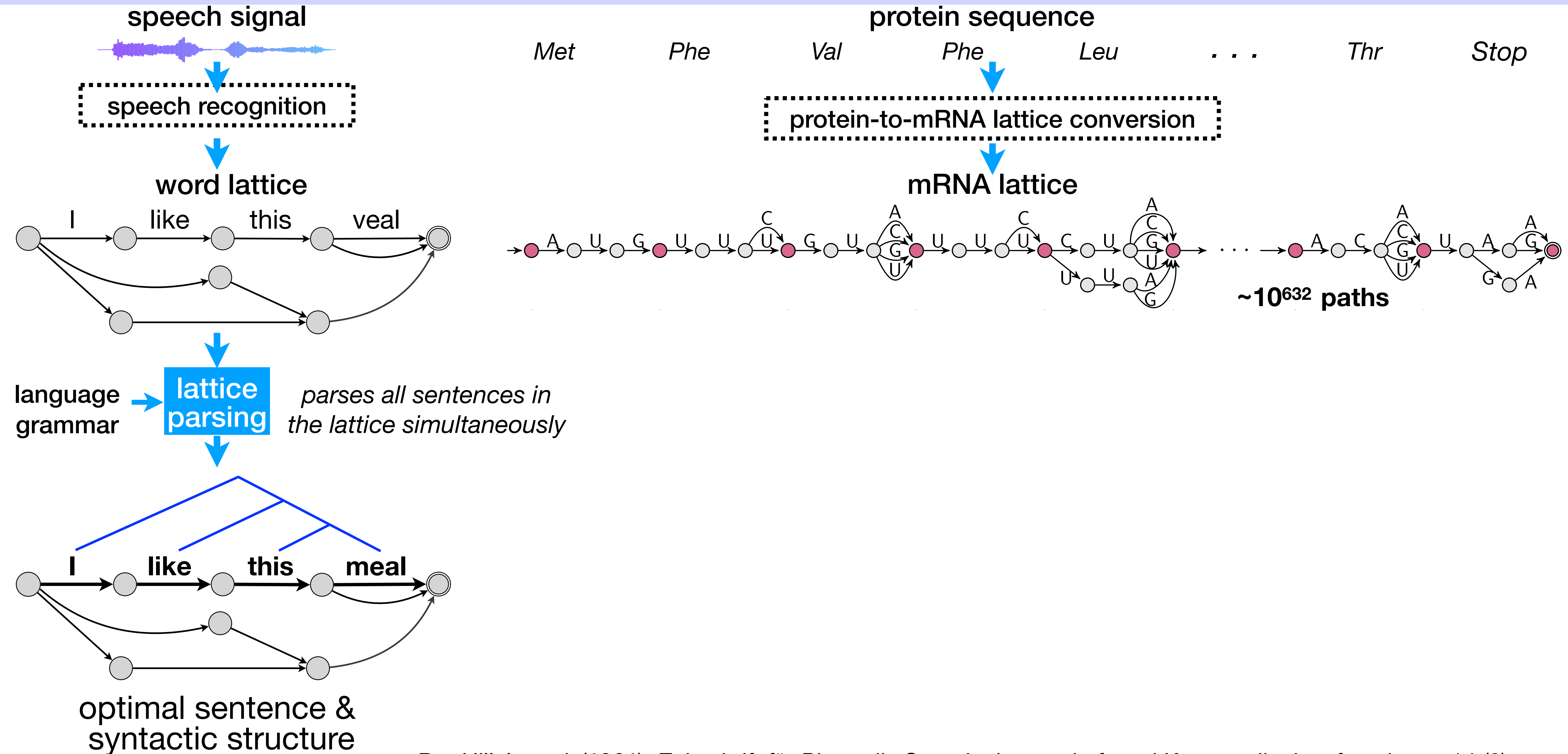
word lattice



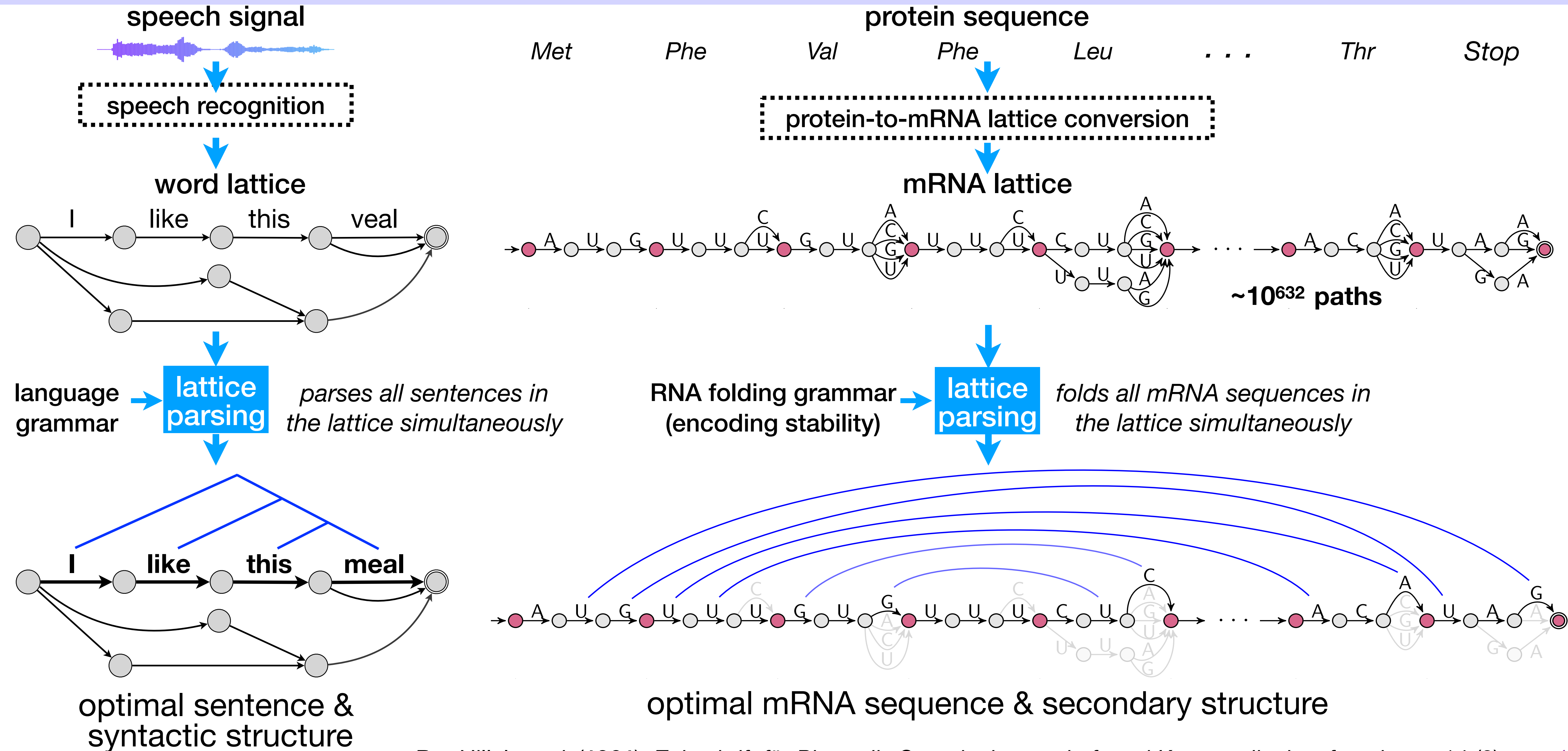
Linear Design Algorithm: Lattice Parsing



Linear Design Algorithm: Lattice Parsing



Linear Design Algorithm: Lattice Parsing

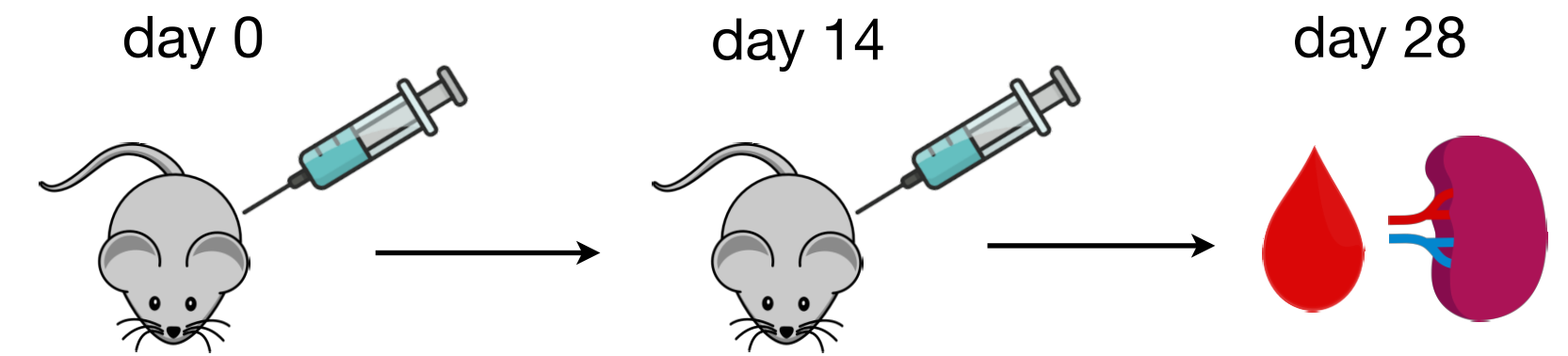


LinearDesign Results on SARS-CoV-2 Spike Protein

largely single-stranded
high free energy
easily degraded

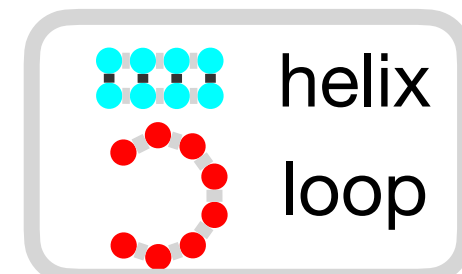
mostly double-stranded
lowest free energy
most stable

~11 minutes of
AI optimization



COVID mRNA vaccine:
up to 128x antibody response
compare to codon optimization
(e.g., BioNTech, Moderna, CureVac)

Wildtype
(63.4% paired)



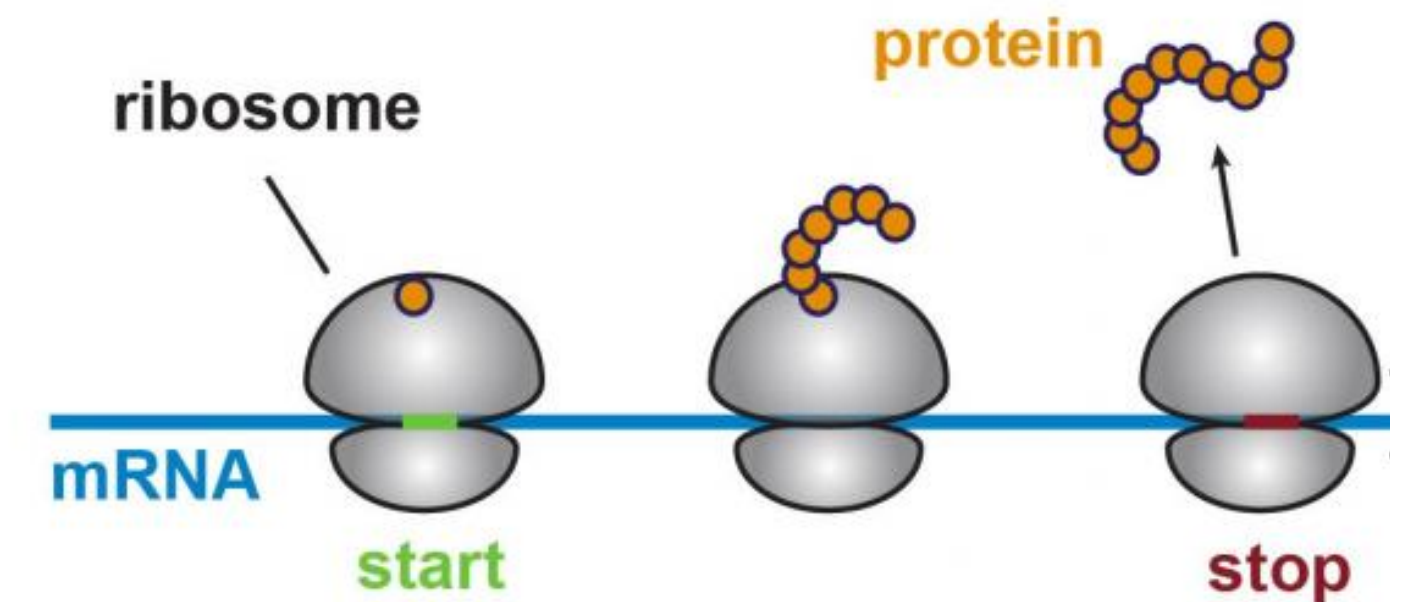
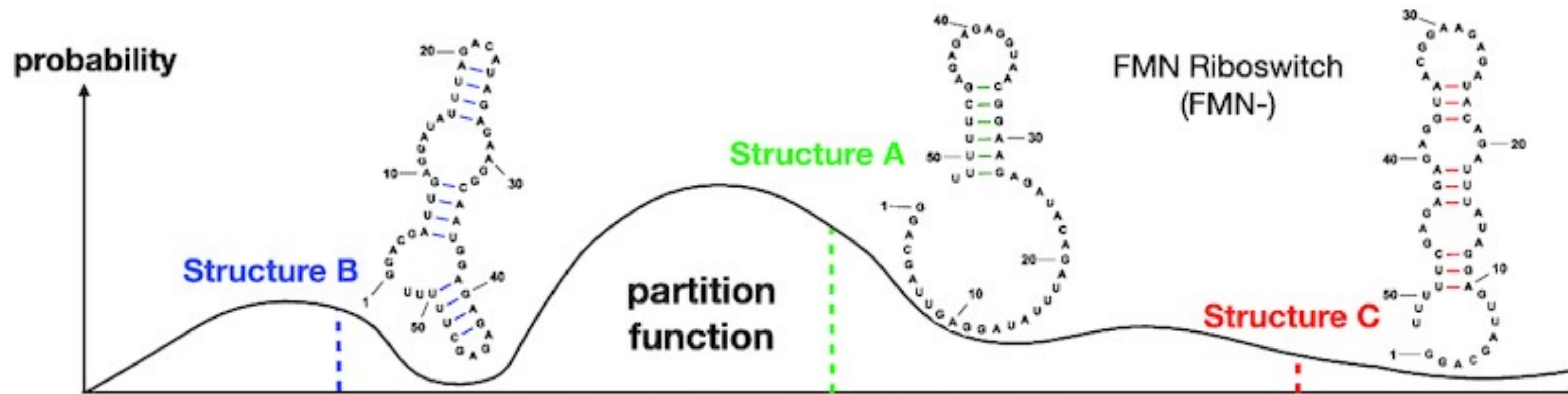
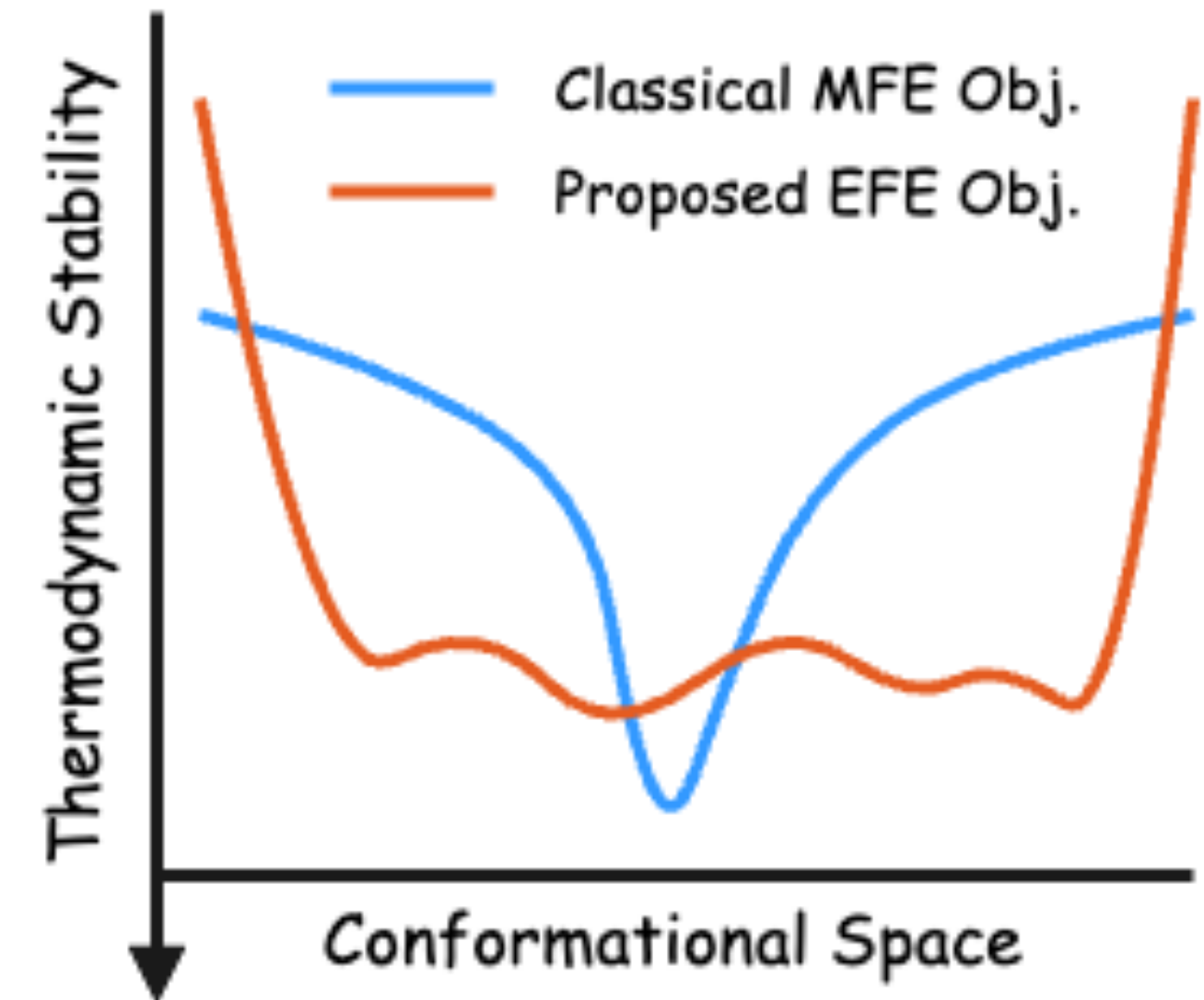
optimally stable mRNA
(83.6% paired)



From LinearDesign to EnsembleDesign

From Single MFE Structure to Whole Ensemble

- an RNA folds into millions of alternative structures in equilibrium (much more flexible than protein)
- an mRNA needs to constantly unfold and refold as ribosome goes through it in translation
- an ideal mRNA should be **stable and flexible**: with many stable structures but flatter energy landscape
- optimize “**ensemble stability**” instead of “**MFE stability**”



New Objective: Ensemble Free Energy (EFE)

- ensemble stability of mRNA across all conformations

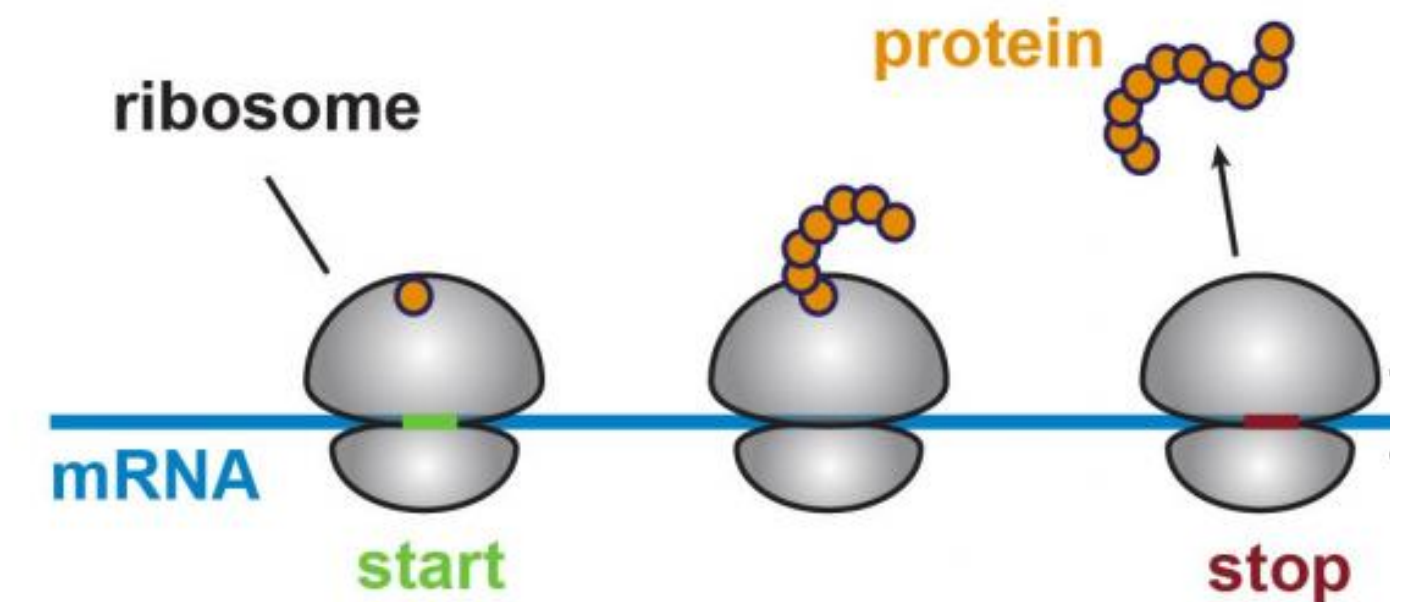
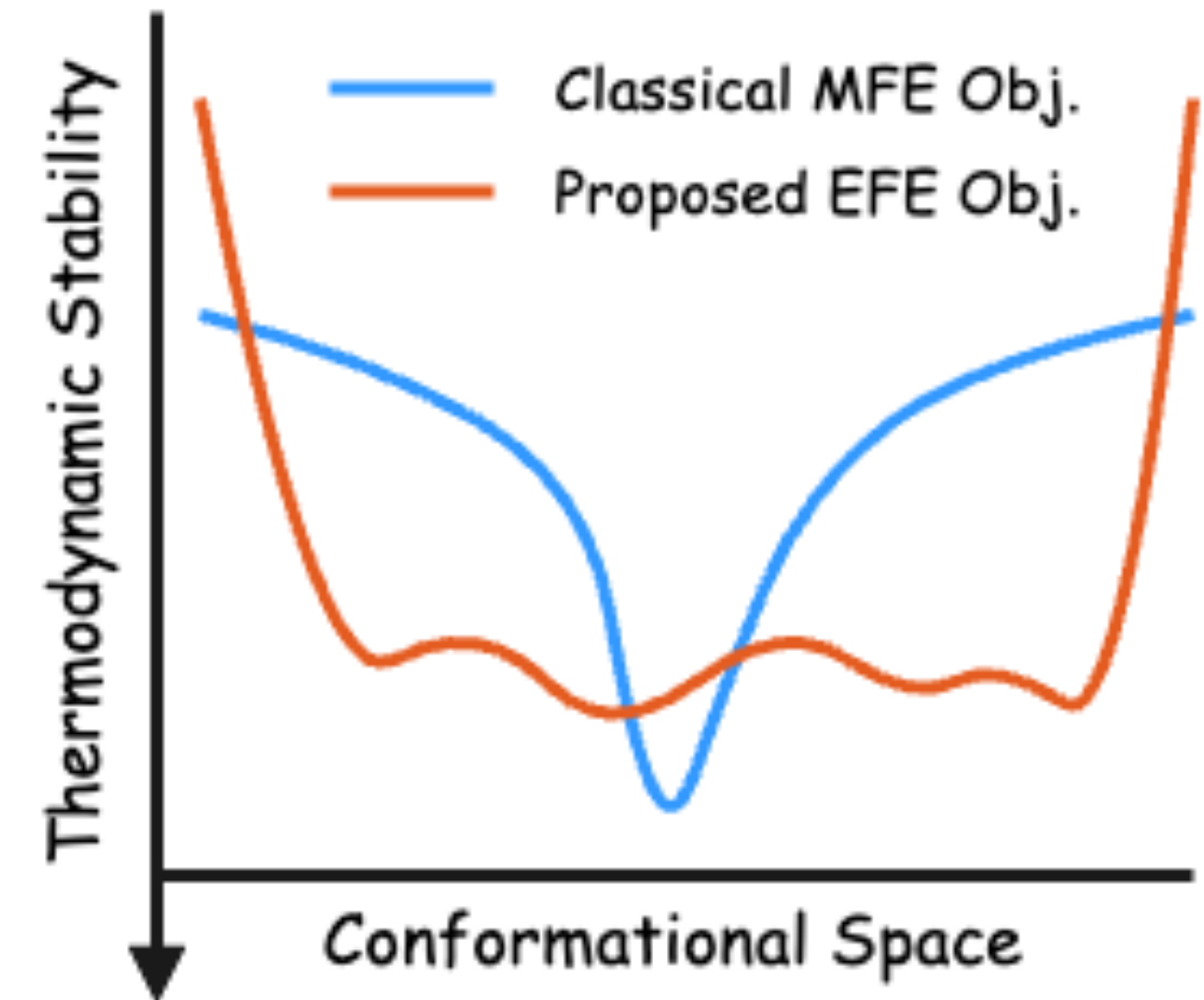
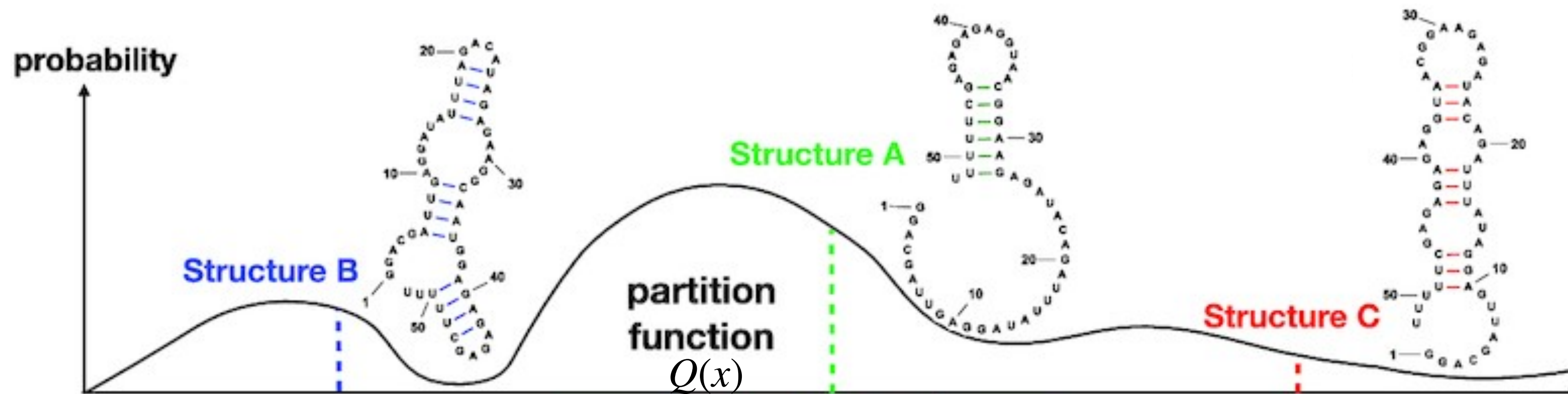
- Classical MFE Objective (most stable structure)

$$\min_{x \in \mathcal{X}(p)} \text{MFE}(x) = \min_{x \in \mathcal{X}(p)} \min_{y \in \mathcal{Y}(x)} \Delta G^\circ(x, y)$$

- Proposed EFE Objective (sum over all structures)

$$\min_{x \in \mathcal{X}(p)} \Delta G_{\text{ens}}^\circ(x) = \min_{x \in \mathcal{X}(p)} -RT \log \sum_{y \in \mathcal{Y}(x)} e^{\frac{-\Delta G^\circ(x, y)}{RT}}$$

partition function $Q(x)$



New Objective: Ensemble Free Energy (EFE)

- ensemble stability of mRNA across all conformations

- Classical MFE Objective (most stable structure)

$$\min_{x \in \mathcal{X}(p)} \text{MFE}(x) = \min_{x \in \mathcal{X}(p)} \min_{y \in \mathcal{Y}(x)} \Delta G^\circ(x, y)$$

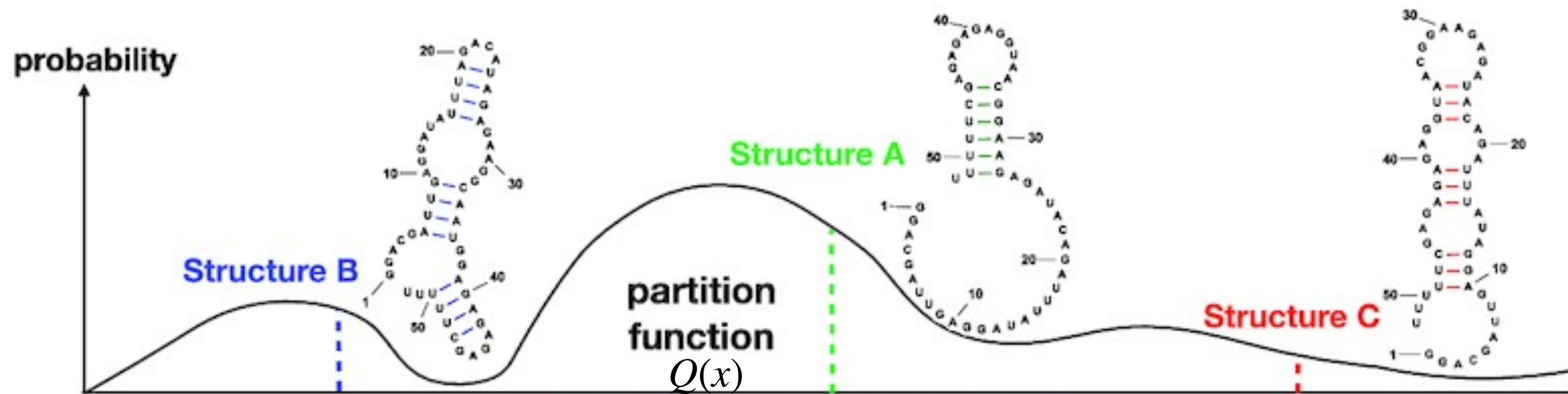
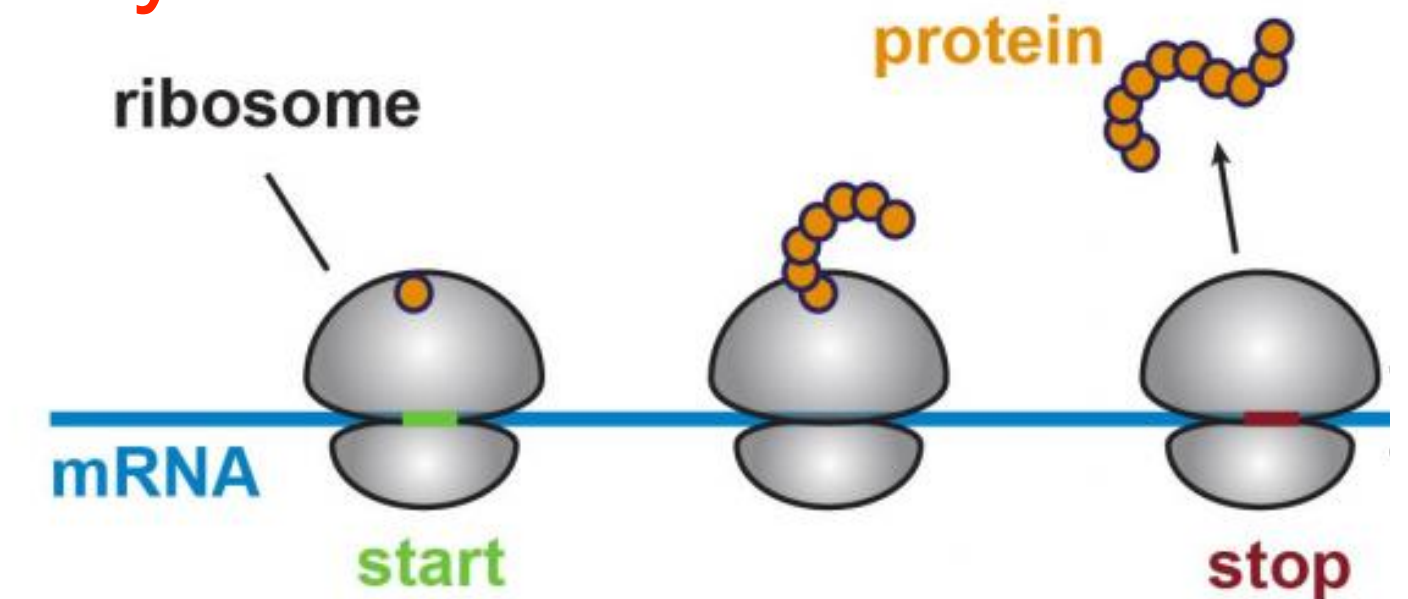
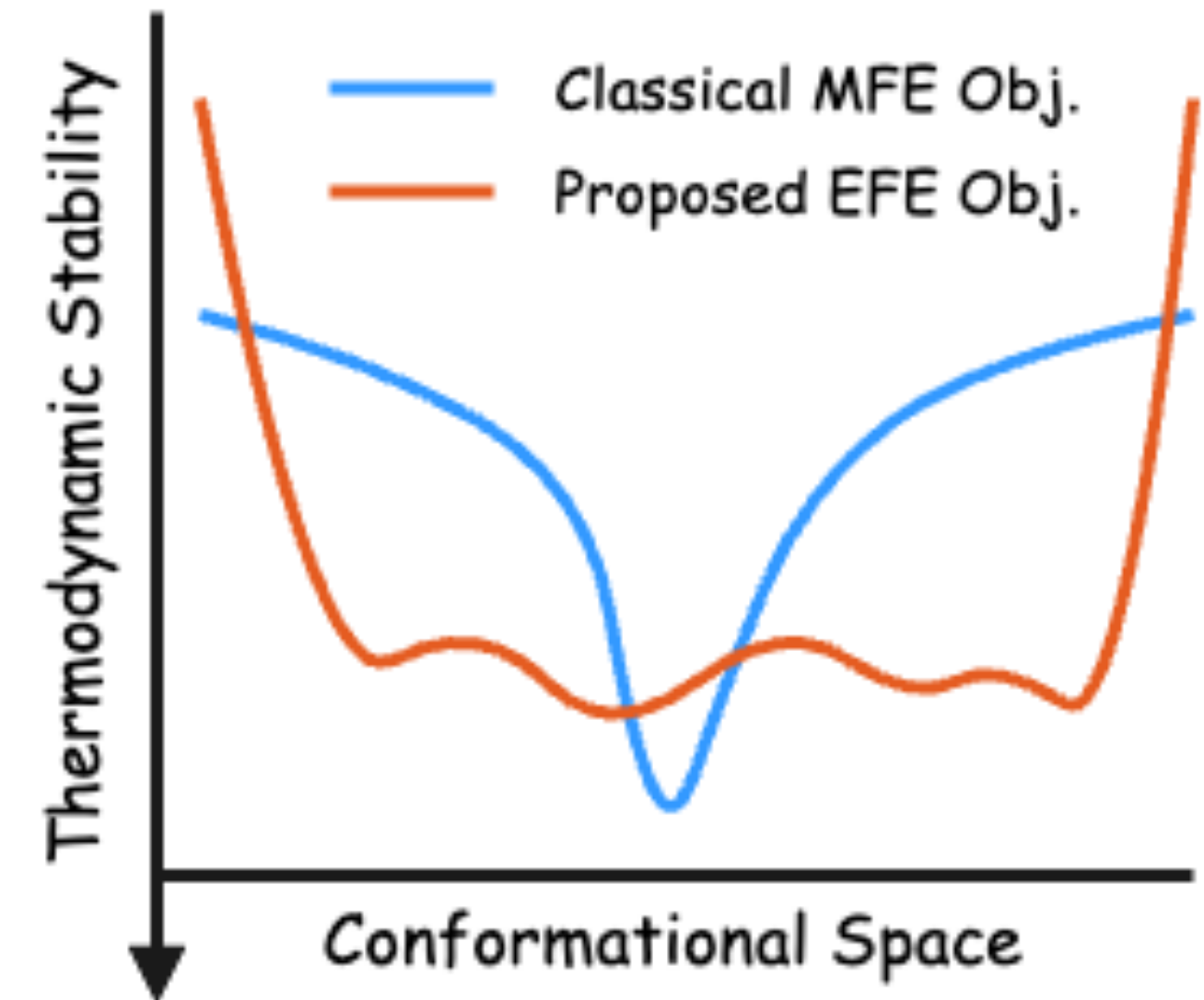
min over min
dynamic program

- Proposed EFE Objective (sum over all structures)

$$\min_{x \in \mathcal{X}(p)} \Delta G_{\text{ens}}^\circ(x) = \min_{x \in \mathcal{X}(p)} -RT \log \sum_{y \in \mathcal{Y}(x)} e^{\frac{-\Delta G^\circ(x, y)}{RT}}$$

min over log sum
beyond DP, likely NP-hard

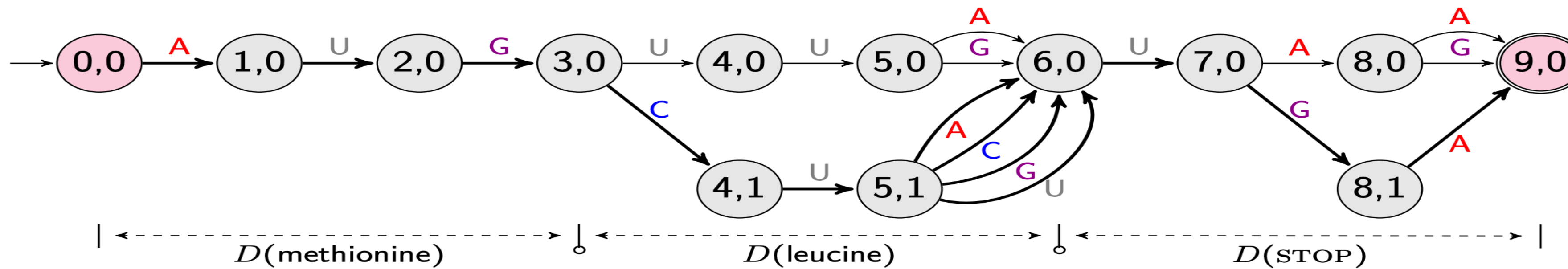
partition function $Q(x)$



Discrete Optimization => Continuous Optimization

- LinearDesign: $\min_{x \in \mathcal{X}(p)} \text{MFE}(x) = \min_{x \in \mathcal{X}(p)} \min_{y \in \mathcal{Y}(x)} \Delta G^\circ(x, y)$

min over min
dynamic programming
via lattice parsing



- EnsembleDesign: $\min_{x \in \mathcal{X}(p)} \Delta G_{\text{ens}}^\circ(x) = \min_{x \in \mathcal{X}(p)} -RT \log \sum_{y \in \mathcal{Y}(x)} e^{\frac{-\Delta G^\circ(x, y)}{RT}}$ ← partition function $Q(x)$

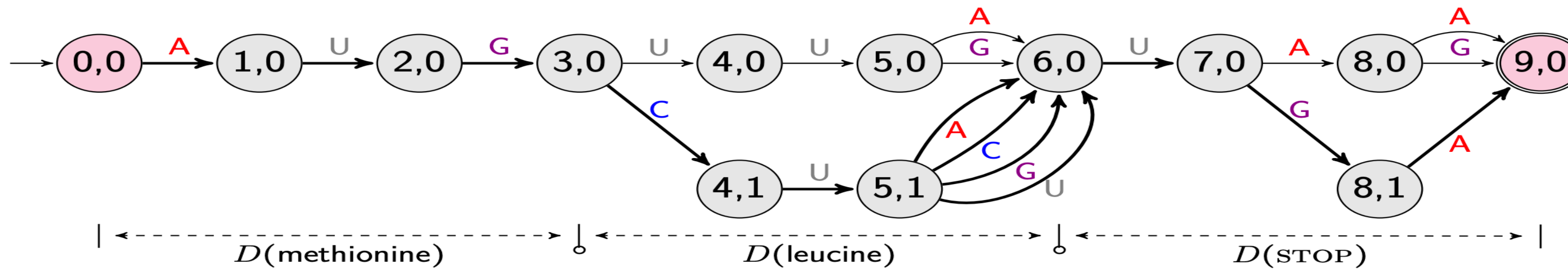
min over log sum
likely NP-hard

- instead use a probabilistic distribution over all designs

Discrete Optimization => Continuous Optimization

- LinearDesign: $\min_{x \in \mathcal{X}(p)} \text{MFE}(x) = \min_{x \in \mathcal{X}(p)} \min_{y \in \mathcal{Y}(x)} \Delta G^\circ(x, y)$

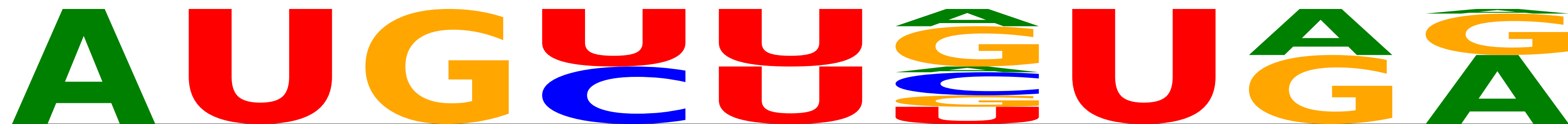
min over min
dynamic programming
via lattice parsing



- EnsembleDesign: $\min_{x \in \mathcal{X}(p)} \Delta G_{\text{ens}}^\circ(x) = \min_{x \in \mathcal{X}(p)} -RT \log \sum_{y \in \mathcal{Y}(x)} e^{\frac{-\Delta G^\circ(x, y)}{RT}}$ ← partition function $Q(x)$

min over log sum
likely NP-hard

- instead use a probabilistic distribution over all designs



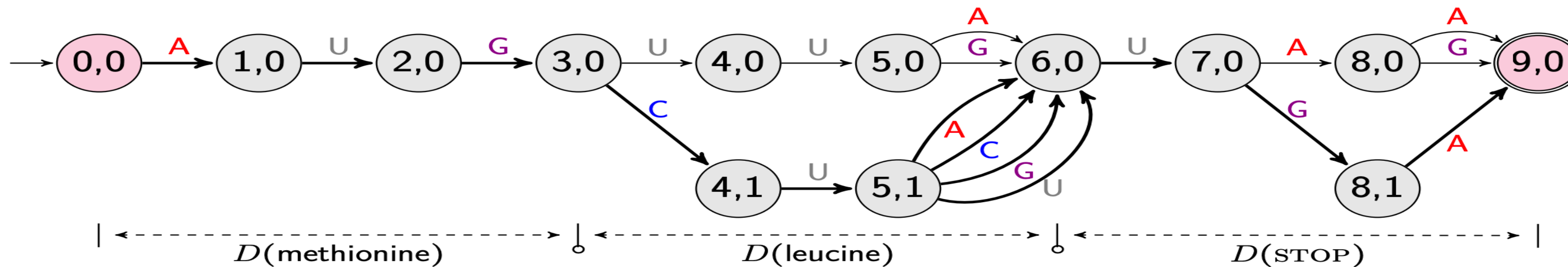
$$\mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^\circ(x)] = -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)]$$

optimize *expected* EFE
over whole distribution

Discrete Optimization => Continuous Optimization

- LinearDesign: $\min_{x \in \mathcal{X}(p)} \text{MFE}(x) = \min_{x \in \mathcal{X}(p)} \min_{y \in \mathcal{Y}(x)} \Delta G^\circ(x, y)$

min over min
dynamic programming
via lattice parsing



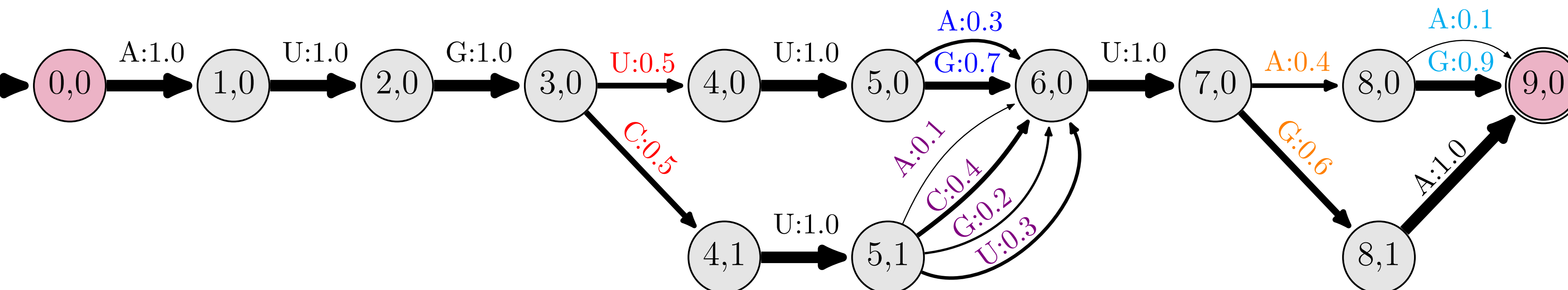
- EnsembleDesign: $\min_{x \in \mathcal{X}(p)} \Delta G_{\text{ens}}^\circ(x) = \min_{x \in \mathcal{X}(p)} -RT \log \sum_{y \in \mathcal{Y}(x)} e^{\frac{-\Delta G^\circ(x, y)}{RT}}$ ← partition function $Q(x)$

min over log sum
likely NP-hard

- instead use a probabilistic distribution over all designs



$$\mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^\circ(x)] = -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)]$$



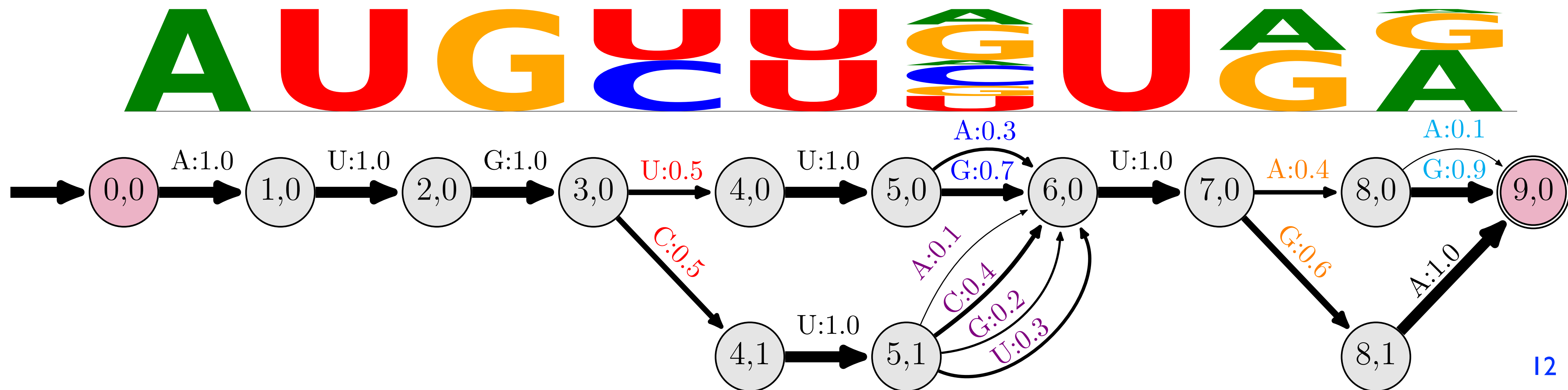
optimize *expected* EFE
over whole distribution

Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

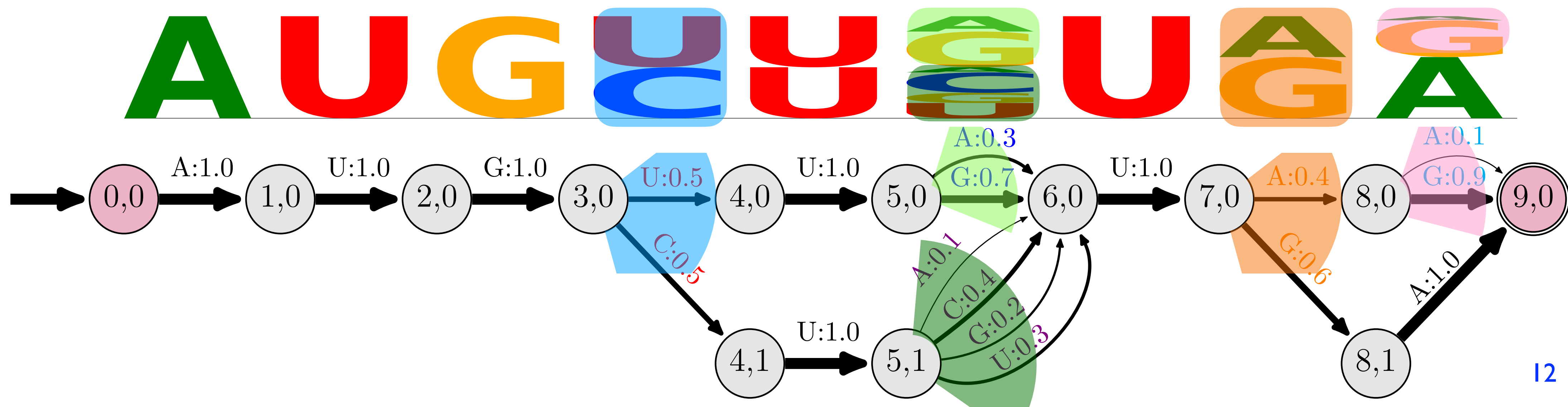


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

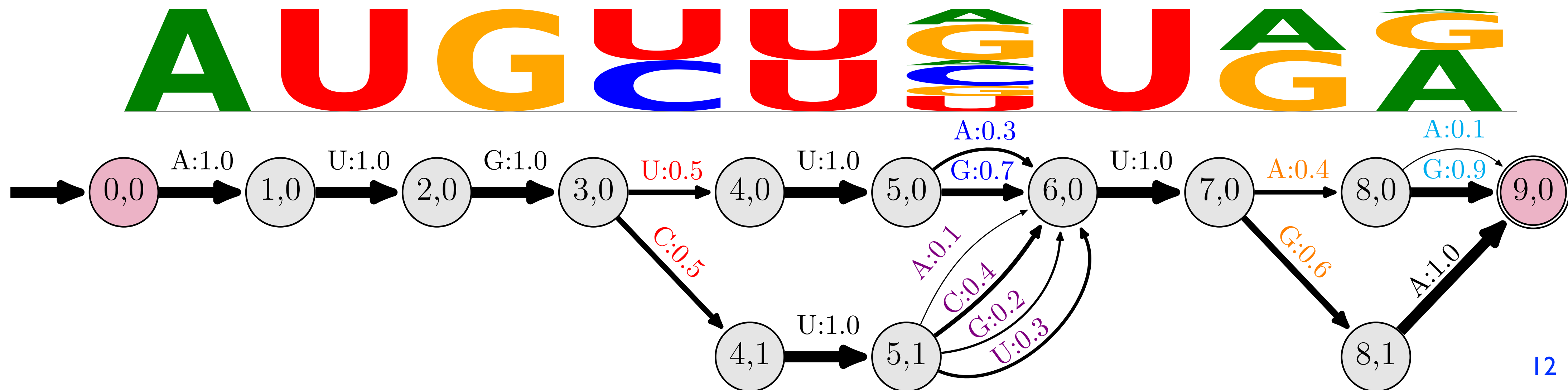


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

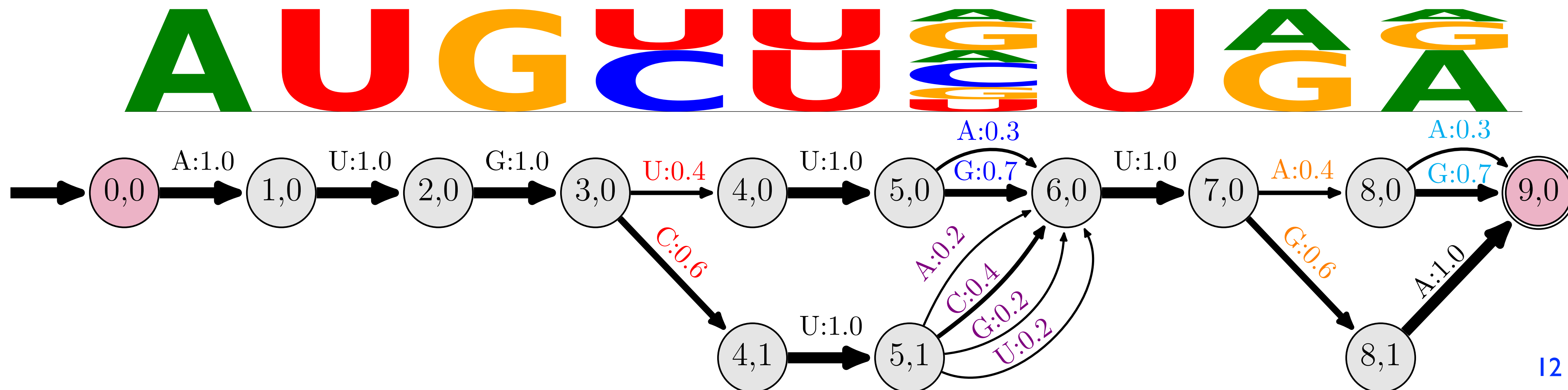


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

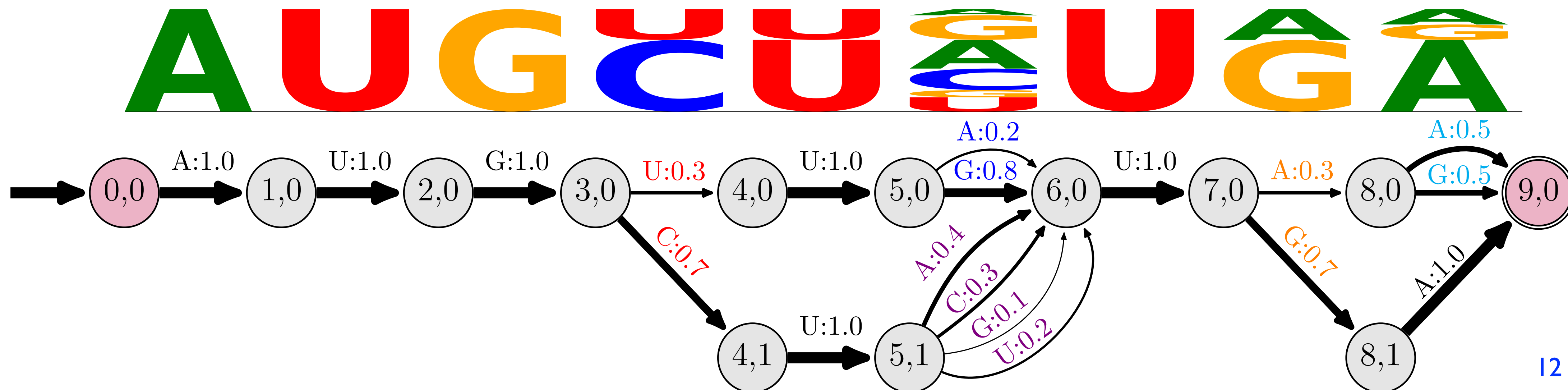


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

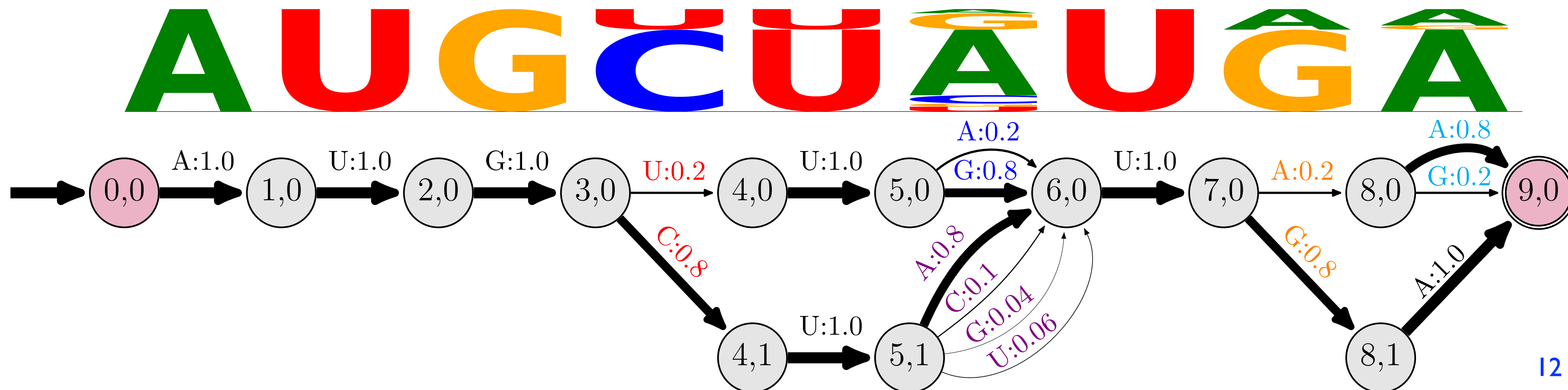


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E}$ [EFE]) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot

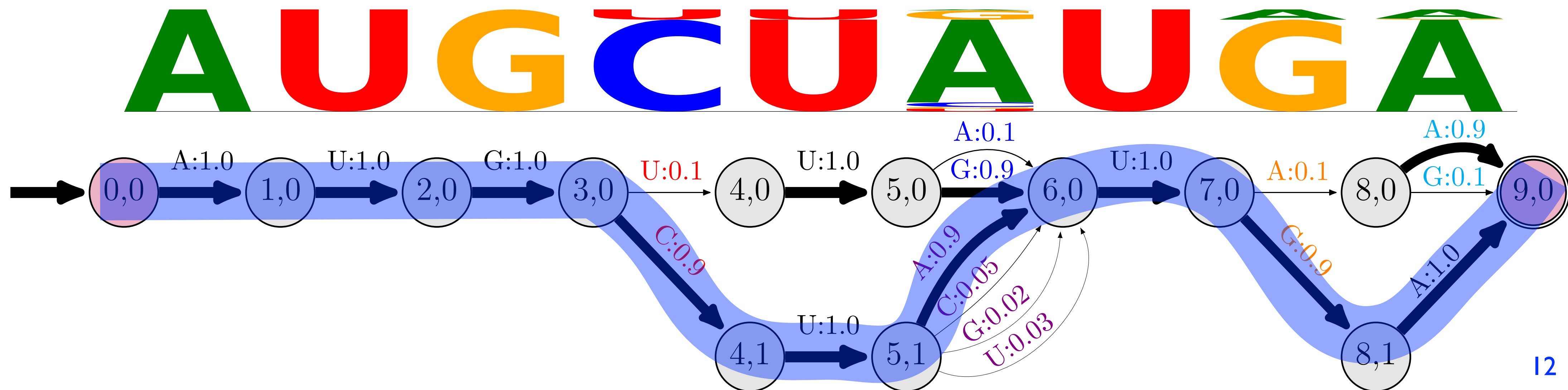


Distribution on Lattice: Probabilistic Lattice

- each (branching) node in the lattice has a local distribution
- sequence distribution $\mathbb{D}$ via product and weighted sum of local distributions
- optimize **expected** ensemble free energy ($\mathbb{E} [\text{EFE}]$) by gradient descent

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\Delta G_{\text{ens}}^{\circ}(\mathbf{x})] = -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})]$$

- distribution gradually shrinks in entropy and converges to one-hot



EnsembleDesign via Probabilistic Lattice Parsing

$$\mathbb{E} [\text{EFE}] : \quad -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)] \approx -RT \log \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [Q(x)]$$


EnsembleDesign via Probabilistic Lattice Parsing

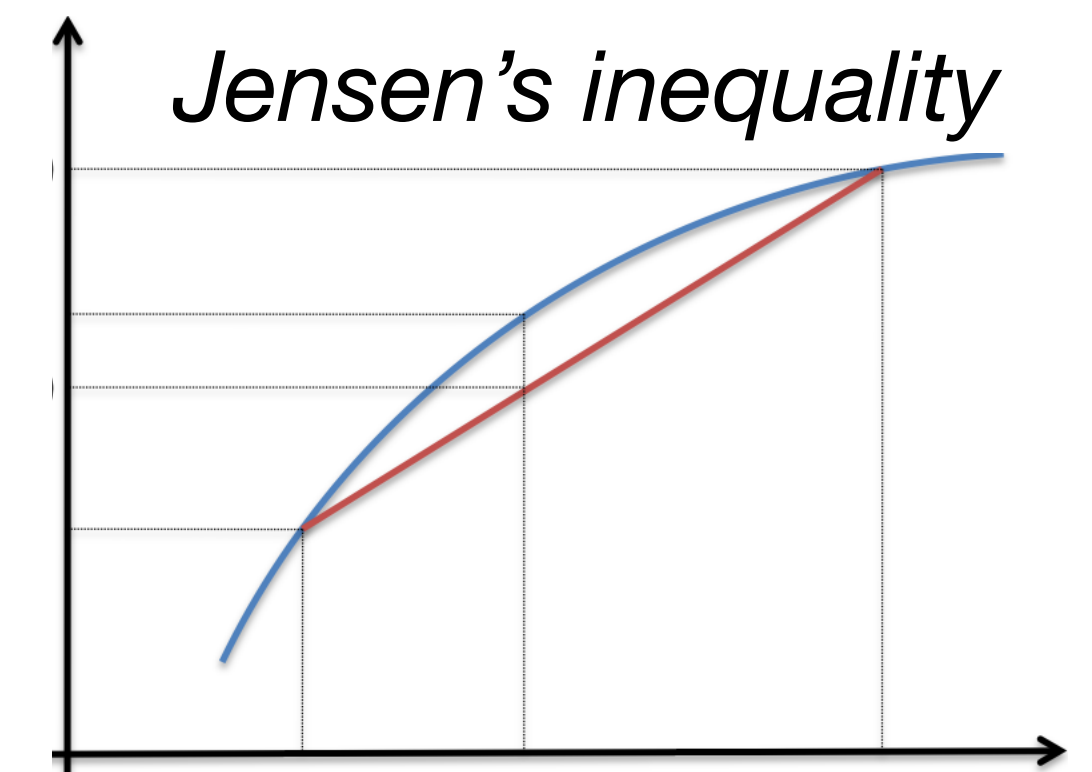

$$\mathbb{E} [\text{EFE}] : \quad -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)] \approx -RT \log \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [Q(x)]$$

- the continuous version is still hard to minimize, due to $\mathbb{E}$ over $\log$

EnsembleDesign via Probabilistic Lattice Parsing

$$\mathbb{E} [\text{EFE}] : \quad -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)] \approx -RT \log \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [Q(x)]$$

- the continuous version is still hard to minimize, due to $\mathbb{E}$ over $\log$
- further apply Jensen's inequality to swap the order b/w $\mathbb{E}$ and $\log$

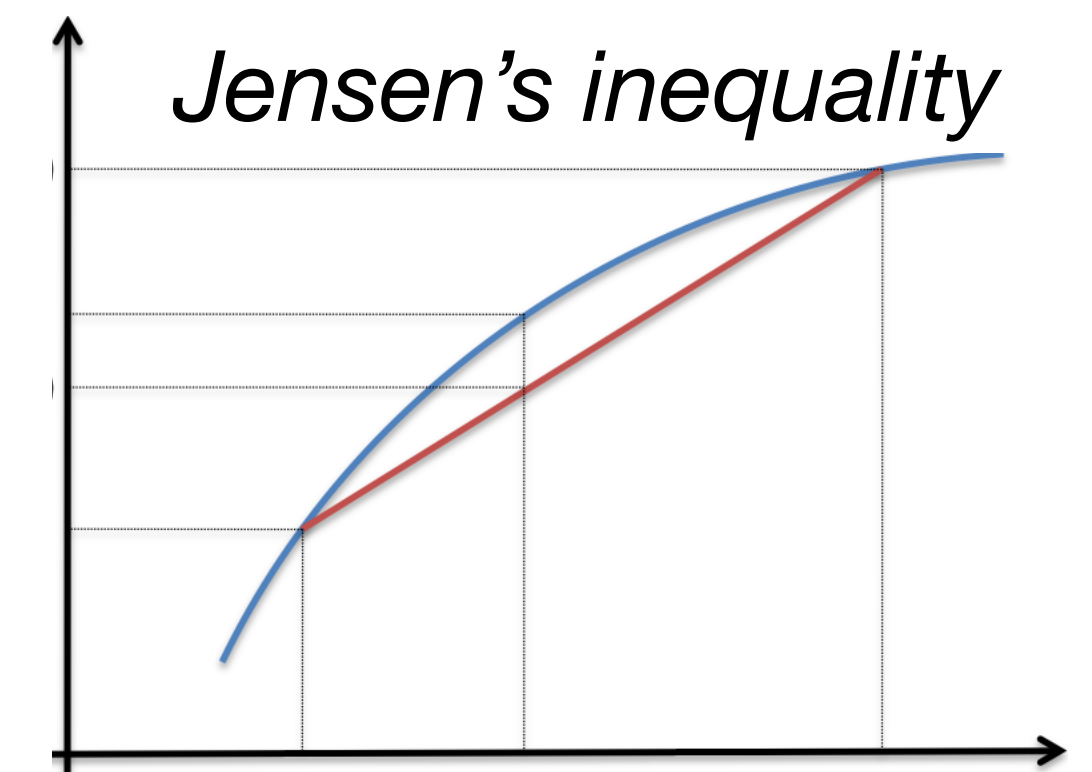


EnsembleDesign via Probabilistic Lattice Parsing

$$\mathbb{E} [\text{EFE}] : \quad -RT \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [\log Q(\mathbf{x})] \approx -RT \log \mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [Q(\mathbf{x})]$$

- the continuous version is still hard to minimize, due to $\mathbb{E}$ over $\log$
- further apply Jensen's inequality to swap the order b/w $\mathbb{E}$ and $\log$
- minimizing this surrogate $\Leftrightarrow$ maximizing **expected partition function**:

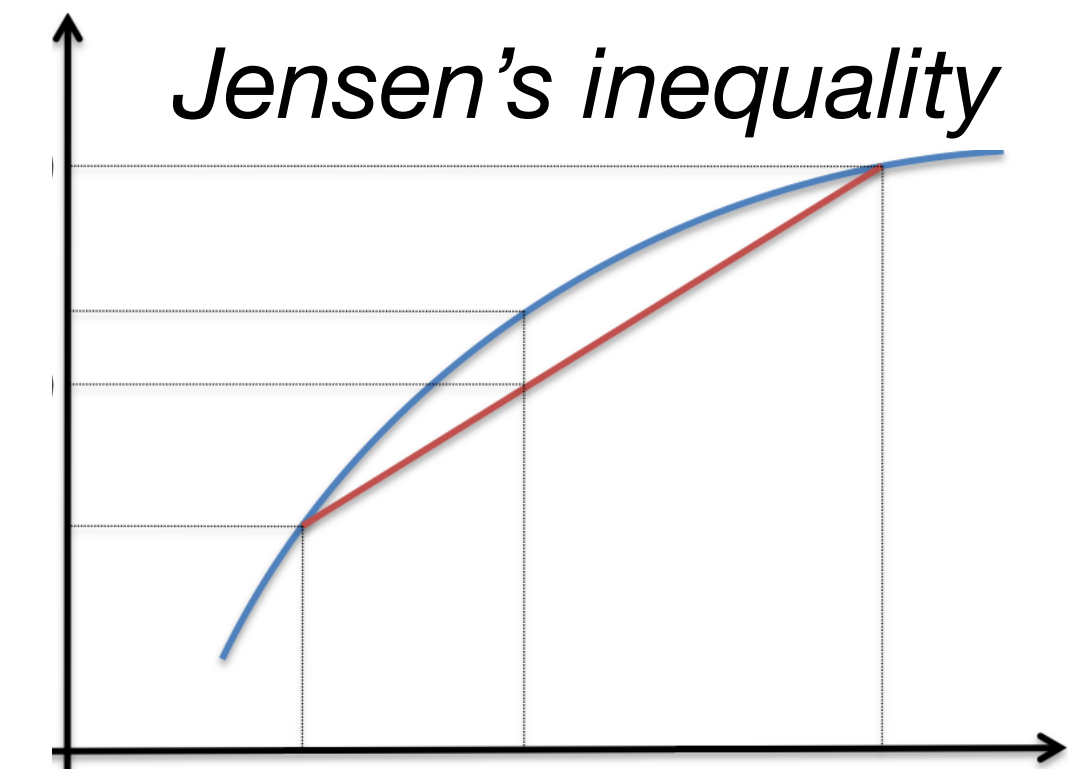
$$\mathbb{E}_{\mathbf{x} \sim \mathbb{D}(\cdot)} [Q(\mathbf{x})] = \sum_{\mathbf{x} \in \mathcal{X}(p)} \mathbb{D}(\mathbf{x}) \left(\sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} e^{-\Delta G^\circ(\mathbf{x}, \mathbf{y})/RT} \right) \quad \text{sum over sum}$$



EnsembleDesign via Probabilistic Lattice Parsing

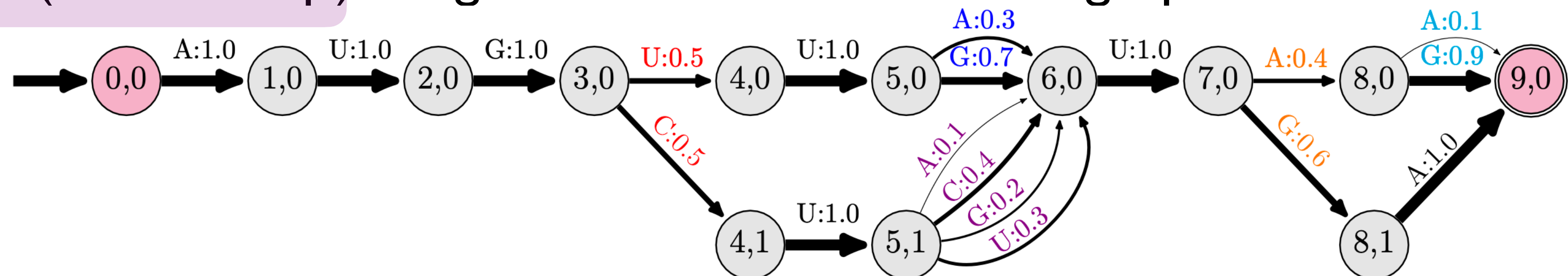
$$\mathbb{E} [\text{EFE}] : -RT \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [\log Q(x)] \approx -RT \log \mathbb{E}_{x \sim \mathbb{D}(\cdot)} [Q(x)]$$

- the continuous version is still hard to minimize, due to $\mathbb{E}$ over $\log$
- further apply Jensen's inequality to swap the order b/w $\mathbb{E}$ and $\log$
- minimizing this surrogate $\Leftrightarrow$ maximizing **expected partition function**:



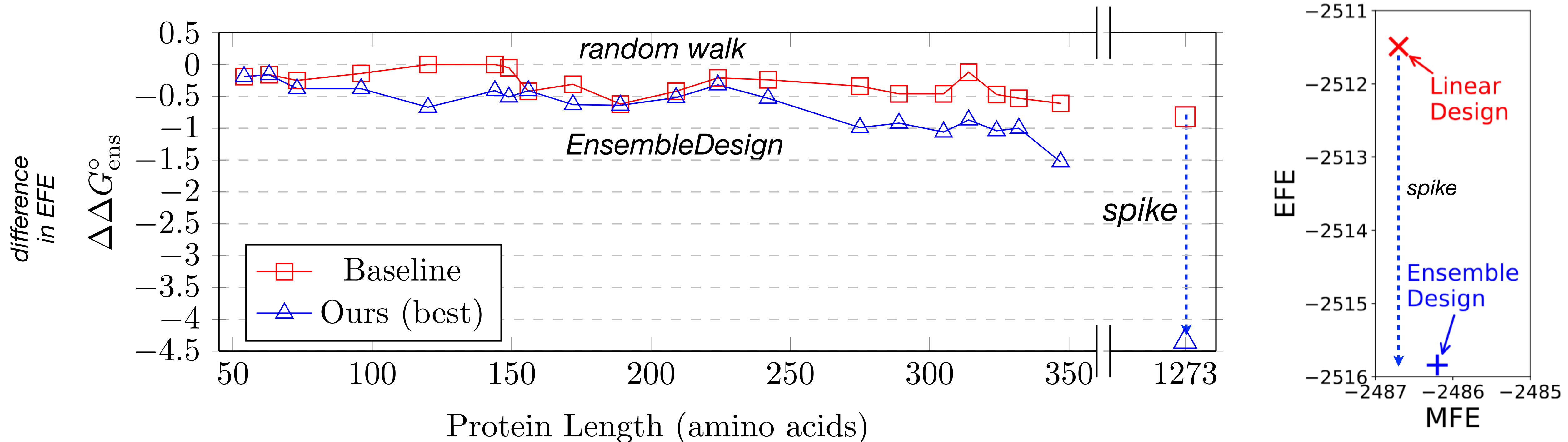
$$\mathbb{E}_{x \sim \mathbb{D}(\cdot)} [Q(x)] = \sum_{x \in \mathcal{X}(p)} \mathbb{D}(x) \left(\sum_{y \in \mathcal{Y}(x)} e^{-\Delta G^\circ(x,y)/RT} \right) \quad \text{sum over sum}$$

- (weighted) sum over sum can be efficiently solved by extending LinearDesign in two steps
- inner sum (over all structures)**: partition function (not MFE) version of LinearDesign
- outer weighted sum (over all seqs)**: weighted version of LinearDesign: probabilistic lattice parsing



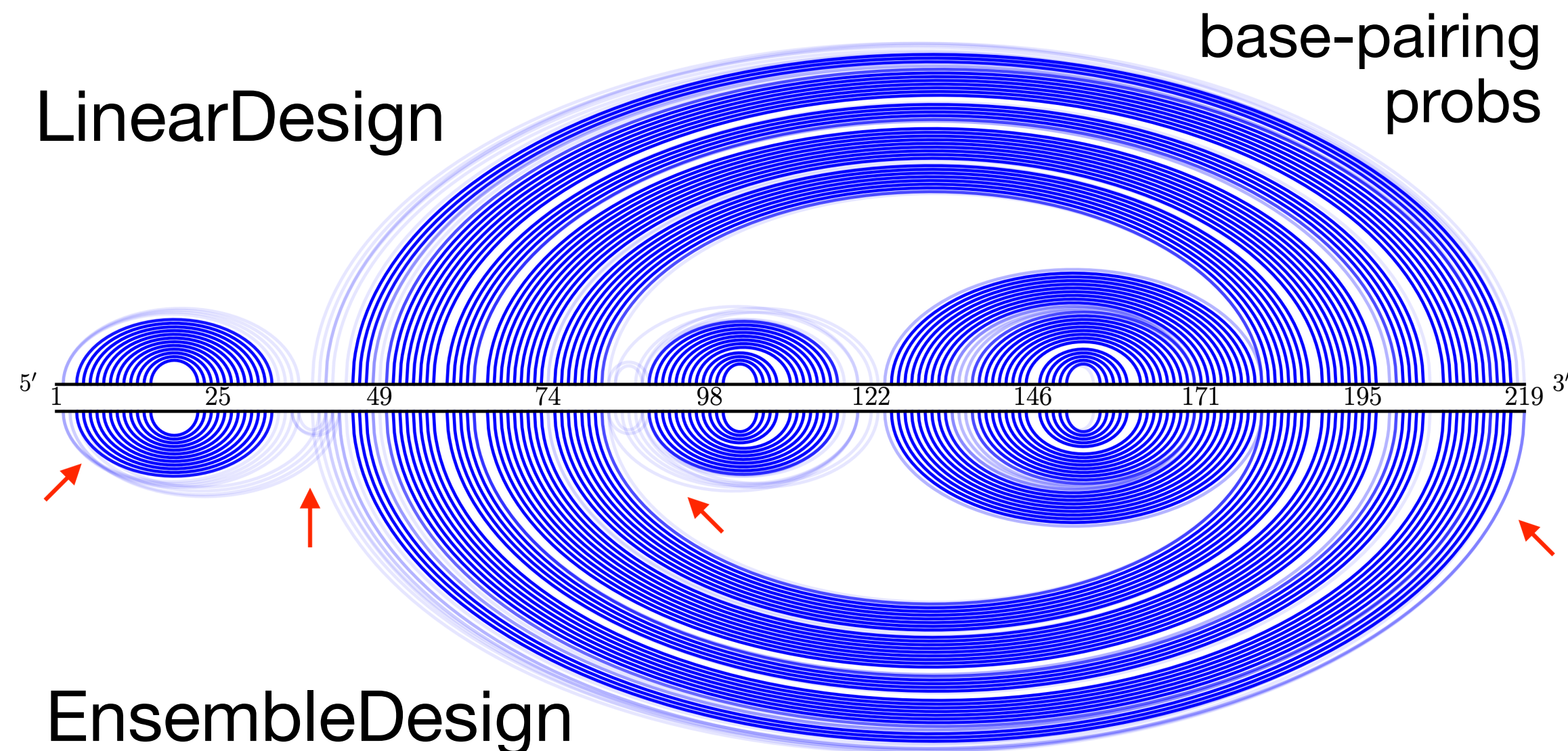
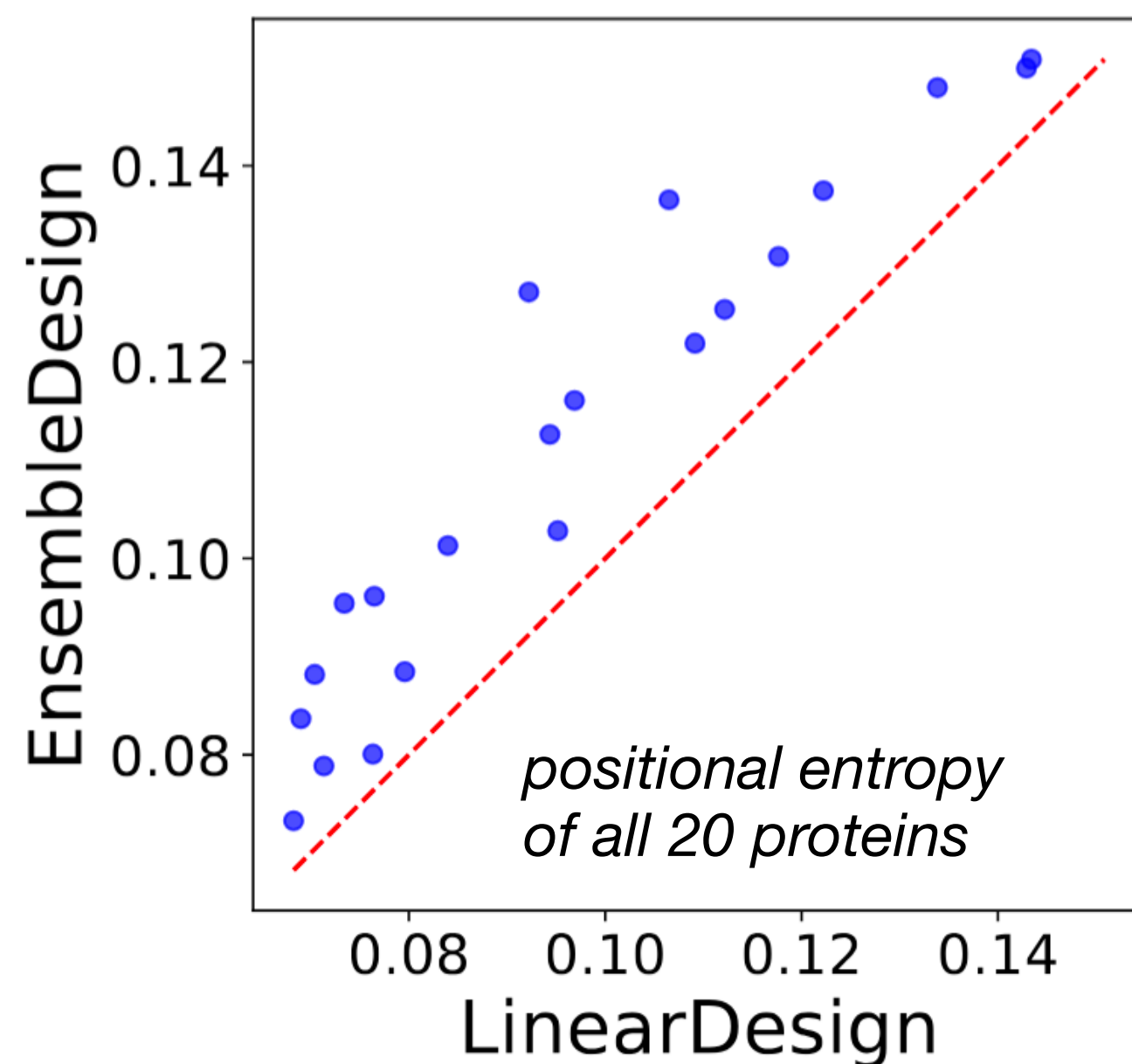
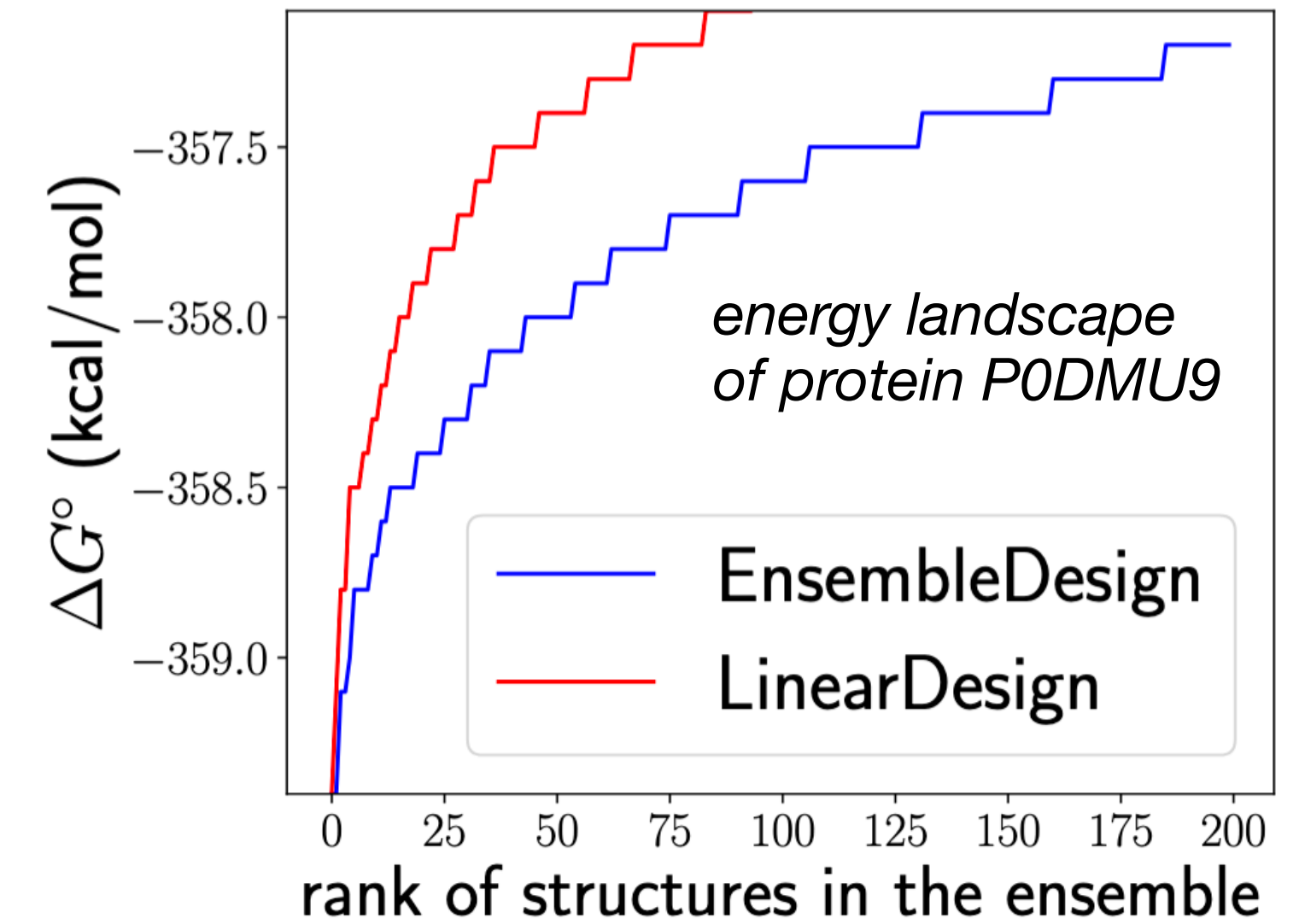
Main Evaluation Results

- EnsembleDesign consistently finds sequences with lower (better) ensemble free energy
 - bigger advantage on longer proteins over random walk baseline
 - substantially better ensemble stability with negligible sacrifice on MFE stability
 - data: 20 UniProt proteins (50~350 aa) and SARS-CoV-2 spike protein (1273 aa)



Byproducts: Flatter Distribution and Lower AUP

- optimizing ensemble stability also results in
 - flatter energy landscape
 - higher positional entropy (more structural flexibility)
 - lower average unpaired probability (“AUP”)
 - AUP is related to RNA degradation (...)



	AUP
LinearDesign	0.1973
Random Walk	0.1966
EnsembleDesign	0.1949

Thank you!



Proceedings Paper



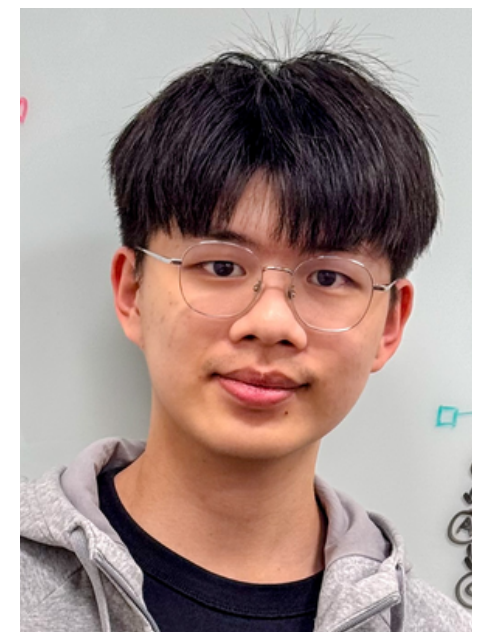
GitHub Code



Ning Dai



Tianshuo Zhou



Wei Yu Tang
(now PhD student
at USC QCB)



Oregon State
University

Poster C-247
SamplingDesign
for RNA Design