

mRNA Vaccine

Sequence & Structure Design and Optimization

Advances and Challenges (Jin et al., 2025)

◇ ◇ ◇

Application détaillée

LinearDesign — algorithme d'optimisation
de séquences ARNm pour les vaccins SARS-CoV-2 et VZV

Angenot Valentin

Année académique 2025–2026

Table des matières

I	Introduction	3
I.1	Contexte général : la vaccination, un enjeu majeur de santé publique	3
I.2	Les vaccins à ARNm : avantages	3
I.3	Procédure de production d'un ARNm vaccinal et principe d'un vaccin à ARNm	4
I.4	Défis actuels des vaccins à ARNm	5
I.5	Principes de l'optimisation computationnelle des séquences ARNm	5
I.6	Principales réalisations récentes et choix de l'application	6
II	Application : LinearDesign — optimisation algorithmique des séquences d'ARNm vaccinaux	8
II.1	Contexte et problématique	8
II.2	L'élégance de l'analogie linguistique : du mot au codon	8
II.3	Résultats expérimentaux : vaccins SARS-CoV-2 et VZV	11
II.3.1	Vaccin contre le SARS-CoV-2 (protéine spike)	11
II.3.2	Vaccin contre le virus varicelle-zona (VZV, protéine gE)	13
II.4	Comparaison avec les approches concurrentes	15
III	Discussion et conclusion	16
III.1	Rappel de la problématique	16
III.2	Solutions apportées par LinearDesign	16
III.3	Points forts et limites	17
III.4	Perspectives futures et actualité	17
III.5	Conclusion	18
	Références bibliographiques	20

I Introduction

I.1 Contexte général : la vaccination, un enjeu majeur de santé publique

La vaccination constitue l'une des interventions de santé publique les plus significatives dans la réduction de la mortalité mondiale depuis 1974. Selon l'Organisation mondiale de la Santé, la vaccination à l'échelle globale a permis de prévenir environ **154 millions de décès** au cours des cinquante dernières années (Shattock *et al.*, 2024). Le principe fondamental de la vaccination repose sur l'entraînement du système immunitaire à reconnaître les antigènes produits par des agents pathogènes ou des cellules malignes, afin de stimuler une réponse immunitaire adaptative robuste contre les infections ou les cancers (Pardi *et al.*, 2018 ; Chaudhary *et al.*, 2021).

Les vaccins conventionnels, qu'ils soient inactivés, atténués ou à sous-unités recombinantes, ont permis de lutter efficacement contre de nombreuses maladies infectieuses. Néanmoins, leur développement reste souvent long et coûteux, nécessitant la manipulation de pathogènes infectieux et impliquant des risques de contamination (Pardi *et al.*, 2018). C'est dans ce contexte que la technologie des vaccins à **ARN messenger (ARNm)** a émergé comme une alternative révolutionnaire, offrant des avantages considérables en termes de rapidité de développement, de sécurité et de flexibilité (Pardi *et al.*, 2018 ; Chaudhary *et al.*, 2021).

I.2 Les vaccins à ARNm : avantages

Les avancées récentes en technologie de l'ARN messenger ont permis de délivrer des molécules d'ARNm codant pour des antigènes directement dans les cellules hôtes, facilitant ainsi l'expression des antigènes au sein de l'organisme humain (Pardi *et al.*, 2018). Cette approche présente plusieurs avantages distincts par rapport aux vaccins conventionnels.

Rapidité de développement : L'ARNm peut être conçu et fabriqué de manière rapide, évolutive et rentable grâce aux rendements élevés de la technique de **transcription *in vitro* (IVT)**. Le vaccin mRNA-1273 de Moderna est entré en phase de tests cliniques seulement **deux mois** après l'identification de la séquence génétique du SARS-CoV-2 (Chaudhary *et al.*, 2021).

Sécurité accrue : Ce type de vaccin ne contient aucun agent infectieux, éliminant ainsi tout risque d'infection lors de la vaccination. L'environnement de fabrication *in vitro* sans cellules minimise les problèmes de sécurité liés aux impuretés d'origine cellulaire (Pardi *et al.*, 2018).

Un aspect crucial du succès des vaccins ARNm réside dans leur système de délivrance. Les **nanoparticules lipidiques (LNP)**, composées de lipides ionisables, de lipides auxiliaires, de cholestérol et de lipides PEGylés, protègent l'ARNm de la dégradation enzymatique et facilitent son entrée dans les cellules cibles par endocytose (Hou *et al.*, 2021). Le développement des LNP a été déterminant pour la traduction clinique des vaccins ARNm, comme en témoignent les formulations utilisées dans les vaccins **BNT162b2** (Pfizer/BioNTech) et **mRNA-1273** (Moderna) (Hou *et al.*, 2021).

I.3 Procédure de production d'un ARNm vaccinal et principe d'un vaccin à ARNm

La production d'un ARNm vaccinal commence à partir de la séquence génétique de la protéine antigénique cible, obtenue par séquençage du génome du pathogène. Cette information permet de concevoir une séquence ADN optimisée, notamment par rapport à l'usage préférentiel des codons chez l'humain. L'ADN correspondant est ensuite synthétisé, généralement par automate, puis cloné sous forme de **pDNA** : un plasmide amplifiable en laboratoire dans des bactéries afin d'obtenir un grand nombre de copies. Ce pDNA sert finalement de matrice aux ARN polymérases pour la transcription *in vitro* d'ARNm artificiels. L'ARNm simplifié produit par IVT se compose de **cinq éléments fonctionnels essentiels**, chacun jouant un rôle spécifique dans l'expression de la protéine antigénique (Jin *et al.*, 2025 ; Kim *et al.*, 2024). La Figure 1 illustre le processus complet, de la conception de la matrice ADN jusqu'à la réponse immunitaire adaptative.

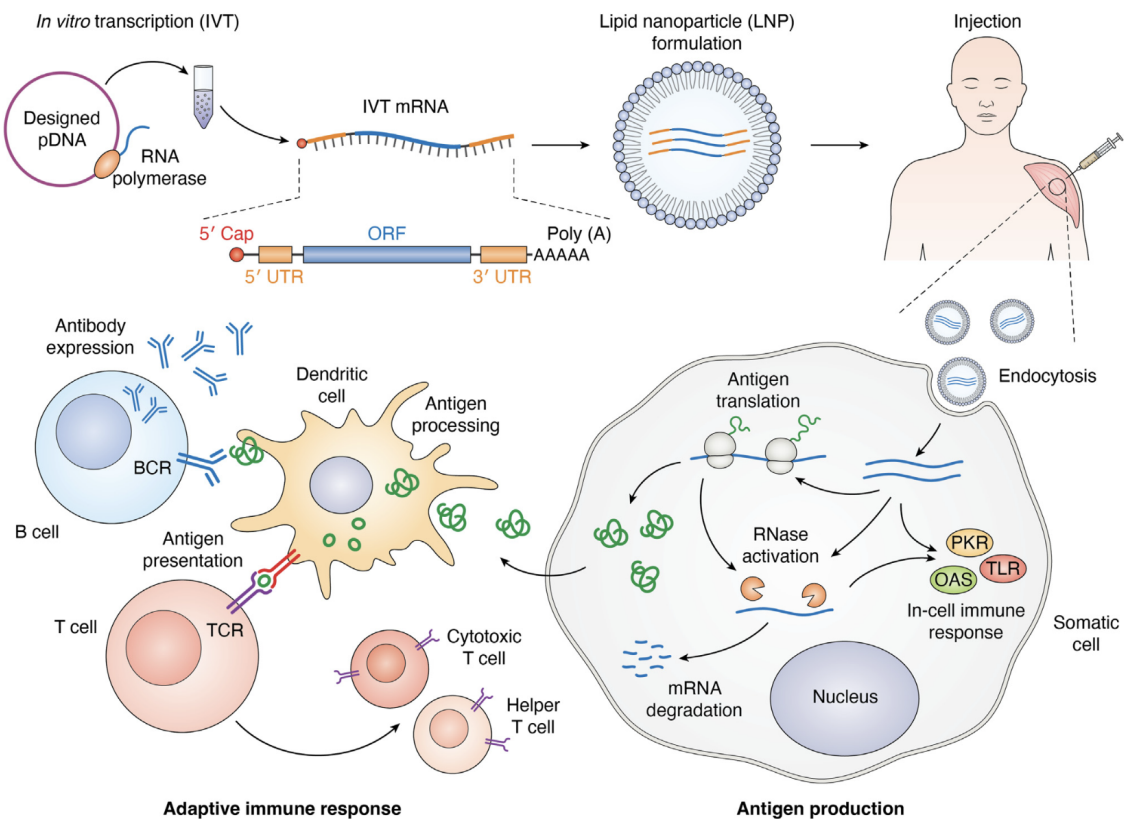


Figure 1 : Procédure simplifiée de fabrication et mécanisme d'action d'un vaccin à ARNm. À partir de la séquence génétique de la protéine cible, une séquence ADN optimisée est conçue, synthétisée puis clonée sous forme de pDNA. Ce plasmide sert de matrice aux ARN polymérases lors de la transcription *in vitro*, produisant des ARNm artificiels ensuite encapsulés dans des nanoparticules lipidiques (LNP). Après injection intramusculaire, les LNP facilitent l'entrée des ARNm dans les cellules somatiques par endocytose. Les ARNm vaccinaux libérés dans le cytoplasme sont traduits en protéines antigéniques par les ribosomes de la cellule. Ces antigènes stimulent les réponses immunitaires adaptatives, activant les lymphocytes B et T contre les infections potentielles. Les ARNm injectés sont ensuite dégradés par les ribonucléases cellulaires (RNases). Adapté de Jin *et al.* (2025).

Les cinq composants structuraux de l'ARNm vaccinal

- I. **La coiffe 5' (5' cap)** : guanosine modifiée (m^7G) qui stabilise l'ARNm contre la dégradation exonucléolytique en 5' et participe à l'initiation de la traduction par les ribosomes dans la cellule.
- II. **La région 5' non traduite (5' UTR)** : contient des éléments régulateurs cruciaux (site de liaison au ribosome (RBS) et site interne d'entrée du ribosome [IRES]) contrôlant la traduction en recrutant le ribosome et d'autres facteurs.
- III. **Le cadre ouvert de lecture (ORF)** : situé entre les codons d'initiation et de terminaison, il contient la **séquence codante (CDS)** qui encode la protéine antigénique cible. La dégénérescence du code génétique offre un vaste espace d'optimisation (Xia, 2021).
- IV. **La région 3' non traduite (3' UTR)** : protège l'ARNm contre la dégradation enzymatique dans le cytoplasme en recrutant les protéines de liaison au poly(A) (PABP).
- V. **La queue poly(A)** : séquence de résidus adényliques en 3' essentielle pour la stabilité de l'ARNm et facilitant le chargement du ribosome via les facteurs d'initiation eucaryotiques (eIF) lors de la circularisation.

I.4 Défis actuels des vaccins à ARNm

Malgré le succès remarquable des vaccins ARNm dans la prévention de la COVID-19, plusieurs défis subsistent pour leurs applications plus larges, en particulier dans le traitement du cancer, domaine qui a initialement motivé leur développement (Jin *et al.*, 2025). Les essais cliniques en oncologie, bien que prometteurs, n'ont pas encore abouti à une approbation réglementaire d'un vaccin ARNm thérapeutique anticancéreux, même si plus de 120 essais cliniques sont en cours à travers le monde (Yaremenko *et al.*, 2025).

01

Efficacité d'expression

Les ARNm artificiels ne permettent pas toujours de susciter une réponse immunitaire durable, nécessitant de multiples rappels (Jin *et al.*, 2025).

02

Instabilité thermique

L'ARNm exige un stockage à très basse température, posant des problèmes logistiques pouvant entraver la distribution en pandémie (Leppek *et al.*, 2022).

03

Réactogénicité

L'ARNm artificiel, ses vecteurs et des traductions protéiques indésirables peuvent entraîner des effets secondaires inattendus (Jin *et al.*, 2025).

I.5 Principes de l'optimisation computationnelle des séquences ARNm

Les propriétés d'un vaccin ARNm étant largement déterminées par sa séquence primaire, les efforts actuels d'amélioration de l'efficacité se concentrent principalement sur le design et l'optimisation de cette séquence (Kim *et al.*, 2024). Le recours aux approches computationnelles, incluant l'intelligence artificielle et l'apprentissage automatique, a profondément transformé la manière dont les séquences

vaccinales sont conçues, en permettant d'explorer des espaces séquentiels inaccessibles aux méthodes traditionnelles (Imani *et al.*, 2025 ; Hederman *et al.*, 2023).

Jin *et al.* (2025) identifient **quatre principes clés** pour un design optimal :

Principe 1 Optimisation des UTR

Le design des régions 5' et 3' non traduites améliore l'efficacité de chargement des ribosomes et l'efficacité de la traduction. Des modèles d'apprentissage automatique comme Optimus 5-Prime (Sample *et al.*, 2019) et UTR-LM (Chu *et al.*, 2024) ont été développés pour prédire et optimiser les séquences 5' UTR.

Principe 2 Optimisation des codons de l'ORF

Le choix des codons synonymes influence significativement l'efficacité de traduction et la durée de vie intracellulaire de l'ARNm. L'**indice d'adaptation des codons (CAI)** quantifie ce biais d'usage des codons (Sharp & Li, 1987). Les vaccins approuvés BNT162b2 et mRNA-1273 atteignent des valeurs de CAI $\approx 0,95$ – $0,98$, reflétant une forte adaptation aux gènes hautement exprimés.

Principe 3 Amélioration de la stabilité structurale globale

L'ARN, contrairement à l'ADN, est sujet à l'hydrolyse catalysée par les bases, plus probable dans les régions simple brin où les groupes chimiques sensibles sont exposés. La conception de séquences favorisant les structures secondaires réduit la susceptibilité à l'hydrolyse. Cette stabilité est mieux capturée par la **probabilité moyenne de nucléotides non appariés (AUP)** que par la seule énergie libre minimale de la structure, et son optimisation permet d'augmenter significativement la demi-vie des ARNm, avec des implications directes pour leur stockage et leur distribution (Wayment-Steele *et al.*, 2021).

Principe 4 Ajustement des structures locales

Les structures locales de l'ARNm (boucles, épingles à cheveux, jonctions) influencent la durée de vie intracellulaire, la vitesse locale de traduction et la fidélité de traduction. Certains motifs activant les endoribonucléases ou stimulant des décalages du cadre de lecture doivent être évités (Jin *et al.*, 2025).

I.6 Principales réalisations récentes et choix de l'application

Ces dernières années, plusieurs outils computationnels ont été développés pour mettre en œuvre ces principes :

Outil	Description
CDSfold	Programmation dynamique pour les structures secondaires stables ; complexité $O(L^3)$ limitante pour les longues séquences (Teraï <i>et al.</i> , 2016).
RiboTree	Recherche arborescente de Monte Carlo couplée à LinearPartition (un algorithme de calcul rapide des probabilités d'appariement de chaque nucléotide) pour minimiser l'AUP, un prédicteur de stabilité plus fin que la seule MFE car il intègre l'ensemble des structures possibles. Sans garantie d'optimalité globale (Wayment-Steele <i>et al.</i> , 2021).
DegScore	Régression linéaire estimant le taux de dégradation en solution (Leppek <i>et al.</i> , 2022).
RNA-FM, UTR-LM	Modèles de langage ARN pour la prédiction séquence–fonction (Jin <i>et al.</i> , 2025).

Parallèlement, le *deep learning* transforme plus largement la conception vaccinale, notamment à travers la prédiction de structures protéiques et l'analyse des répertoires immunitaires (Hederman *et al.*, 2023).

Choix de l'application : LinearDesign

Parmi ces développements, **LinearDesign** (Zhang *et al.*, 2023) se distingue par sa capacité à combiner l'optimisation de l'usage des codons et la stabilité structurale dans un cadre computationnel efficace, en réduisant le temps de calcul de **plusieurs heures à quelques minutes**. Les résultats expérimentaux montrent que les séquences conçues par LinearDesign peuvent **trippler l'expression protéique** et **multiplier les titres d'anticorps par un facteur allant jusqu'à 128** chez la souris, par rapport au benchmark d'optimisation de codons seule. Le présent travail détaille cette innovation car elle illustre de manière remarquable l'alliance entre les sciences de l'ingénieur et les sciences biomédicales.

II Application : LinearDesign — optimisation algorithmique des séquences d'ARNm vaccinaux

II.1 Contexte et problématique

La conception d'un ORF pour un vaccin ARNm nécessite de trouver un équilibre entre l'**efficacité de traduction** (via l'optimisation des codons) et la **stabilité structurale** (mesurée par la MFE ou l'AUP), ces objectifs pouvant impliquer des compromis selon les stratégies de conception (Jin *et al.*, 2025), bien que certains travaux récents suggèrent qu'ils peuvent, dans certains cas, être optimisés conjointement (Leppek *et al.*, 2022).

En raison de la redondance du code génétique, un ORF codant pour une protéine antigénique comme la protéine spike du SARS-CoV-2 pourrait théoriquement présenter plus de 10^{632} variations de CDS possibles (Zhang *et al.*, 2023). L'exploration exhaustive de cet espace séquentiel étant computationnellement impossible, des algorithmes efficaces sont nécessaires.

Les outils existants avant LinearDesign, tels que CDSfold (Terai *et al.*, 2016), présentaient des limitations majeures. CDSfold nécessite plusieurs heures pour analyser un ORF de quelques milliers de codons en raison de sa complexité temporelle en $O(L^3)$ et spatiale en $O(L^2)$, un frein considérable au développement rapide de vaccins en situation de pandémie.

II.2 L'élégance de l'analogie linguistique : du mot au codon

LinearDesign, développé par Zhang *et al.* (2023), représente une avancée majeure en adressant la complexité computationnelle inhérente aux tâches d'optimisation de séquences. L'innovation fondamentale réside dans l'adaptation d'un concept issu de la **linguistique computationnelle** : le **lattice parsing** (analyse sur treillis), permettant de traiter simultanément un ensemble exponentiel de séquences candidates.

La méthode se base sur l'hypothèse que le problème de la conception d'ARNm possède une structure formellement identique à ce problème classique du traitement automatique du langage naturel.

En reconnaissance vocale, un signal audio ambigu peut correspondre à plusieurs phrases candidates. Par exemple, le son « *I like this meal* » pourrait aussi être interprété comme « *alike his veal* » ou d'autres combinaisons phonétiquement proches. Le système doit alors déterminer, parmi toutes ces alternatives, la phrase la plus *grammaticalement correcte* et la plus *probable*. Pour ce faire, les linguistes computationnels représentent l'ensemble des alternatives sous forme d'un **automate fini déterministe (DFA)**, c'est-à-dire un graphe orienté où chaque chemin du début à la fin correspond à une phrase candidate. Un algorithme de *parsing* sur ce treillis évalue ensuite simultanément toutes les phrases possibles à l'aide d'une grammaire formelle, sans avoir à les énumérer une par une (Zhang *et al.*, 2023).

Zhang *et al.* montrent que la conception d'un ARNm obéit exactement à la même logique, à travers une correspondance terme à terme entre les deux domaines :

Correspondance linguistique $\longleftrightarrow$ biologique

Mot	$\longleftrightarrow$	Codon (triplet de nucléotides)
Synonymes phonétiques	$\longleftrightarrow$	Codons synonymes encodant le même acide aminé
Phrase	$\longleftrightarrow$	Séquence ARNm complète (CDS)
Treillis de mots (DFA)	$\longleftrightarrow$	Treillis de codons (mRNA DFA)
Grammaire du langage	$\longleftrightarrow$	Modèle thermodynamique de repliement de l'ARN (SCFG)
Phrase la plus grammaticale	$\longleftrightarrow$	ARNm ayant la structure secondaire la plus stable (MFE minimale)
Probabilité du mot	$\longleftrightarrow$	Adaptivité relative du codon $w(c)$ pour le CAI

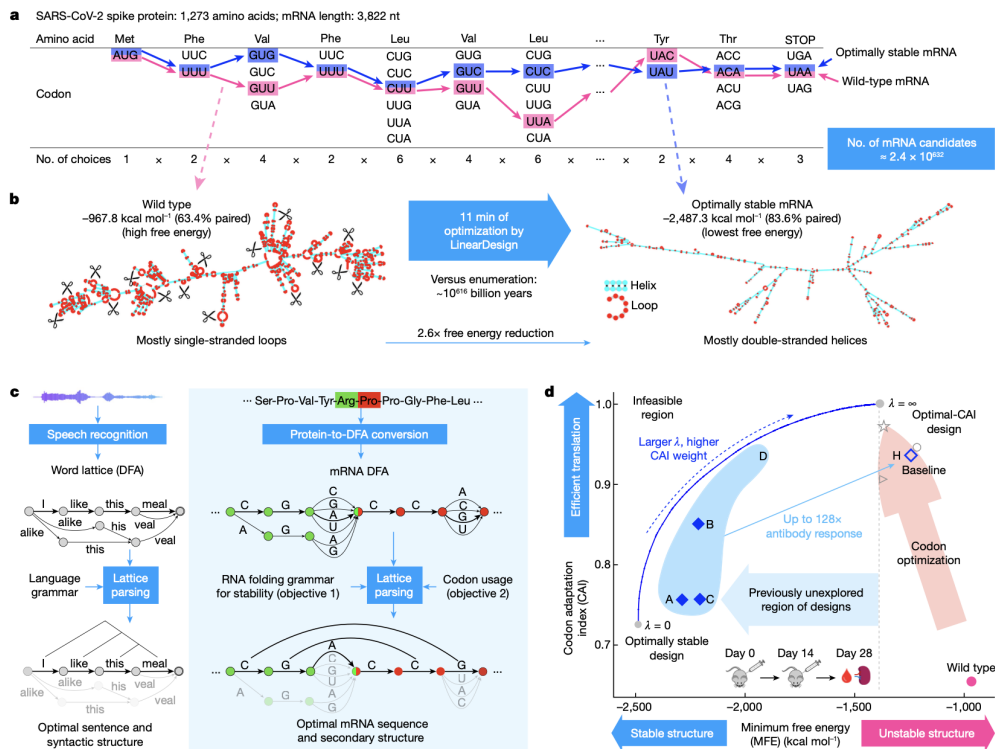


Figure 2: Vue d'ensemble du design de la région codante d'un ARNm optimisant simultanément stabilité et usage des codons, illustrée sur la protéine spike du SARS-CoV-2. **(a)** Espace combinatoire des séquences synonymes ($\approx 2,4 \times 10^{632}$ candidats). Les chemins rose et bleu représentent respectivement la séquence sauvage et la séquence optimalement stable. **(b)** Structures secondaires de l'ARNm sauvage (majoritairement simple brin, MFE = $-967,8$ kcal mol⁻¹) et de l'ARNm optimisé par LinearDesign (majoritairement double brin, MFE = $-2\,487,3$ kcal mol⁻¹). **(c)** Correspondance entre le *lattice parsing* en linguistique computationnelle (gauche) et son adaptation au design d'ARNm (droite) : le treillis de mots devient un DFA de codons, la grammaire du langage devient le modèle thermodynamique de repliement (SCFG). **(d)** Espace de design bidimensionnel MFE–CAI. L'optimisation classique des codons (flèche rose) n'explore pas la région haute stabilité ; LinearDesign (courbe bleue, paramètre λ) y accède et produit des réponses anticorps jusqu'à 128 fois supérieures chez la souris. Reproduit de Zhang *et al.*(2023).

La profondeur de cette analogie est remarquable. En linguistique, chaque mot possède un certain nombre de synonymes phonétiques (homophones ou quasi-homophones) qui peuvent s'y substituer dans le signal audio. De manière parfaitement parallèle, chaque acide aminé de la protéine cible peut être encodé par un nombre variable de **codons synonymes** — de un seul (méthionine : AUG) jusqu'à six (leucine, sérine, arginine) — en raison de la dégénérescence du code génétique. C'est cette redondance qui engendre l'espace combinatoire colossal de $2,4 \times 10^{632}$ séquences pour la seule protéine spike.

La construction du **DFA de codons** procède en deux étapes (Zhang *et al.*, 2023). D'abord, pour chaque acide aminé de la protéine cible, un petit automate fini est construit : ses arêtes sont étiquetées par les nucléotides, et chaque chemin du début à la fin du mini-graphe correspond à l'un des codons synonymes de cet acide aminé. Ensuite, ces automates élémentaires sont *concaténés* bout à bout pour former un unique **DFA global** représentant l'intégralité de l'espace de conception. Chaque chemin à travers ce graphe global, de l'état initial à l'état final, correspond à une séquence d'ARNm candidate encodant la protéine cible.

L'étape clé est alors le **lattice parsing** proprement dit. Le repliement de l'ARN peut être interprété comme un problème d'analyse structurale, similaire à ceux rencontrés en linguistique, visant à associer à une séquence une structure optimale. Des modèles probabilistes permettent ainsi d'évaluer l'énergie associée aux différentes structures secondaires possibles.

D'un point de vue formel, ce problème peut être représenté à l'aide de **grammaires stochastiques hors contexte (SCFG)**, dans lesquelles chaque règle de production est associée à un terme d'énergie libre (Zhang *et al.*, 2023). L'analyse d'une séquence unique correspond alors à un problème classique de programmation dynamique, résolu notamment par des algorithmes de type CKY (Cocke–Kasami–Younger), qui évaluent efficacement l'ensemble des structures possibles en temps cubique $O(n^3)$.

L'innovation de LinearDesign est de généraliser cette analyse à l'ensemble des séquences représentées dans le treillis. Au lieu d'évaluer chaque séquence individuellement, l'algorithme traite simultanément toutes les possibilités en exploitant le partage de sous-structures entre les différents chemins du graphe.

Extension au CAI : le DFA pondéré Pour intégrer l'optimisation des codons, Zhang *et al.* introduisent une seconde innovation. Le DFA est transformé en **DFA pondéré** : chaque arête du graphe reçoit un poids $-\log w(c)$, où $w(c)$ est l'adaptivité relative du codon c . Le coût total d'un chemin dans ce DFA pondéré est alors proportionnel à $-\log(\text{CAI})$ de la séquence correspondante. Le *lattice parsing* opère alors sur l'intersection pondérée entre la SCFG (stabilité) et le DFA pondéré (usage des codons), résolvant ainsi le problème d'optimisation conjointe avec garantie d'optimalité.

Ce pont inattendu entre linguistique et biologie offre non seulement un algorithme efficace, puisque le temps de calcul ne croît empiriquement que quadratiquement avec la longueur de la séquence, grâce à la compacité du DFA ; il apporte aussi une élégance conceptuelle qui transforme l'immensité de l'espace de conception en un *avantage* plutôt qu'un obstacle, en offrant une liberté de design auparavant inaccessible (Zhang *et al.*, 2023).

Analogie fondamentale Identifier la séquence d'ARNm optimale parmi toutes les combinaisons de codons synonymes est comparable à sélectionner la phrase la plus probable parmi un ensemble d'alternatives similaires en linguistique computationnelle. Chaque acide aminé correspond à un « mot » pouvant être encodé par plusieurs codons synonymes. Le treillis représente l'ensemble de ces alternatives, et l'algorithme recherche le « chemin » optimal à travers ce treillis.

Concrètement, LinearDesign utilise un algorithme de **programmation dynamique** amélioré, fonctionnant de gauche à droite (de 5' vers 3'), dans le cadre du *lattice parsing*. Cet algorithme intègre simultanément deux critères d'optimisation : la sélection de codons favorables (CAI élevé) et le repliement vers une structure secondaire stable (MFE basse).

FONCTION DE SCORE COMBINÉE

$$\text{Score} = \text{MFE} - \lambda \cdot |p| \cdot \log(\text{CAI}) \quad (1)$$

λ — paramètre de pondération contrôlant le compromis stabilité / traduction

$|p|$ — longueur de la séquence protéique (nombre de codons)

MFE — énergie libre minimale de la structure secondaire

CAI — indice d'adaptation des codons

En faisant varier λ , LinearDesign explore différentes solutions optimales correspondant à différents compromis entre stabilité structurale et efficacité de traduction.

II.3 Résultats expérimentaux : vaccins SARS-CoV-2 et VZV

II.3.1 Vaccin contre le SARS-CoV-2 (protéine spike)

Avant de discuter les performances obtenues par LinearDesign, il est utile de rappeler ce que représente la **protéine Spike (S)**. Dans le virion SARS-CoV-2, Spike est une glycoprotéine trimérique exposée à la surface de l'enveloppe virale. Elle joue un rôle central dans les premières étapes du cycle infectieux : elle reconnaît le récepteur **ACE2** à la surface des cellules hôtes, puis, après activation par des protéases cellulaires telles que TMPRSS2 ou les cathepsines, permet l'entrée du virus par fusion membranaire ou par endocytose. Cette étape conduit à la libération du génome viral dans le cytoplasme, point de départ de la réplication virale. C'est pourquoi Spike constitue une cible antigénique majeure des vaccins à ARNm contre la COVID-19 : les anticorps anti-Spike peuvent neutraliser les virions libres en bloquant l'attachement ou l'entrée du virus dans les cellules cibles (Peng *et al.*, 2021 ; Chaudhary *et al.*, 2021).

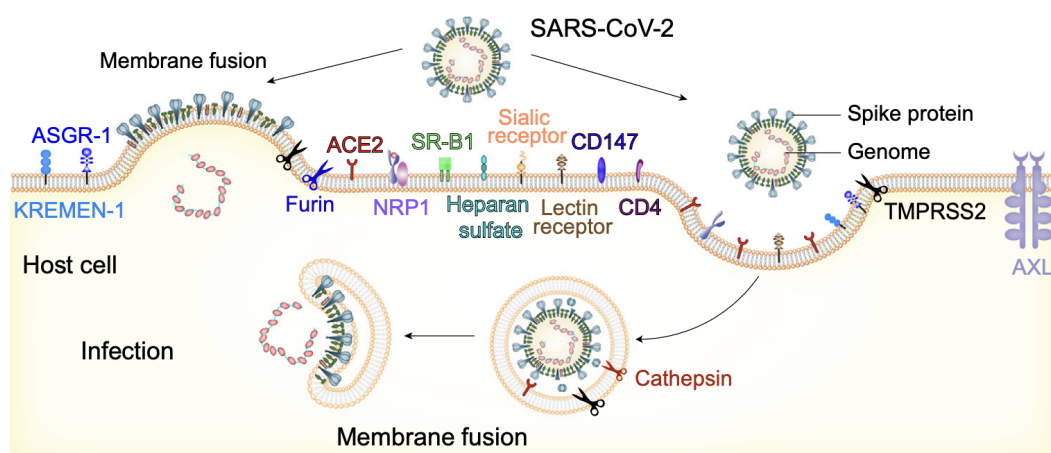


Figure 3: Voies d'entrée du SARS-CoV-2 dans la cellule hôte. La protéine Spike (S), située à la surface du virion, permet la reconnaissance du récepteur ACE2 et participe à l'entrée du virus, soit par fusion directe avec la membrane plasmique, soit après endocytose. Cette étape permet la libération du génome viral dans le cytoplasme et constitue le point de départ du cycle réplcatif. Reproduit de Peng *et al.* (2021).

Pour le design du vaccin ciblant la protéine spike du SARS-CoV-2, LinearDesign a identifié des séquences d'ARNm capables d'atteindre une expression protéique jusqu'à trois fois supérieure dans les cellules, comparé aux vaccins établis tels que BNT162b2 (Pfizer/BioNTech) et mRNA-1273 (Moderna) (Zhang *et al.*, 2023). De plus, les titres d'anticorps mesurés chez la souris ont été significativement augmentés, avec un facteur d'amélioration allant jusqu'à 128 fois par rapport au benchmark d'optimisation de codons seule (Zhang *et al.*, 2023).

Un résultat clé illustré par la Figure 4 est que la séquence offrant la meilleure réponse immunitaire adaptative ne correspond ni au CAI le plus élevé ni à la MFE la plus basse. Les vaccins approuvés Moderna mRNA-1273 et Pfizer BNT162b2 se situent dans la zone d'optimisation des codons (CAI élevé), tandis que les séquences générées par LinearDesign explorent une région intermédiaire où la stabilité structurale et l'efficacité de traduction sont équilibrées. Cette observation confirme que l'optimisation multi-objectif est essentielle pour maximiser les performances vaccinales.

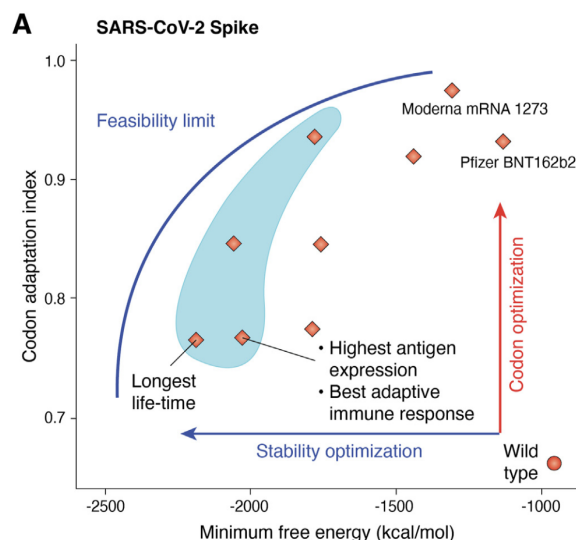


Figure 4: Espace de design MFE–CAI pour la protéine spike du SARS-CoV-2. Les vaccins approuvés Moderna mRNA-1273 et Pfizer BNT162b2 se situent dans la zone d'optimisation des codons (CAI élevé, flèche rouge). Les séquences générées par LinearDesign (zone cyan) explorent la région haute stabilité, inaccessible par optimisation des codons seule. La séquence offrant la meilleure réponse immunitaire adaptative se situe dans la zone intermédiaire. La courbe bleue représente la *feasibility limit* (λ variant de 0 à ∞). Reproduit de Jin *et al.* (2025).

II.3.2 Vaccin contre le virus varicelle-zona (VZV, protéine gE)

Le **virus varicelle-zona (VZV)** est un alphaherpèsvirus enveloppé dont le virion comprend une nucléocapside icosaédrique contenant un génome d'ADN double brin, un tégment protéique et une enveloppe lipidique portant plusieurs glycoprotéines de surface. Parmi celles-ci, la **glycoprotéine E (gE)** est la plus abondante et l'une des glycoprotéines essentielles du VZV. Exposée à la surface du virion et des cellules infectées, elle participe à l'infektivité virale, à l'enveloppement secondaire et surtout à la dissémination de cellule à cellule, notamment en interaction avec la glycoprotéine I (gI). Son exposition à la surface du virion, son caractère immunodominant et son rôle central dans le cycle infectieux expliquent qu'elle constitue une cible vaccinale pertinente (Oliver *et al.*, 2016 ; Liu *et al.*, 2025).

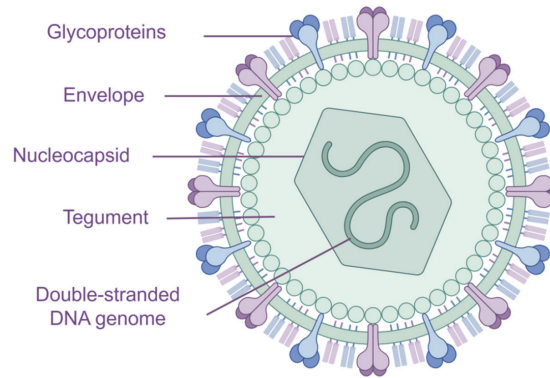


Figure 5: Structure schématique du virion VZV. Le virus varicelle-zona est un virus enveloppé contenant une nucléocapside, un tégment protéique et un génome ADN double brin. Son enveloppe porte plusieurs glycoprotéines virales, dont la glycoprotéine E (gE), cible antigénique majeure des vaccins contre le zona. gE participe notamment à l'infectivité virale et à la dissémination de cellule à cellule, ce qui explique son intérêt dans le design vaccinal. Adapté de Liu *et al.* (2025) et Oliver *et al.* (2016).

Des résultats similaires ont été obtenus pour le design de vaccins ARNm ciblant la protéine gE du virus varicelle-zona (VZV). Comme le montre la Figure 6, la séquence induisant la meilleure réponse immunitaire adaptative se situe également dans la zone intermédiaire entre optimisation des codons et optimisation de la stabilité. La séquence présentant l'expression antigénique la plus élevée et celle offrant la durée de vie la plus longue correspondent à des paramètres différents, soulignant l'importance du paramètre λ dans le réglage fin de l'algorithme (Zhang *et al.*, 2023).

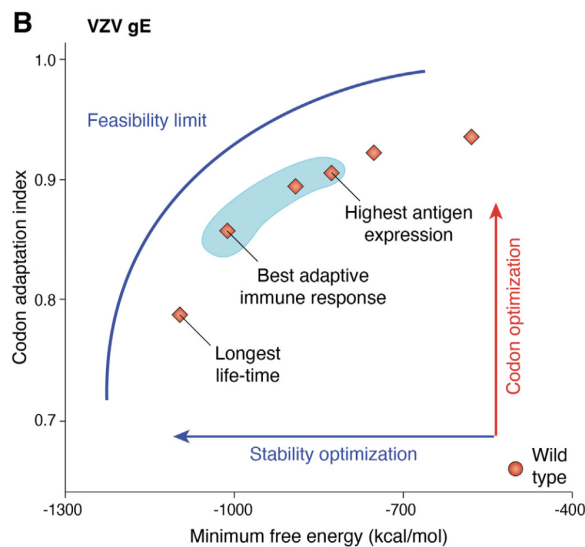


Figure 6: Espace de design MFE-CAI pour la protéine gE du VZV. Un schéma similaire à celui de la protéine spike est observé : la séquence induisant la meilleure expression antigénique et celle offrant la durée de vie la plus longue correspondent à des valeurs de λ différentes, confirmant la nécessité d'une optimisation multi-objectif spécifique à chaque antigène. Reproduit de Jin *et al.* (2025).

Il est également intéressant de noter que les valeurs optimales de λ diffèrent entre les cibles protéiques du SARS-CoV-2 et du VZV, suggérant que l'équilibre idéal entre CAI et MFE est spécifique à chaque antigène cible. Cette observation justifie l'approche de LinearDesign consistant à explorer

une gamme de valeurs de λ plutôt que de se fier à une valeur unique (Zhang *et al.*, 2023).

II.4 Comparaison avec les approches concurrentes

Le tableau 1 résume les principales caractéristiques de LinearDesign comparé à d'autres outils d'optimisation de séquences d'ORF pour les vaccins ARNm.

Table 1: Comparaison des outils computationnels d'optimisation d'ORF pour les vaccins ARNm.

Outil	Métrique	Temps	Multi-obj.	Référence
CDSfold	MFE	Heures	<i>Limité</i>	Terai <i>et al.</i> , 2016
RiboTree	AUP	Variable	Non	Wayment-Steele <i>et al.</i> , 2021
DegScore	Dégradation	Rapide	Non	Leppek <i>et al.</i> , 2022
LinearDesign	MFE + CAI	Minutes	Oui	Zhang <i>et al.</i> , 2023

LinearDesign se distingue par sa capacité à réduire drastiquement le temps de calcul (de plusieurs heures à quelques minutes) tout en intégrant une optimisation multi-objectif via la fonction $MFE - \lambda|p| \log(CAI)$.

III Discussion et conclusion

III.1 Rappel de la problématique

La conception de vaccins ARNm efficaces requiert l'optimisation simultanée de multiples paramètres souvent antagonistes : efficacité de traduction, stabilité en solution, durée de vie intracellulaire et fidélité d'expression. La complexité combinatoire de l'espace séquentiel rend les approches naïves d'optimisation impraticables (Jin *et al.*, 2025). Cette problématique ne se limite pas aux vaccins anti-infectieux : les vaccins ARNm anticancéreux, qui encodent des néoantigènes spécifiques aux tumeurs, nécessitent les mêmes optimisations de séquences pour maximiser l'expression antigénique et la réponse immunitaire (Lorentzen *et al.*, 2022).

III.2 Solutions apportées par LinearDesign

LinearDesign apporte des réponses concrètes à cette problématique en proposant un algorithme capable de naviguer efficacement dans l'immense espace séquentiel des codons synonymes. En transposant le problème de l'optimisation biologique vers un cadre de linguistique computationnelle, l'algorithme parvient à identifier des séquences qui équilibrent stabilité structurale et efficacité de traduction, le tout en un temps de calcul considérablement réduit (Zhang *et al.*, 2023). Les résultats expérimentaux obtenus pour les vaccins ciblant le SARS-CoV-2 et le VZV montrent que cette approche permet d'identifier des séquences présentant des améliorations significatives en termes d'expression protéique et de réponse immunitaire par rapport à des stratégies d'optimisation classiques (Zhang *et al.*, 2023).

Nature de la validation immunitaire. Un vaccin antiviral efficace ne doit pas seulement induire des anticorps. Il doit idéalement activer deux branches complémentaires de l'immunité adaptative.

La première est la **réponse humorale**, c'est-à-dire la production d'anticorps par les lymphocytes B. Ces anticorps peuvent neutraliser les virions libres en se fixant aux protéines de surface utilisées par le virus pour entrer dans les cellules. Ils empêchent ainsi le virus d'infecter de nouvelles cellules.

La seconde est la **réponse cellulaire**, notamment portée par les lymphocytes T CD8⁺ cytotoxiques. Ceux-ci reconnaissent les cellules déjà infectées grâce aux fragments d'antigènes viraux présentés par le CMH I à leur surface, puis déclenchent leur mort. Cette réponse est importante car une cellule infectée peut devenir une véritable usine à nouveaux virions. Le même raisonnement vaut pour les vaccins anticancéreux : l'objectif est alors d'aider le système immunitaire à reconnaître et éliminer des cellules tumorales exprimant des antigènes ou néoantigènes spécifiques.

Dans l'étude de Zhang *et al.* (2023), les auteurs ne mesurent donc pas uniquement la réponse humorale. Pour le modèle SARS-CoV-2, ils évaluent des IgG anti-Spike, des anticorps neutralisants, mais aussi des lymphocytes T spécifiques de Spike capables de sécréter l'IFN- γ . Cela indique l'activation d'une réponse T antigène-spécifique, plutôt orientée Th1. Pour le modèle VZV-gE, la figure principale met surtout en avant l'augmentation des IgG anti-gE, mais les méthodes indiquent également une évaluation des cellules T spécifiques de l'antigène par ELISpot IFN- γ .

Il faut toutefois nuancer : une sécrétion d'IFN- γ montre qu'une réponse T spécifique est activée, mais ne prouve pas directement une activité cytotoxique LT8 au sens strict. Pour démontrer formellement une cytotoxicité CD8⁺, il faudrait par exemple identifier précisément les cellules CD8⁺ impliquées et

mesurer leur capacité de destruction des cellules cibles, via des marqueurs comme perforine/granzyme ou un test de lyse cellulaire.

III.3 Points forts et limites

+ Points forts

Rapidité

Design de la protéine spike en 11 minutes, contre plusieurs heures pour CDSfold. Compatible avec le développement en urgence pandémique.

Validation expérimentale

Deux cibles antigéniques (SARS-CoV-2 et VZV), trois attributs mesurés (stabilité, expression, immunogénicité), résultats cohérents.

Multi-objectif

Première méthode à optimiser conjointement CAI et MFE dans un cadre de programmation dynamique avec garantie d'optimalité.

Adoption industrielle

Sanofi a obtenu une licence non exclusive pour intégrer LinearDesign dans son pipeline de développement de vaccins ARNm (Zhang *et al.*, 2023).

— Limites

Fidélité de traduction

L'algorithme ne prend pas en compte les structures locales affectant la fidélité de traduction. Jusqu'à 30 % de protéines altérées par décalage du cadre de lecture ont été observées avec les vaccins ARNm COVID-19 (Jin *et al.*, 2025 ; Mulroney *et al.*, 2024).

Paramètre λ

La valeur optimale de λ diffère selon les protéines cibles, rendant sa détermination non triviale. L'exploration d'une gamme de λ est nécessaire pour chaque nouvel antigène (Zhang *et al.*, 2023).

Stabilité $\neq$ expression

La stabilité globale en solution corrèle faiblement avec la durée de vie intracellulaire. Des structures trop stables dans l'ORF peuvent ralentir la progression ribosomale et activer des mécanismes de dégradation (Jin *et al.*, 2025).

Facteurs exclus

LinearDesign n'intègre ni les modifications nucléosidiques (m¹Ψ), ni l'optimisation des LNP, ni l'ingénierie des UTR, trois facteurs pourtant déterminants pour l'efficacité vaccinale (Nance & Meier, 2021 ; Hou *et al.*, 2021).

III.4 Perspectives futures et actualité

Les perspectives futures pour l'optimisation computationnelle des vaccins ARNm sont multiples. Les modèles intégrés devront considérer simultanément la structure globale (stabilité en solution), les structures locales (fidélité de traduction et durée de vie intracellulaire), et les interactions entre les différents segments de l'ARNm (UTR et ORF) (Jin *et al.*, 2025). L'expansion des bases de données

expérimentales portant sur la stabilité, la dégradation et l'efficacité de traduction des ARNm, comme celle initiée par PERSIST-seq (Leppek *et al.*, 2022), sera essentielle pour entraîner des modèles d'apprentissage profond plus performants.

L'IA AU-DELÀ DE L'OPTIMISATION DE SÉQUENCES

1. Structure 3D des antigènes

AlphaFold et ses successeurs permettent d'inférer la conformation tridimensionnelle d'une protéine à partir de sa séquence. Cette information est cruciale car la structure d'un antigène influence directement sa reconnaissance par le système immunitaire (Hederman & Ackerman, 2023).

2. Répertoires immunitaires

Le *deep learning* est mobilisé pour analyser les répertoires d'anticorps et de récepteurs T, dont la diversité est générée par la recombinaison V(D)J, la diversité jonctionnelle et, pour les anticorps, l'hypermutation somatique. Il permet ainsi de modéliser les interactions antigène-récepteur et de prédire l'immunogénicité de séquences candidates avant toute validation expérimentale (Imani *et al.*, 2025).

3. Optimisation des LNP

L'apprentissage automatique est utilisé pour optimiser les formulations de nanoparticules lipidiques, en améliorant leur stabilité, leur biodistribution et leur efficacité de délivrance intracellulaire, ce qui représente un facteur aussi déterminant que la séquence elle-même (Imani *et al.*, 2025).

Ces avancées convergentes ouvrent la voie à une approche holistique du design vaccinal, intégrant séquence ARNm, structure antigénique et vecteur de livraison.

Signal clinique Du côté thérapeutique, les vaccins ARNm anticancéreux progressent rapidement. Le vaccin personnalisé **mRNA-4157 (V940)** de Moderna, qui encode des néoantigènes tumoraux propres à chaque patient, est administré en combinaison avec le *pembrolizumab*, un anticorps monoclonal anti-PD-1 utilisé comme inhibiteur de point de contrôle immunitaire. Cette combinaison a réduit le risque de récurrence de **44 %** chez des patients atteints de mélanome à haut risque en essai de phase IIb, et est actuellement en **phase III** (Yaremenko *et al.*, 2025). Ces résultats illustrent le potentiel croissant de la plateforme ARNm au-delà des maladies infectieuses.

III.5 Conclusion

La conception de vaccins à ARNm s'impose aujourd'hui comme un problème intrinsèquement multidimensionnel, à l'interface de la biologie moléculaire, de la biophysique et de l'informatique. L'optimisation des séquences codantes, longtemps abordée sous l'angle unique de l'efficacité de traduction, apparaît désormais comme un problème d'optimisation multi-objectif, impliquant des compromis subtils entre stabilité structurale, fidélité de traduction et expression protéique.

Néanmoins, les limites identifiées soulignent que l'optimisation de l'ORF, bien que centrale, ne saurait à elle seule capturer la complexité du système. Les interactions entre structures locales et globales, les effets des modifications nucléosidiques, la dynamique de traduction ainsi que les contraintes liées à la délivrance intracellulaire constituent autant de facteurs déterminants encore partiellement

modélisés. La performance d'un vaccin ARNm résulte d'un système intégré, où chaque niveau, de la séquence primaire à la réponse immunitaire, interagit de manière non linéaire.

Les perspectives résident dans le développement de modèles unifiés capables d'intégrer ces différentes dimensions au sein d'un cadre prédictif cohérent. L'essor des bases de données expérimentales à grande échelle et des approches d'apprentissage profond ouvre la voie à une nouvelle génération d'outils capables de capturer ces interactions et de guider la conception de manière plus rationnelle.

Plus largement, l'intégration de l'intelligence artificielle dans le design vaccinal marque une transition vers une ingénierie biomédicale fondée sur la donnée, où la conception des antigènes, l'optimisation des séquences et celle des systèmes de délivrance convergent vers une approche systémique. Les vaccins à ARNm ne constituent pas seulement une innovation technologique, mais le point d'ancrage d'un paradigme émergent dans lequel la modélisation computationnelle devient un levier central de la médecine de précision.

Références bibliographiques

- Chaudhary, N., Weissman, D., & Whitehead, K. A. (2021). mRNA vaccines for infectious diseases: Principles, delivery and clinical translation. *Nature Reviews. Drug Discovery*, 20(11), 817–838. <https://doi.org/10.1038/s41573-021-00283-5>
- Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L., Zhang, J., & Wang, M. (2024). A 5' UTR Language Model for Decoding Untranslated Regions of mRNA and Function Predictions. *Nature Machine Intelligence*, 6(4), 449–460. <https://doi.org/10.1038/s42256-024-00823-9>
- Depledge, D. P., Sadaoka, T., & Ouwendijk, W. J. D. (2018). Molecular Aspects of Varicella-Zoster Virus Latency. *Viruses*, 10(7), 349. <https://doi.org/10.3390/v10070349>
- Hederman, A. P., & Ackerman, M. E. (2023). Leveraging Deep Learning to Improve Vaccine Design. *Trends in Immunology*, 44(5), 333–344. <https://doi.org/10.1016/j.it.2023.03.002>
- Hou, X., Zaks, T., Langer, R., & Dong, Y. (2021). Lipid nanoparticles for mRNA delivery. *Nature Reviews Materials*, 6(12), 1078–1094. <https://doi.org/10.1038/s41578-021-00358-0>
- Imani, S., Li, X., Chen, K., Maghsoudloo, M., Jabbarzadeh Kaboli, P., Hashemi, M., Khoushab, S., & Li, X. (2025). Computational biology and artificial intelligence in mRNA vaccine design for cancer immunotherapy. *Frontiers in Cellular and Infection Microbiology*, 14. <https://doi.org/10.3389/fcimb.2024.1501010>
- Jin, L., Zhou, Y., Zhang, S., & Chen, S.-J. (2025). mRNA vaccine sequence and structure design and optimization: Advances and challenges. *The Journal of Biological Chemistry*, 301(1), 108015. <https://doi.org/10.1016/j.jbc.2024.108015>
- Kim, Y.-A., Mousavi, K., Yazdi, A., Zwierzyna, M., Cardinali, M., Fox, D., Peel, T., Collier, J., Aggarwal, K., & Maruggi, G. (2024). Computational design of mRNA vaccines. *Vaccine*, 42(7), 1831–1840. <https://doi.org/10.1016/j.vaccine.2023.07.024>
- Leppek, K., Byeon, G. W., Ber, W., Arber, A., Torber, E., Peri, A., Caliskan, N., Sarnow, P., & Barna, M. (2022). Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nature Communications*, 13, 1536. <https://doi.org/10.1038/s41467-022-28776-w>
- Lorentzen, C. L., Haanen, J. B., Met, Ö., & Svane, I. M. (2022). Clinical advances and ongoing trials on mRNA vaccines for cancer treatment. *The Lancet. Oncology*, 23(10), e450–e458. [https://doi.org/10.1016/S1470-2045\(22\)00372-2](https://doi.org/10.1016/S1470-2045(22)00372-2)
- Mulroney, T. E., Pöyry, T., Yam-Puc, J. C., Rust, M., Harvey, R. F., Kalmar, L., Horner, E., Booth, L., Ferreira, A. P., Stoneley, M., Sawarkar, R., Mentzer, A. J., Lilley, K. S., Smales, C. M., von der Haar, T., Turtle, L., Dunachie, S., Klenerman, P., Thaventhiran, J. E. D., & Willis, A. E. (2024). N1-methylpseudouridylation of mRNA causes +1 ribosomal frameshifting. *Nature*, 625(7993), 189–194. <https://doi.org/10.1038/s41586-023-06800-3>
- Nance, K. D., & Meier, J. L. (2021). Modifications in an Emergency: The Role of N1-Methylpseudouridine in COVID-19 Vaccines. *ACS Central Science*, 7(5), 748–756. <https://doi.org/10.1021/acscentsci.1c00197>
- Oliver, S. L., Yang, E., & Arvin, A. M. (2016). Varicella-Zoster Virus Glycoproteins: Entry, Replication, and Pathogenesis. *Current Clinical Microbiology Reports*, 3(4), 204–215. <https://doi.org/10.1007/s40588-016-0044-4>
- Pardi, N., Hogan, M. J., Porter, F. W., & Weissman, D. (2018). mRNA vaccines—A new era in vaccinology. *Nature Reviews. Drug Discovery*, 17(4), 261–279. <https://doi.org/10.1038/nrd.2017.243>
- Sample, P. J., Wang, B., Reid, D. W., Presnyak, V., McFadyen, I. J., Morris, D. R., & Seelig, G. (2019). Human

- 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology*, 37(7), 803–809. <https://doi.org/10.1038/s41587-019-0164-5>
- Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3), 1281–1295. <https://doi.org/10.1093/nar/15.3.1281>
- Shattock, A. J., Johnson, H. C., Sim, S. Y., Carter, A., Lambach, P., Hutubessy, R. C. W., ... Bar-Zeev, N. (2024). Contribution of vaccination to improved survival and health: Modelling 50 years of the Expanded Programme on Immunization. *The Lancet*, 403(10441), 2307–2316. [https://doi.org/10.1016/S0140-6736\(24\)00850-X](https://doi.org/10.1016/S0140-6736(24)00850-X)
- Terai, G., Kamegai, S., & Asai, K. (2016). CDSfold: An algorithm for designing a protein-coding sequence with the most stable secondary structure. *Bioinformatics*, 32(6), 828–834. <https://doi.org/10.1093/bioinformatics/btv678>
- Wayment-Steele, H. K., Kim, D. S., Choe, C. A., Nicol, J. J., Wellington-Oguri, R., Watkins, A. M., Parra Sperberg, R. A., Huang, P.-S., Participants, E., & Das, R. (2021). Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Research*, 49(18), 10604–10617. <https://doi.org/10.1093/nar/gkab764>
- Xia, X. (2021). Detailed Dissection and Critical Evaluation of the Pfizer/BioNTech and Moderna mRNA Vaccines. *Vaccines*, 9(7), 734. <https://doi.org/10.3390/vaccines9070734>
- Yaremenko, A. V., Khan, M. M., Zhen, X., Tang, Y., & Tao, W. (2025). Clinical advances of mRNA vaccines for cancer immunotherapy. *Med*, 6(1), 100562. <https://doi.org/10.1016/j.medj.2024.11.015>
- Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., Liu, B., Ma, X., Zhao, F., Jiang, H., Chen, C., Shen, H., Li, H., Mathews, D. H., Zhang, Y., & Huang, L. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, 621(7978), 396–403. <https://doi.org/10.1038/s41586-023-06127-z>
- Liu, H., Cui, L., Zhang, S., Wang, H., Xue, W., Li, H., Zhang, Y., Chen, L., Gu, Y., Li, T., Xia, N., & Li, S. (2025). Research Progress on Varicella-Zoster Virus Vaccines. *Vaccines*, 13(7), 730. <https://doi.org/10.3390/vaccines13070730>