bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Directed Evolution in Codon Space

James Heuschkel[1,2*], Laura Kingsley[1*], Jon Reed[1], Di Li[1], Matthew Warner[1], Noah Pefaur[1] and Steven Cramer[3*]

[1]Biotherapeutics Discovery Department, Boehringer Ingelheim Pharmaceutical Inc., Ridgefield, CT 06877, United States

[2] Department of Biochemistry and Biophysics and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, United States

[3] Department of Chemical and Biological Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, United States

*Corresponding authors. Emails: james.heuschkel@boehringer-ingelheim.com; laura.kingsley@boehringer-ingelheim.com; crames@rpi.edu;

Abstract

Directed evolution is commonly used in protein engineering, where mature molecules are routinely improved through iterative local search of amino acid space. Here, we extend this principle to coding DNA. We developed a language-model-guided framework that iteratively refined industry-optimized coding sequences of clinical-stage therapeutics through synonymous exploration of codon space. Across 23 antibody-based therapeutics, SynCodonLM-guided refinement significantly increased recombinant expression in CHO cells for 17 molecules (74% responder rate), without significant compromise of product-quality or biophysical attributes. Moreover, changes in model likelihood predicted expression gains more effectively than heuristic statistical or mRNA-structure descriptors, despite no explicit expression objective. Codon-level likelihood also tracked temporal progression in influenza A H1N1 sequences, indicating the model captures evolutionary signal. These results show that even production-optimized sequences retain accessible fitness in synonymous codon space, establishing directed

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

evolution as a practical strategy to improve biologic expression, a key manufacturing bottleneck, without altering protein sequence.

Introduction

The degeneracy of the genetic code creates a large design space in which many synonymous sequences encode the same protein, yet these choices are not functionally equivalent[1,2]. Synonymous codon usage influences translation kinetics[3], mRNA stability[4], co-translational folding[5], protein conformation and function[6], and product quality[7] - with downstream consequences in vaccines[8–10], oncolytic viruses[8], AAV payloads and capsids[9–11], and therapeutic antibody yield[12–14]. Synonymous substitutions are therefore not biologically silent.

This complexity has made codon optimization a routine component of recombinant protein and nucleic acid design, including applications in vaccines, gene therapies and therapeutic biologics[15]. Yet codon optimization remains methodologically unsettled. Classical approaches have relied heavily on host-frequency heuristics and related summary statistics, whereas newer machine-learning models generate host-specific coding sequences from protein input using context-aware sequence models[16–22].

A central limitation is that codon optimization is typically framed as a one-shot design problem, producing a fixed construct. This contrasts with modern protein engineering, where functional molecules are improved through iterative rounds of localized change, and where protein language models have shown that a small number of evolutionarily plausible edits can improve even highly mature antibodies[23].

Recent advances in codon language modeling suggest that an analogous framework may now be possible directly in coding DNA space[20,24–28]. Language models trained on coding DNA

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

sequences rather than amino acid sequences have been shown to capture biologically informative signals in downstream tasks, indicating that coding sequences contain useful information beyond the protein sequence they encode[26]. Codon language models trained under synonym-constrained prediction further raise the possibility that codon-level sequence modeling can isolate DNA-level constraints from amino-acid-level semantics and thereby provide a principled basis for post-design synonymous refinement[24].

Here, we show that codon-level likelihood reflects evolutionary structure, significantly tracking temporal progression in several influenza A (H1N1) proteins, establishing that the model encodes a biologically meaningful signal rather than a spurious statistical artifact. Leveraging this encoding, we apply our directed evolution approach to 23 patented, clinically deployed antibody-based therapeutics, treating each disclosed coding sequence as a high-fitness parental baseline for iterative synonymous refinement rather than as a fully optimized endpoint. At each step, a context-aware codon model proposes the synonymous substitution predicted to most improve local sequence compatibility, the sequence is updated, and the modified context is rescored before the next edit is selected. Most importantly, we demonstrate that these clinically validated therapeutic coding sequences remain locally improvable in codon space.

Results

Reframing codon optimization as an iterative search process

Whereas protein engineering routinely improves functional molecules through iterative rounds of localized change (**Fig. 1a**), codon optimization is typically performed as a one-time design step (**Fig. 1b**). We therefore reframed synonymous sequence design as an iterative search problem,

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

treating clinically validated coding sequences as high-fitness starting points for further refinement in codon space (**Fig. 1c**).

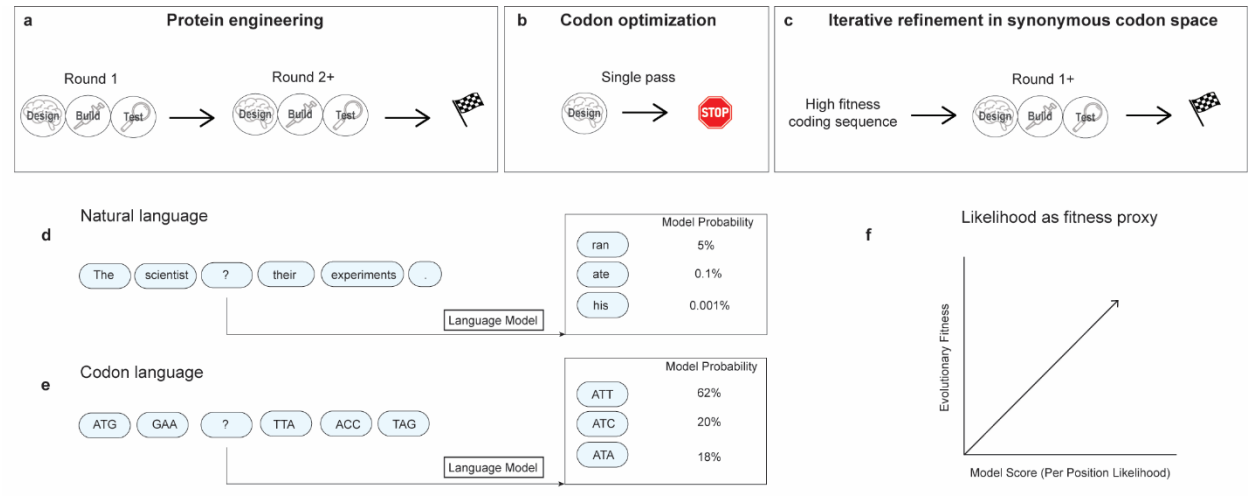


Fig. 1: Reframing codon optimization as an iterative, language-model-guided refinement in codon space. (a) Protein engineering improves functional molecules through iterative rounds of design-build-test cycles, progressively enhancing sequence performance. (b) Conventional codon optimization, by contrast, is performed as a single-pass design step, producing a fixed coding sequence that is not revisited. (c) Here, clinically validated coding sequences are treated as high-fitness starting points and iteratively refined in codon space through context-aware synonymous substitutions guided by a codon language model, without altering the encoded protein. (d) A natural-language model assigns probabilities to candidate words conditioned on their surrounding context. (e) Analogously, a codon language model assigns probabilities to synonymous codons conditioned on their local sequence context, scoring a masked position over its synonymous alternatives. (f) Model-assigned likelihood is used as a proxy for evolutionary fitness, consistent with prior work[23,29,30] showing that language-model likelihoods correlate with sequence fitness in protein engineering.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

To guide this search, we drew on an analogy to natural language modeling. Just as a natural-language model assigns probabilities to candidate words based on their surrounding context (**Fig. 1d**), a codon language model assigns probabilities to synonymous codons conditioned on local sequence context (**Fig. 1e**). Because language-model likelihoods have been shown to correlate with sequence fitness in protein engineering[23,29,30], we used codon-level likelihood as a proxy for sequence compatibility, defining an evolutionary score ("Evo score") for each coding sequence (**Fig. 1f**).

A codon language model captures evolutionary context

To first validate that SynCodonLM encodes biologically meaningful evolutionary information, we evaluated model-derived Evo scores on an independent viral evolution benchmark spanning Influenza A (H1N1) nucleoprotein, hemagglutinin, and neuraminidase coding sequences. Critically, this analysis was performed without task-specific fine-tuning, enabling direct assessment of whether codon-level likelihood alone reflects evolutionary trajectories.

Across all three proteins, SynCodonLM Evo scores increased significantly with sampling time ($r = 0.41–0.79$; all $p < 1e\text{-}3$; **Fig. 2a, c, e**), indicating that sequences observed later in viral evolution are assigned higher model compatibility. Nucleoprotein additionally showed a clear temporal shift, coincident with the 2009 pandemic strain A/California/07/2009, followed by sustained elevation in subsequent years. We also evaluated full-vocabulary (non-synonym-constrained) Evo scores (**Supplementary Fig. 1–3**). The temporal signal was reproduced for nucleoprotein and hemagglutinin, but was absent for neuraminidase, which retained a significant positive trend only under synonym-constrained scoring. This indicates that the synonym constraint can be necessary to isolate the codon-level evolutionary signal from amino-acid-level variation.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

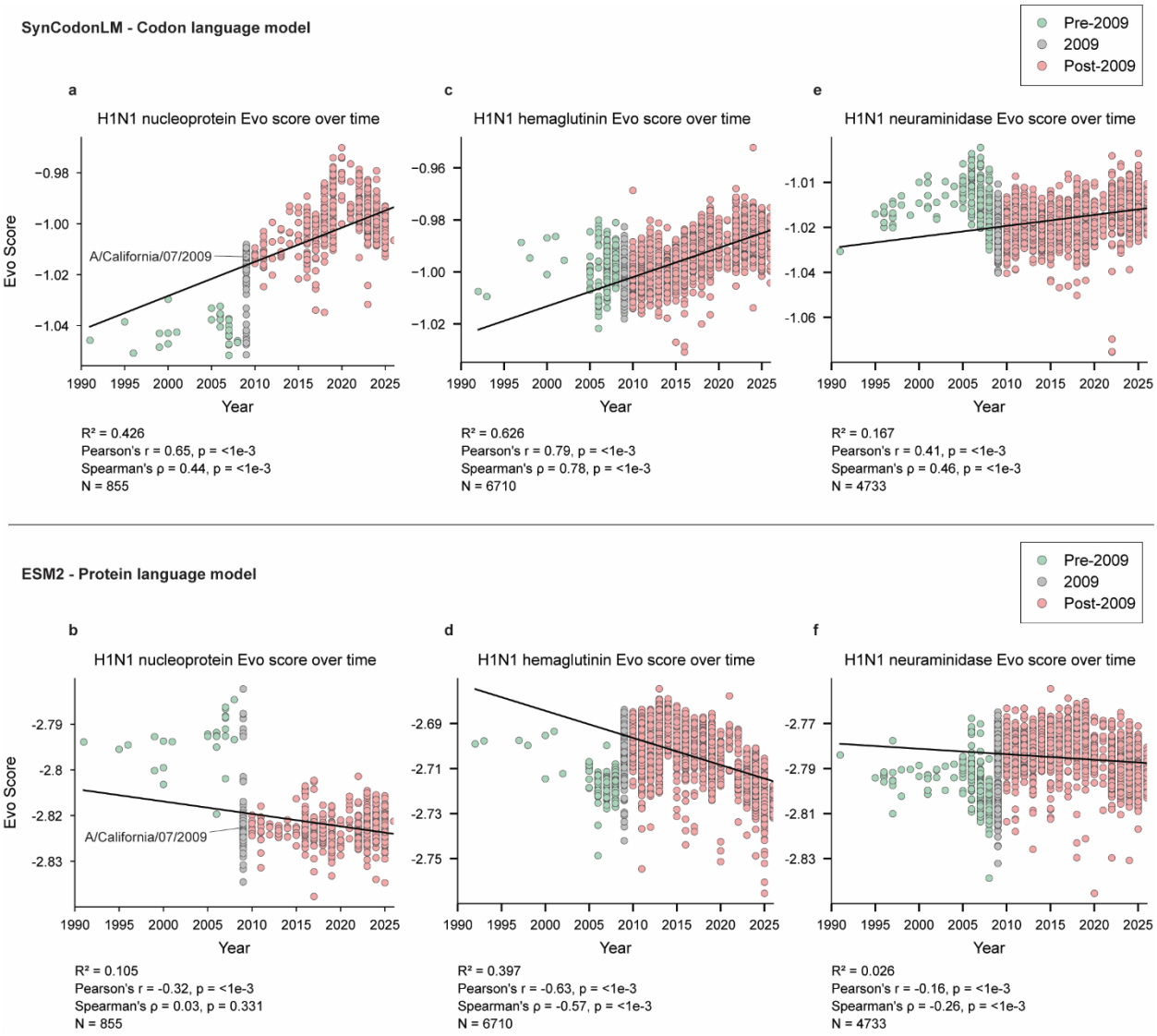


Fig. 2: Codon-level likelihood tracks forward viral evolution, whereas protein-level likelihood trends in the opposite direction. (a, c, e) SynCodonLM Evo scores plotted against sampling year for Influenza A H1N1 nucleoprotein (a), hemagglutinin (c), and neuraminidase (e) coding sequences. Scores increase significantly with time in all three proteins (Pearson r = 0.65, 0.79, and 0.41, respectively; all $p < 1 \times 10^{-3}$), indicating that later-emerging sequences are assigned higher model likelihood. Nucleoprotein shows a pronounced shift, coincident with the 2009 pandemic strain A/California/07/2009, sustained in subsequent years. (b, d, f) ESM-2 pseudo-log-likelihoods on the matched translated sequences show the opposite trend (r = −0.32,

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

−0.63, and −0.16; all $p < 1 \times 10^{-3}$), decreasing over time. All sequences correspond to unique viral isolates with non-synonymous diversity; purely synonymous variants were excluded. No task-specific fine-tuning was applied to either model.

In contrast, ESM-2 exhibited the opposite temporal behavior on the matched translated sequences (r = −0.16 to −0.63; all $p <$ 1e-3; **Fig. 2b, d, f**).

While absolute score scales are not directly comparable due to differences in tokenization and training objectives, the directionality of these trends is striking; codon-level modeling captures a forward-in-time signal that is not recoverable from the matched protein sequences, where ESM-2 likelihood instead trends opposite to sampling year.

Importantly, all sequences analyzed corresponded to distinct viral isolates with non-synonymous diversity, with purely synonymous variants excluded to prevent artificial inflation of SynCodonLM performance relative to protein language models. Despite this conservative design, SynCodonLM systematically assigned higher likelihood to sequences that emerged later in viral evolution.

These findings demonstrate that a simple, unsupervised codon-level likelihood score contains substantial predictive information about real-world viral evolution, indicating that synonymous codon usage encodes context-dependent biological constraints learnable by a language model. If these learned constraints reflect genuine functional information rather than an artifact of natural sequence statistics, then optimizing sequences toward higher codon-level likelihood should yield measurable biological benefits. Therefore, we next ask whether SynCodonLM-guided synonymous edits improve recombinant protein expression.

A language-model-guided framework for directed evolution in codon space

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We developed a language-model-guided framework for directed evolution in synonymous codon space, treating clinically validated coding sequences as high-fitness parental baselines and iteratively refining them while preserving the encoded protein.

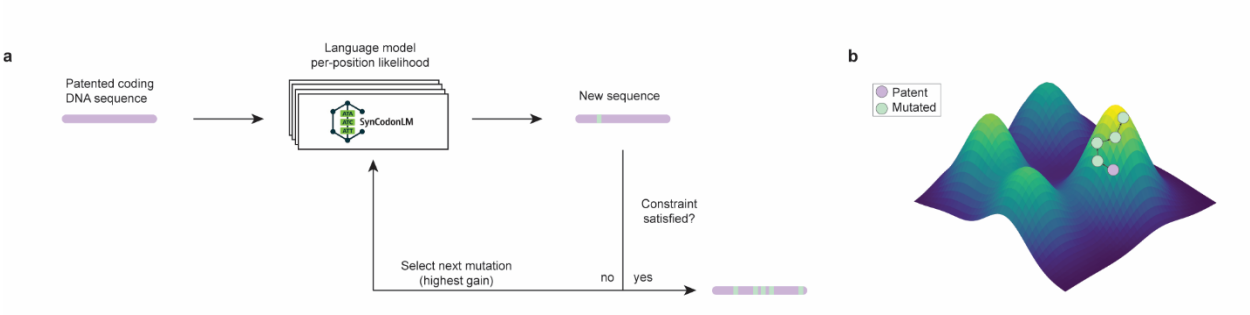


Fig. 3: A language-model-guided framework for directed evolution in codon space. (a) Patented antibody coding sequences are passed to SynCodonLM[24], and the substitution predicted to most increase local likelihood is applied at each iteration. The newly generated sequence is passed back into the model for finding the next most preferable mutation, up to a set limit of mutations or when the sequence is at a maxima. (b) Directed evolution drives the sequence up a local fitness gradient in codon space, defined as Evo score, while preserving the encoded protein.

Directed evolution proceeded by repeatedly selecting the substitution predicted to most improve local sequence compatibility, updating the context, and re-scoring before the next edit (**Fig. 3a**). This allowed successive substitutions to interact through changing context, traversing codon space toward higher-fitness configurations without altering the encoded protein (**Fig. 3b**).

To test this framework, we assembled a panel of 23 patented, clinically deployed antibody-based therapeutics and treated their disclosed coding sequences as high-fitness parental baselines. The panel was deliberately constructed to span industrially relevant format and isotype diversity, including conventional monospecific IgGs, bispecific architectures, an antibody–drug conjugate, antibody fusion proteins, and molecules with significant Fc engineering (**Supplementary Table**

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

1). SynCodonLM-guided refinement was capped at 10% of codon positions per sequence, limiting divergence from the clinically validated parent while allowing sufficient synonymous change to test whether local codon-space refinement could produce measurable effects.

SynCodonLM-refined sequences increase recombinant antibody titers

We next tested whether iterative synonymous refinement produced measurable changes in recombinant antibody expression. Parental (patented) and SynCodonLM-refined coding sequences were cloned into pTT5-based mammalian expression vectors[31] with a fixed signal peptide and stop codon and transiently expressed in CHO-3E7[32] cultures in biological triplicate. After six days, culture supernatants were harvested for titer quantification, and expressed antibodies were purified for downstream analytical characterization.

Importantly, across the panel of 23 clinical-stage antibody-based therapeutics, SynCodonLM-refined sequences increased secreted antibody titer relative to their parental patented counterparts in 17 of 23 cases (74% responder rate, 15 individually significant, paired t-test $p < 0.05$), corresponding to a mean (geometric) 1.34-fold increase (**Fig. 4a**; paired t-test $p < 0.001$; Wilcoxon signed-rank $p < 0.001$). The magnitude of positive responses substantially exceeded that of the few negative ones (**Fig. 4a**). Raw titers are provided in **Supplementary Fig. 4**.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

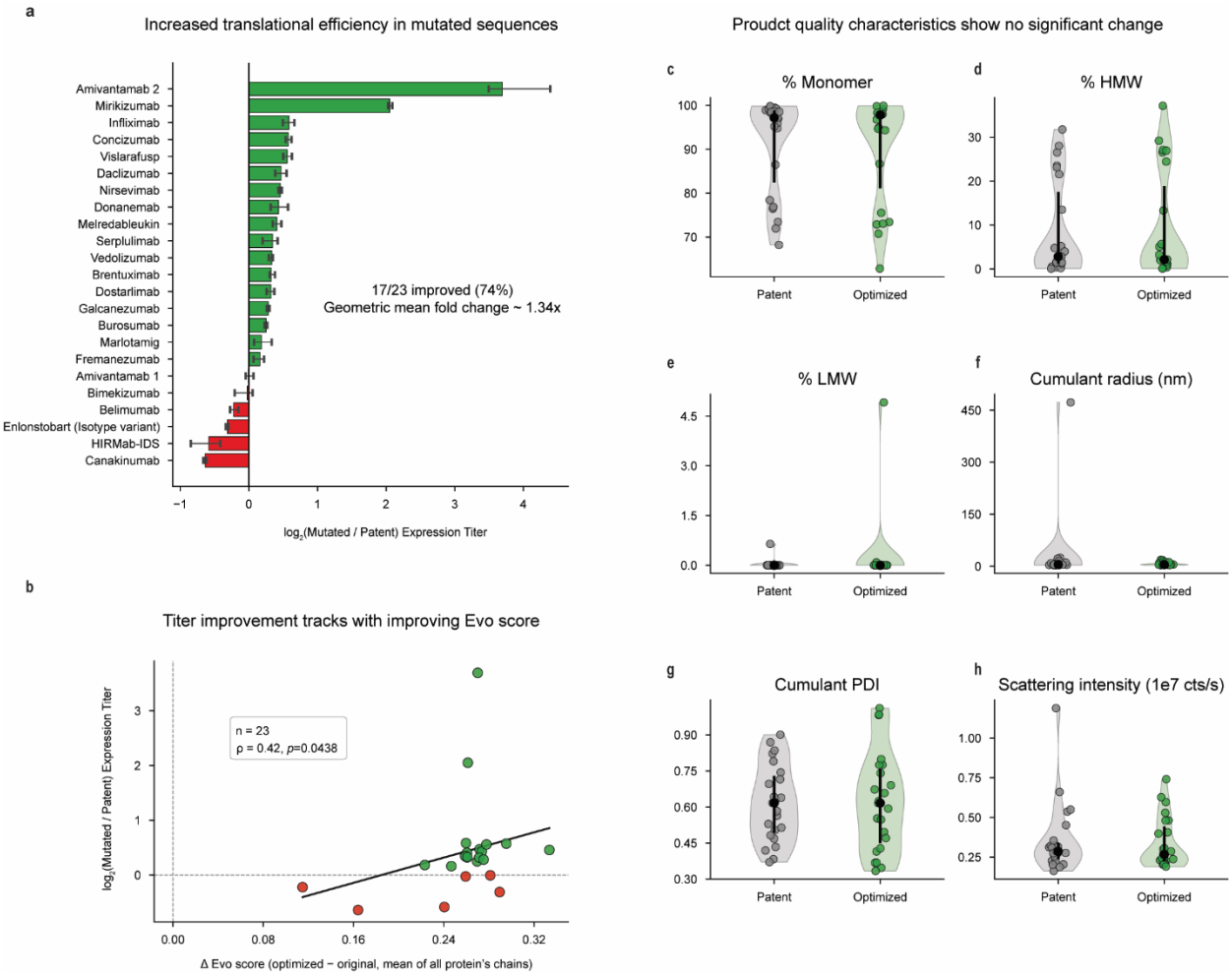


Fig. 4: Clinically optimized therapeutic coding sequences remain improvable without compromising product quality. (a) Per-molecule $\log_2$-fold change in secreted CHO titer for SynCodonLM-refined sequences relative to patented parental counterparts across 23 clinical-stage antibody-based therapeutics (mean ± SD, biological triplicates). Green, improved constructs (17/23; 74% responder rate, 15 individually significant by paired t-test, $p < 0.05$); red, reduced-titer constructs. Panel-wide mean (geometric) fold increase ~1.34 (paired t-test $p <$ 0.001; Wilcoxon signed-rank $p < 0.001$). (b) $\Delta\log_2$(titer) versus Δ Evo score (refined − parental, averaged across expressed chains); Spearman's rank correlation coefficient (ρ) = 0.42, $p =$ 0.0438, n = 23. All 23 constructs show positive Δ Evo score, confirming that refinement

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

systematically raised model likelihood. (c–e) Analytical size-exclusion chromatography (aSEC) of Protein A–purified material: percent monomer (c), high-molecular-weight species (%HMW, d), and low-molecular-weight species (%LMW, e), for parental (gray) and refined (green) groups; violin plots with median lines. Distributions show no significant difference between groups (paired Wilcoxon signed-rank, $p > 0.05$ for all measures). (f–h) Dynamic light scattering (DLS) of the same material: cumulant hydrodynamic radius (f), polydispersity index (g), and scattering intensity (h). Again, distributions are closely matched between groups, indicating no detectable change in aggregation propensity or homogeneity (paired Wilcoxon signed-rank, $p > 0.05$).

To ask whether these titer changes tracked with the model's own compatibility signal, we compared the change in mean chain-level Evo score (ΔEvo = refined − parental, averaged across all expressed chains of a given molecule) with the observed $\log_2$-fold change in titer (**Fig. 4b**). Across the panel, Δ Evo score was significantly positively correlated with titer improvement ($\rho = 0.42$, $p = 0.0438$, n = 23), and all 23 constructs, including the negative responders, showed positive ΔEvo values, confirming that iterative refinement systematically raised codon-language-model likelihood as intended. This relationship indicates that the magnitude of model-implied improvement provides genuine, quantitative information about the likelihood of a titer gain. Chain-resolved correlations were stronger for heavy chains (**Supplementary Fig. 5**), suggesting that heavy-chain optimization may contribute disproportionately to expression gains under the transient expression conditions used here.

Because synonymous changes can affect co-translational folding[5] and post-translational modifications[7], we confirmed product identity by intact LC-MS on Protein A–purified material. Each pair of molecules (parental and refined) was evaluated by LC-MS at both the intact and

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

subunit levels to (1) verify sequences and complexation, and (2) gauge the impact(s) on the distribution of abundant PTMs (ex., glycosylation, C-terminal lysine truncation) that comprise the majority of observed proteoforms.

For all molecule pairs, intact and subunit molecular weights matched one another at the global level (per molecule conditions, **Supplementary Table 2**; spectra and summary, **Supplementary Data)**, ruling out coincidental modifications or subtle sequence deviations as sources of the observed titer differences. Additionally, PTM distributions were nearly identical within each pair, but dissimilar across the different molecules, indicating that the amino acid sequence (and not codon differences within the nucleotide sequences) is the major driver for proteoform distribution.

Consistent with this preserved molecular identity, biophysical characterization of the purified material showed no significant adverse shifts in product-quality attributes. Analytical size-exclusion chromatography (aSEC) revealed comparable distributions of monomer content (**Fig. 4c**), high-molecular-weight species (**Fig. 4d**), and low-molecular-weight species (**Fig. 4e**) between parental and refined groups. Complementary dynamic light scattering (DLS) analysis showed comparable distributions of cumulant hydrodynamic radius (**Fig. 4f**), polydispersity index (**Fig. 4g**), and scattering intensity (**Fig. 4h**). Across all six attributes, no significant difference was detected between parental and refined sequences (two-sided Wilcoxon signed-rank test, all $p > 0.05$).

These results demonstrate that SynCodonLM-guided refinement systematically improves CHO titer, with improvement magnitude tracking the model's likelihood signal, while preserving product identity (including native PTMs) and biophysical quality. More broadly, they show that clinically optimized therapeutic coding sequences remain evolvable despite extensive prior

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

development, supporting iterative refinement as a practical strategy for engineering coding DNA independently of protein sequence.

Mechanistic insight into differential expression

To probe the mechanistic basis of the observed titer improvements, we asked (i) whether SynCodonLM-guided refinement systematically shifted established codon-usage descriptors in directions heuristically associated with higher translational efficiency, and (ii) whether the SynCodonLM Evo score predicted per-construct titer improvements more effectively than classical single-metric heuristics.

We first evaluated the direction and significance of change for a broad panel of >23 established codon-usage descriptors (full list and references in Methods)[33] computed on parental and refined sequence pairs. For each metric, the paired difference (Δ = refined − parental) was tested with a Wilcoxon signed-rank test, and each metric was annotated with a biologically expected direction for improved translational efficiency drawn from prior literature. Of the metrics with a defined directional expectation, the overwhelming majority shifted in the favorable direction (**Fig. 5a**), including CAI[16], RSCU[34], ENc[35–37], GC content[38], CpG and UpA O/E enrichment[39,40], CBI[41], and FOP[42] (all $p < 0.001$); no directional descriptor shifted significantly in the opposing direction. Despite optimizing purely for codon-language-model likelihood, with no explicit metric-based objective, SynCodonLM refinement recovered the directional expectations of canonical codon-optimization heuristics across all descriptors at once.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

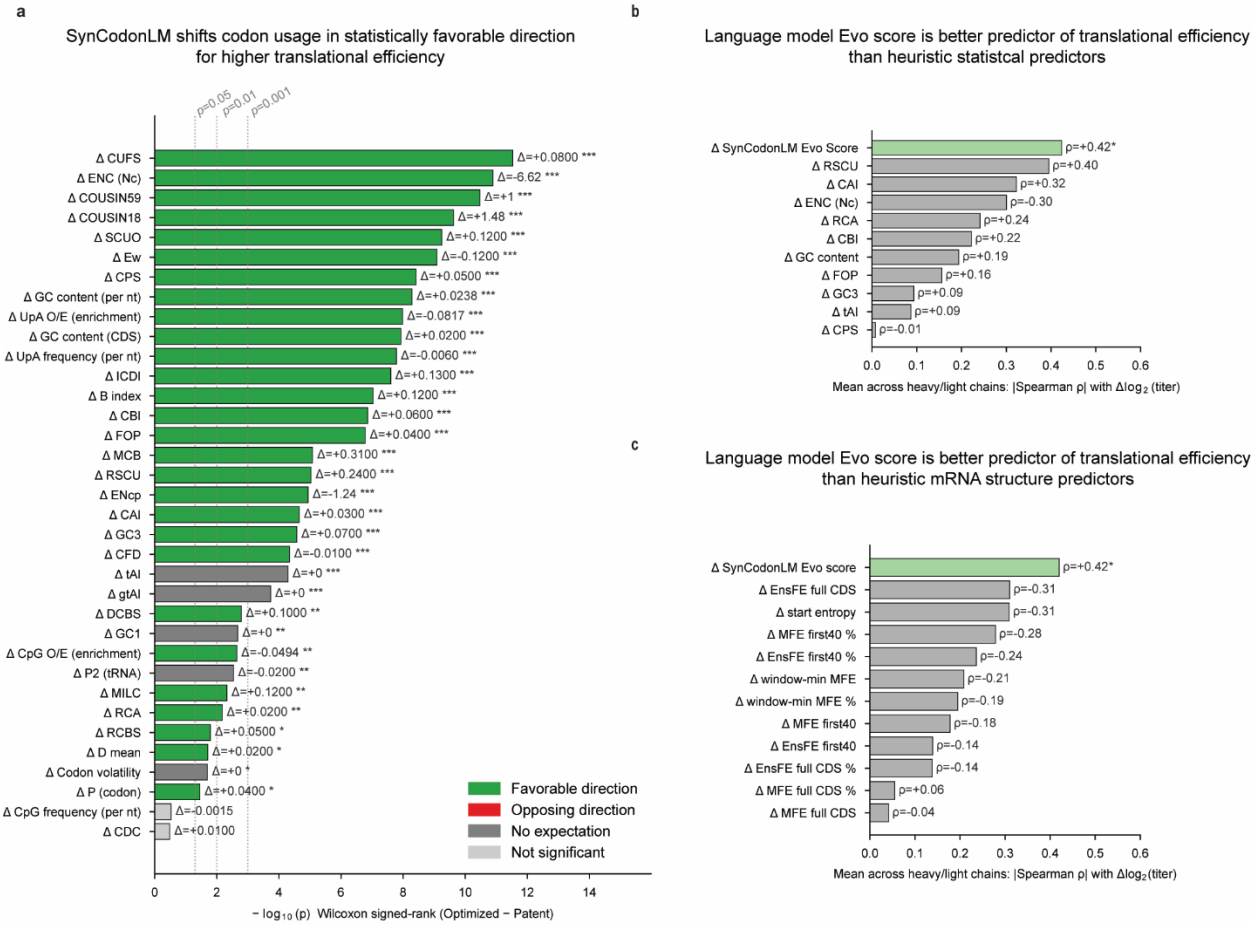


Fig. 5: Codon language model likelihood outperforms classical codon-usage and mRNA-structure heuristics. (a) Paired Wilcoxon signed-rank comparison of established codon-usage descriptors on parental versus SynCodonLM-refined sequences (n = 23). The x-axis reports $-\log_{10}(p)$; dashed guides mark $p = 0.05$, 0.01, and 0.001. Bars are annotated with the paired difference (refined − parental) and colored by the expected direction for improved translational efficiency: green, favorable shift; red, opposing shift; dark gray, no defined expectation; light gray, not significant. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. (b) Mean absolute Spearman's rank correlation coefficient (|ρ|, averaged across heavy and light chains) between per-descriptor change and $\Delta\log_2$(titer). Signed ρ values are annotated on each bar. The SynCodonLM Evo score (pastel green) is the strongest single predictor ($\rho = +0.42$, $p < 0.05$), exceeding all classical

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

codon-usage heuristics (gray). (c) Analogous analysis using ViennaRNA-derived mRNA-structure descriptors. SynCodonLM Evo score again ranks first.

We next asked whether these classical descriptors were quantitatively predictive of the observed titer improvements. For each construct, we computed the mean absolute Spearman's rank correlation across heavy and light chains between the per-metric change and the observed $\Delta\log_2$(titer) (**Fig. 5b).** The change in SynCodonLM Evo score was the strongest single predictor ($\rho = +0.42$, $p < 0.05$), outperforming every classical codon-usage descriptor evaluated ($|\rho| \leq 0.40$; **Fig. 5b**). Notably, several widely used metrics (CAI, tAI[43], GC3, FOP) showed weak correlations with translational efficiency, underscoring that individual heuristic descriptors capture only limited information about expression outcomes, even when they respond systematically in aggregate. A graph including all the statistical descriptors and their correlation with $\Delta\log_2$(titer) is available in **Supplementary Figure 6**. A nonparametric bootstrap (5,000 iterations; **Supplementary Fig. 7**) yielded broad confidence intervals for all descriptors, with substantial overlap among metrics, although the SynCodonLM Evo score retained the highest point estimate.

Because expression can also be shaped by mRNA secondary-structure features around the ribosome-loading region, a signal established by Kudla et al. in *E. coli*[44], we performed a similar analysis using a panel of ViennaRNA-derived[45] mRNA-structure descriptors (**Fig. 5c**; full descriptor definitions in Methods). The SynCodonLM Evo score again emerged as the strongest predictor of $\Delta\log_2$(titer) ($\rho = +0.42$, $p < 0.05$). Among the mRNA-structure descriptors, ensemble free-energy and sequence-entropy metrics showed the strongest associations with expression ($|\rho|$ up to 0.31), whereas the canonical 5′ folding-window descriptors like that of Kudla et al.,

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

exhibited weaker correlations ($|\rho| \leq 0.28$). Correlation strengths varied depending on whether the signal peptide was included in the folding calculation (**Supplementary Fig. 8**).

We further asked whether SynCodonLM's suggested substitutions were concentrated at specific positions along the CDS, for example enriched near the 5′ ribosome-loading region where structure-mediated effects on translation initiation and early elongation are strongest[46,47]. We observed a clear positional bias, with edits enriched toward the 5′ end of the coding sequence and declining toward the 3′ end (**Supplementary Fig. 9**). To test the classical hypothesis that synonymous codon choice is coordinated with local protein structure[5,48], we then examined editing rates relative to secondary-structure elements while controlling for codon degeneracy throughout each sequence. Codons at the start (N-cap) of β-sheet strands were significantly less likely to be edited than expected, consistent with the idea that codon usage at some structural boundaries is under selective constraint (**Supplementary Fig. 10**).

These results point to a non-obvious conclusion: therapeutic coding sequences retain accessible optimization potential beyond what can be captured by any single classical codon-usage or mRNA-structure heuristic. Although our iterative refinement strategy systematically moves sequences in the favorable direction of nearly all canonical codon-usage descriptors, no single descriptor predicts per-construct titer gains, whereas codon-language-model likelihood does. This suggests that residual coding-sequence optimization potential is distributed across many context-dependent features that classical heuristics capture only in isolation.

Discussion

We reframed codon optimization as an iterative, model-guided directed-evolution process operating strictly within synonymous codon space. Applied to the patented coding DNA of 23

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

clinically deployed antibody-based therapeutics, sequences already refined through years of industrial development, SynCodonLM improved CHO titer in 17 of 23 cases (~1.34-fold change on average) with no significant degradation of monomer content, colloidal behavior or post-translational modification. The central result is that clinically validated coding sequences remain evolvable in synonymous codon space.

This parallels modern protein engineering, where protein language models improve even highly mature antibodies through small numbers of localized edits[23]; our results extend that logic to the synonymous layer of the genetic code. Practically, this reframes codon-optimized constructs not as finished artifacts locked in at DNA synthesis, but as starting points that remain open to further model-guided refinement, where small synonymous edits accumulate into meaningful expression gains without altering the encoded protein.

The mechanistic analysis in **Fig. 5** clarifies why language-model-guided refinement outperforms static, single-metric heuristics. In aggregate, SynCodonLM shifted nearly every classical codon-usage descriptor (CAI, RSCU, CBI, FOP, ENc, GC content, and CpG and UpA enrichment) toward values associated with efficient translation, despite being trained only on a synonym-constrained masked-language-modeling objective, with no explicit knowledge of any of these metrics. This convergence indicates that codon-language-model likelihood implicitly integrates the same statistical regularities that decades of classical codon-optimization work have each captured in isolation. Consistent with this, the SynCodonLM Evo score showed the strongest per-construct association with relative expression ($\rho = 0.42$, $p < 0.05$), with the classical and mRNA-structure descriptors tested here reaching comparable but somewhat weaker associations ($|\rho| \leq 0.40$). Together, these observations suggest that the advantage of a codon language model is

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

compositional: rather than relying on any single heuristic, it captures a distributed, context-dependent combination of features that each classical descriptor represents only partially.

That codon-level likelihood improves expression is consistent with our finding that it also encodes genuine evolutionary signal. On an Influenza A H1N1 benchmark spanning nucleoprotein, hemagglutinin, and neuraminidase, SynCodonLM Evo scores increased significantly with time ($r = 0.65, 0.79, 0.41$; all $p < 1e\text{-}3$), whereas ESM-2[49] scores on matched translated sequences showed the opposite trend. This grounds the refinement results: the same likelihood that tracks real viral evolution is what drives the observed titer gains.

Several limitations warrant acknowledgment. First, our data derive from small-scale transient CHO cultures; although transient titers have been shown in previous studies to preserve rank-order in stable clones[50], translation to stable-pool or clonal production settings remains to be tested. Second, refinement was capped at 10% of codon positions per construct, leaving open whether deeper traversal or non-greedy acquisition strategies would yield further gains or introduce product-quality trade-offs. Third, our analytical package (intact LC-MS, SEC, DLS) detected no significant shifts in molecular weight, glycoform envelope, monomer content, aggregation, or colloidal behavior, but did not include manufacturing-depth characterization (released N-glycan mapping, icIEF/CEX charge-variant analysis, or site-specific peptide mapping). Fourth, the mechanistic analysis in **Fig. 5** is correlative and panel-level; it does not isolate which mechanistic axis is causal in any given molecule. Fifth, the viral evolution benchmark is subject to known geographic and temporal sampling biases and should be interpreted as evidence that codon-level likelihood tracks evolutionary progression, not as a claim that codon-level fitness causally shapes viral evolution.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Looking forward, the framework introduced here suggests several natural extensions. Iterative synonymous refinement is model-agnostic at the framework level: as codon language models continue to improve, through larger context windows, multi-species conditioning, or objectives explicitly coupled to downstream fitness signals, the same directed-evolution loop can be plugged in without changing its structure. Coupling refinement to closed-loop experimental feedback, in which measured titers or product-quality attributes inform the next round of edits, is a particularly natural next step and would convert the framework from a single-shot in silico refinement into a genuine active-learning cycle in coding DNA space. Beyond therapeutic antibodies, the same approach should generalize to any recombinantly expressed protein where the coding sequence is a controllable design axis, including vaccine antigens, gene-therapy payloads, and other biologics classes where synonymous choice has been shown to matter but has historically been treated as static.

More broadly, our results argue that the boundary between protein-level and codon-level engineering deserves to be revisited. Protein language models have unlocked iterative, evolutionarily plausible improvement at the amino-acid level[23]; codon language models now appear to offer an analogous capability at the synonymous layer, without any of the regulatory or immunogenicity considerations that accompany changes to the protein sequence itself. Treating already-optimized coding sequences as high-fitness starting points for further iterative refinement, rather than as design endpoints, offers a simple, low-risk, and experimentally validated route to improving clinical biologics, and reframes the coding layer of the genetic code as an actively engineerable, not merely selectable, design space.

Methods

High-fitness coding sequence baseline (clinical-stage reference set)

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A high-confidence benchmark of pre-optimized antibody coding DNA sequences was curated by starting from Thera-SAbDab[51], which tracks WHO-recognized antibody- and nanobody-derived therapeutics and provides associated sequence and metadata. As a first pass, each therapeutic's variable (V) region was used to query the patent nucleotide collection with NCBI BLAST[52] (blastn against nt), retaining only patent-disclosed coding DNA sequences that returned 100% identity across the V region and therefore encoded the same V region as the clinical molecule. For each therapeutic entry, primary patent documents and sequence listings were mined using automated document-review workflows to determine whether full-length coding DNA sequences were disclosed for each expressed chain of the final drug molecule and whether those sequences were described in a CHO expression context. Candidate chains were then cross-mapped to publicly available substance registration resources in the Global Substance Registration System (GSRS)[53] to corroborate identity and chain assignment through sequence-based checks aligned with regulated substance definitions and sequence search capabilities. The resulting collection served as a high-fitness baseline of coding sequences that were plausibly already optimized in an industrial therapeutic development context and therefore provided an empirically grounded starting point for in silico refinement.

In total, 23 clinical antibody-based therapeutics were gathered (comprising 22 distinct therapeutics corresponding to 23 expressed constructs, as the amivantamab bispecific was produced and analyzed as its two half-antibodies), including: monoclonal antibodies spanning IgG1, IgG2 and IgG4 heavy chain isotypes as well as both kappa and lambda light chain isotypes. The heavy-chain isotype distribution was dominated by IgG1 (n = 16) and IgG4 (n = 6), with a single IgG2 representative (fremanezumab), while light-chain use was almost exclusively kappa (n = 22) with lambda represented by a single chain (belimumab); isotype pairings

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

therefore break down as IgG1κ (n = 15), IgG1λ (n = 1), IgG2κ (n = 1), and IgG4κ (n = 6). Nineteen chains were derived from approved medicines, two from Phase II candidates, and two from a Phase I/ candidate (**Supplementary Table 1**). Beyond conventional monospecific IgG monoclonal antibodies (n = 18), the reference set was deliberately enriched for architecturally complex, industrially relevant formats whose coding sequences must nonetheless be engineered for high-fitness expression in CHO: one antibody–drug conjugate (brentuximab vedotin, a chimeric IgG1κ site-conjugated on reduced interchain cysteines via a cathepsin-cleavable valine-citrulline linker to monomethyl auristatin E at a drug-to-antibody ratio of ≈ 4)[54]; two bispecific IgG molecules: amivantamab (a Genmab DuoBody IgG1κ target-bispecific against EGFR × MET, assembled from two half-antibodies bearing complementary F405L/K409R Fc mutations that undergo controlled Fab-arm exchange[55] and produced in low-fucose CHO to enhance ADCC), marlotamig (a Regeneron Veloci-Bi IgG4 with a common kappa light chain and a "chimeric star-Fc" heavy-chain mutation on one arm that ablates Protein A binding to enforce heterodimer-selective purification[56]); and three antibody fusion proteins: melredableukin alfa (a non-targeting IgG1κ scaffold C-terminally fused to an engineered CD25-silent IL-2 variant, presented monovalently via knob-into-hole Fc pairing[57]), AGT-182 (a HIRMab-IDS fusion: an anti-human-insulin-receptor IgG1κ blood–brain-barrier shuttle with iduronate-2-sulfatase fused to the heavy-chain C-terminus for CNS enzyme delivery in MPS II), and vislarafusp alfa (an anti-HER2 IgG1κ with the SIRPα D1 domain appended to the light-chain N-terminus via a GS linker, acting as a locally-delivered CD47 decoy). Several molecules additionally carry defined Fc engineering, including the YTE (M252Y/S254T/T256E) FcRn-affinity mutation that extends serum half-life in nirsevimab[58] and the IgG4 hinge-stabilizing S228P substitution that prevents natural IgG4 Fab-arm exchange present across the IgG4 mAbs[59]. Full per-molecule annotation,

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

format assignments, targets, and the primary patent identifier used to source each coding sequence are provided in **Supplementary Table 1**.

For 5 of the 23 constructs, the coding sequence was obtained from a patent sequence listing assigned to an entity other than the originator of the marketed therapeutic; in each of these cases the disclosed sequence nonetheless encodes the corresponding therapeutic protein, consistent with the sequence-based identity checks described above.

In a small number of constructs, the patent-disclosed coding DNA did not translate to the exact clinical protein sequence at every position. In some of these cases the patent text identified the disclosed sequence as encoding the final therapeutic protein, while the deposited coding DNA translated to a small difference from that protein, consistent with isolated discrepancies within the sequence listings themselves. First, the belimumab and brentuximab coding sequences each carried a C-terminal lysine on the heavy chain that is absent from the final molecule listed in GSRS; C-terminal lysines are, however, routinely clipped from mammalian expression products and are therefore not expected to affect the expressed protein. The bimekizumab coding sequence contained a single-residue difference in the Fc region of its heavy chain at an atypical position. The amivantamab half-antibody 1 light-chain coding sequence was missing the N-terminal alanine. Lastly, the enlonstobart coding sequence encoded an IgG1 backbone rather than the IgG4 molecule listed in GSRS, although its V region and full light chain matched the final drug molecule. In each case the discrepancy was individually verified against the GSRS/INN reference and the coding sequence retained for downstream analysis.

SynCodonLM inference details

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

SynCodonLM was loaded from the Version 2 model checkpoint on HuggingFace[24,60] onto a Nvidia L40S GPU running CUDA 13.0 for all evaluations in this work. Environment versioning is available on the corresponding GitHub codebase released with this work.

SynCodonLM-based evolutionary scoring and greedy synonymous directed evolution

Synonymous codon refinement was performed with SynCodonLM[24], a codon language model trained under a biologically grounded constraint in which masked codons were predicted only from synonymous options consistent with the known amino-acid sequence, thereby disentangling codon-level patterns from protein-level semantics. Each coding sequence was represented as a sequence of codon tokens x = $(x_1, \ldots, x_L)$, where L denotes the number of codons. To quantify "evolutionary" compatibility of a complete sequence using a masked language model, a pseudo-log-likelihood (PLL) score was computed by iteratively masking one position at a time and accumulating the log probability assigned to the true token at that position given the remaining unmasked context. This single-mask PLL scoring paradigm is a standard approach for scoring sequences with masked language models in the absence of an explicit left-to-right factorization. Specifically, for position i, the model produced logits $z_i(\nu)$ over vocabulary tokens $\nu \in V$ when x_i was replaced by a mask token, and the per-position log probability of the true codon token was computed as

Equation 1:

$$\log p\left(x_i \mid x_{\setminus i}\right) = log\ softmax(z_i)_{x_i} = z_i(x_i) - \log\left(\sum_{\nu \in V} e^{z_i(\nu)}\right),$$

where $x_{\setminus i}$ denotes the masked sequence with all positions except i left intact. The sequence-level evolutionary score was reported as the mean PLL across codon positions (in nats),

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Equation 2:

$$PLL_{mean}(x) = \frac{1}{L}\sum_{i=1}^{L} \log p\left(x_i \mid x_{\setminus i}\right),$$

with higher values indicating greater model compatibility.

To propose synonymous edits, scoring at each position was restricted to the set of synonymous codons for the amino acid a_i encoded at position i. Let $\mathrm{S}(a_i)$ denote the synonymous codon set at that position. Scores were normalized over the synonym set through a synonym-restricted softmax, rather than the full-vocabulary softmax, such that

Equation 3:

$$p_{syn}\left(c \mid x_{\setminus i}\right) = \frac{e^{z_i(c)}}{\sum_{c' \in \mathrm{S}(a_i)} e^{z_i(c')}} \quad for\ c \in \mathrm{S}(a_i).$$

This ensured that positions with different synonym degeneracies were compared within their appropriate synonymous choice sets, mirroring the normalization implemented by restricting logits to the synonymous token IDs at each step. Greedy optimization proceeded by iteratively selecting a single synonymous substitution that maximized a head-to-head Δ-margin at one position. For a candidate position i, with current codon $c_{curr} = x_i$ and best alternative synonym $c_{best} \in \mathrm{S}(a_i) \setminus (c_{curr})$, the step score was defined on the logit scale as

Equation 4:

$$\Delta_i = z_i(c_{best}) - z_i(c_{curr}).$$

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Because both terms share the same synonym-restricted normalization constant, Δ_i was equivalently the difference in synonym-normalized log probabilities,

Equation 5:

$$\Delta_i = \log p_{syn}\left(c_{best} \mid x_{\setminus i}\right) - \log p_{syn}\left(c_{curr} \mid x_{\setminus i}\right),$$

and the corresponding odds ratio reported during optimization was

Equation 6:

$$OR_i = \exp(\Delta_i) = \frac{p_{syn}\left(c_{best} \mid x_{\setminus i}\right)}{p_{syn}\left(c_{curr} \mid x_{\setminus i}\right)}.$$

At each iteration, the single position and substitution with the largest admissible Δ_i were applied, the sequence context was updated, and the process was repeated until no admissible Δ remained above threshold.

To constrain divergence from clinically validated parental sequences, the number of codon positions altered relative to the original sequence was capped at ⌊0.10L⌋ (10% of codon positions), with optional tightening of Δ thresholds once the cap was reached to avoid negligible micro-steps. Start and stop boundaries were preserved when configured (retaining an initiating ATG and terminal stop codon), and short-cycle oscillations were mitigated by immediate reversal checks, state-hash cycle detection, and a small tabu memory over recently edited positions. Species conditioning was applied via a species token type, enabling host-contextualized scoring under the model's token-type mechanism when supported.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Cloning, gene synthesis and construct assembly

Parental and SynCodonLM-refined coding sequences were synthesized as single linear DNA fragments (Twist Bioscience) with flanking BsaI recognition sites compatible with Type IIS Golden Gate assembly. Each construct was assembled into a pTT5-based mammalian expression vector[31] containing a fixed signal peptide and stop codon using a BsaI-based Golden Gate cloning kit (New England Biolabs) in a single one-pot reaction per construct, following the manufacturer's recommended thermocycling protocol. Assembly reactions were transformed into chemically competent E. coli, plated on selective medium, and individual colonies were picked for downstream screening.

Colonies were screened by colony PCR using primers flanking the insertion site, and PCR amplicons spanning the full open reading frame (ORF) were verified by Illumina MiSeq short-read sequencing. Each construct was considered sequence-confirmed only upon full-ORF concordance with the intended parental or refined coding sequence, including all synonymous positions edited by SynCodonLM. Sequencing was carried out iteratively until a colony passing full-ORF verification was recovered for every construct.

Sequence-confirmed colonies were then scaled up in selective liquid culture, and endotoxin-reduced plasmid DNA was prepared using the QIAGEN Plasmid Plus midiprep kit (QIAGEN) according to the manufacturer's protocol. Purified plasmid preparations were quantified by UV absorbance (A_{260}), assessed for purity by A_{260}/A_{280} ratio, and used directly for transient CHO transfections.

CHO expression and antibody purification

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Transient expression was performed in CHO-3E7 cells[32], a platform described to enhance recombinant protein production when paired with oriP-bearing pTT vectors and optimized CMV-based expression cassettes.

Cells were maintained in Irvine Scientific BalanCD CHO medium and seeded at a density of 4 × 10^6 cells/mL with viability >99% at the time of transfection. Transfections were carried out using TransIT-PRO (Mirus Bio) in OptiPRO SFM (Thermo Fisher Scientific) according to the manufacturer's recommendations. Heavy- and light-chain expression plasmids were co-transfected at a 1:2 mass ratio (heavy:light), a commonly used configuration intended to promote efficient antibody assembly and secretion[61]. Approximately 16–24 hours post-transfection, cultures were supplemented with CHO Feed B (Irvine Scientific) and Anti-Clumping Agent (Thermo Fisher Scientific) to support cell health and productivity. Temperature was subsequently shifted from 37°C to 30 °C to enhance protein expression and stability during the production phase. Cultures grew for a total of 6 days in a culture volume of 4mL within 24-well plates.

Antibodies were purified from culture supernatants using an AssayMap Bravo platform (Agilent Technologies) equipped with Protein A affinity tips (Agilent Technologies). Purified antibodies were eluted and formulated into 75 mM citrate, 250 mM sodium phosphate, pH 5.4.

Antibody titer quantification using Protein A affinity chromatograph on supernatant

The titer determination experiments were performed using an Agilent Bio-Monolith rProtein A column (4.95 × 5.2 mm; Agilent Technologies) on an Agilent 1290 Infinity II Bio LC system. The column temperature was maintained at 25°C. Detection was carried out using a fluorescence detector set at an excitation wavelength of 280 nm and an emission wavelength of 350 nm. Prior to analysis, cell culture supernatants were subjected to centrifugation at 3200 × g for 20 min,

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

followed by filtration through a 0.2 µm filter plate. Samples were subsequently stored at 4 °C in the autosampler, and an injection volume of 20 µL was used for each run.

Mobile phase A comprised 50 mM sodium phosphate at pH 7.4, and mobile phase B was 100 mM citric acid at pH 2.6, both prepared in Milli-Q water using 1M stock solutions. Chromatographic separation was performed at a flow rate of 1 mL/min using the following gradient: initial equilibration with 100% mobile phase A for 0.8 min, followed by a linear gradient from 0% to 100% mobile phase B from 0.8 to 2.2 min to elute bound antibodies. The column was then re-equilibrated with 100% mobile phase A from 2.2 until the end of a 4.0-min run time. Data acquisition and analysis were performed using Agilent OpenLab software. Antibody titers were quantified by peak area integration against a Rituximab standard curve ranging from 10 to 1000 µg/ml.

Mass spectrometry analyses

Detailed LC-MS parameters are provided in **Supplementary Table 2**. Briefly, intact and subunit level analyses were performed using a 6545XT Q-ToF interfaced with a 1290 Infinity II UHPLC (Agilent Technologies, Santa Clara, CA). Ballistic gradient conditions were used to desalt samples over a reversed-phase column (Poroshell SB 300 C3, 1.0 x 30 mm, 5 µm, 300Å, Agilent Technologies, Santa Clara, CA) prior to introduction into the mass spectrometer. Raw spectra were extracted and deconvoluted using MassHunter v11 software to enable (1) verification of predicted molecular weight values for each molecule and (2) pairwise comparison of the post-translational modifications within each molecule pair (parental and refined) at both the complexed (intact) and reduced subunit levels.

Codon-usage descriptor calculations

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Codon usage descriptors were calculated using the GenScript Rare Codon Analysis Tool[33], utilizing the NCBI Genome Resource[62] reference source, for *Cricetulus griseus* (CHO). Descriptors were computed on the mature coding sequence only; the fixed signal peptide and stop codon were excluded from all calculations, as these positions were held constant across parental and refined pairs and were not subject to synonymous refinement. Metrics computed included CAI[16], tAI[43], gtAI[63], RSCU[34], RCA[64], CBI[41], FOP[42], CFD[65], ENc/ENcp[35–37], GC/GC1/GC3 content[38], CpG and UpA dinucleotide statistics (frequency and observed/expected enrichment)[39,40], CPS/CUFS[66–68], COUSIN18/COUSIN59[69], ICDI[70], MCB[71], MILC[72], DCBS[73], RCBS[74], SCUO[75,76], B index[77], D mean[78], Ew[79], codon volatility[80], P2(tRNA)[81], P(codon)[82], and CDC[83]. Biologically expected directionality for improved translational efficiency (used for color coding in **Fig. 5a**) was assigned from prior literature, metrics without an established directional expectation were annotated accordingly.

mRNA secondary-structure analysis

mRNA secondary structure and associated thermodynamic descriptors were computed with the ViennaRNA package (RNAfold, v2.7.0)[45] via its Python bindings, at 37 °C under default Turner 2004 thermodynamic parameters. DNA sequences were transcribed to RNA prior to folding, and all sequences were analyzed in both "with signal peptide" (wSP) and "without signal peptide" (noSP) configurations; the signal peptide and stop codon were held constant across all parental and refined pairs. For each sequence, minimum free energy (MFE) and partition-function ensemble free energy (EnsFE) were computed on the full coding sequence, and a sliding-window minimum MFE was calculated by scanning a 60-nt window in single-nucleotide steps along the full CDS and retaining the lowest window MFE together with its position; the 60-nt window size was chosen to capture typical mRNA hairpin/stem-loop length scales while preserving positional

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

resolution across the transcript. The same MFE and EnsFE calculations were also performed on the 5′ prefix corresponding to the first 40 codons (120 nt) to characterize structure in the ribosome-loading region. To summarize ensemble-averaged structural content near the start codon, positional pairing probabilities were extracted from the partition-function base-pair probability matrix, converted to a binary paired/unpaired positional entropy, and averaged across the first 30 nucleotides to yield a scalar "start entropy" descriptor. Descriptors were computed for every expressed chain and averaged to yield per-molecule values, and paired differences (Δ = refined − parental) were used for all downstream statistical comparisons. Values reported in **Fig. 5c** correspond to the mature coding sequence without the signal peptide; the analogous wSP analysis is provided in **Supplementary Fig. 8** and yielded qualitatively concordant conclusions.

Protein structure analysis

Three-dimensional structures were predicted for each expressed antibody chain with ESMFold[49] (ESMFold2-Fast) on Tamarind Bio using default settings, folding each heavy and light chain independently as a monomer from its full-length sequence. Per-residue secondary structure and solvent accessibility were assigned with DSSP via Biopython[84]. DSSP's eight states were collapsed to three (helix, strand, loop), relative solvent accessibility (rel_asa) was computed by normalizing each residue's accessible surface area to its maximum reference value[85], and residues below 0.20 were classified as buried. Backbone φ/ψ angles were used to assign α-helical and β-sheet Ramachandran basins, and per-residue pLDDT was retained as a confidence measure. For each contiguous helix and strand, we defined the first two residues as the N-cap (start), the last two as the C-cap (end), and the one to two preceding coil residues as the pre-element flank. Predicted structures were aligned to the coding sequence by translating the

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

parental CDS and matching it to the modeled sequence, so each codon's edit status could be mapped to its residue's features; only chains with at least 20 aligned codons were retained.

To test whether editing was associated with any structural context independently of codon redundancy, each feature was fit in its own logistic regression predicting per-codon edit status, controlling for codon degeneracy (edited ~ feature + degeneracy). Binary features were modeled as context versus all other residues, and continuous features (rel_asa, pLDDT, normalized 5′→3′ position) were standardized and reported per standard deviation. Standard errors were clustered by chain, effects are reported as odds ratios with 95% confidence intervals, and p-values were corrected across the panel using the Benjamini-Hochberg false discovery rate[86] ($q < 0.05$ significant; **Supplementary Fig. 10**). Analyses used Biopython, pandas, NumPy, SciPy, and statsmodels in the same versions listed above.

Statistical analysis

All statistical analyses were performed in Python 3.12. Analyses supporting **Figs. 4** and **5** were run under Python 3.12.0 using SciPy v1.13.0, NumPy v1.26.4, pandas v2.2.3, and statsmodels v0.14.2, while the viral evolution analyses supporting **Fig. 2** were run under Python 3.12.2 in a separate high-performance computing environment using SciPy v1.15.1, NumPy v2.0.1, pandas v2.2.3, statsmodels v0.14.5, PyTorch v2.7.1 (CUDA 12.6), Transformers v4.48.3, and scikit-learn v1.6.1. Random seeds were fixed for all resampling procedures to ensure reproducibility. For the titer analysis in **Fig. 4**, per-construct $\log_2$-fold change was computed as $\log_2$(mean refined titer / mean parental titer) across biological triplicates, and panel-wide comparison of refined versus parental $\log_2$-titer was assessed using both a two-sided paired t-test and a two-sided Wilcoxon signed-rank test. The relationship between the change in mean chain-level SynCodonLM Evo score (ΔEvo = refined − parental, averaged across all expressed chains of a

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

molecule) and the observed $\Delta\log_2$(titer) was quantified by Spearman's rank correlation across n = 23 molecules. For the mechanistic analyses in **Fig. 5**, paired differences (Δ = refined − parental) for each codon-usage descriptor and each ViennaRNA-derived mRNA-structure descriptor were tested using a two-sided Wilcoxon signed-rank test, and per-descriptor correlations with $\Delta\log_2$(titer) were computed as the mean absolute Spearman's ρ across heavy- and light-chain-level computations, with signed ρ values reported alongside. To assess robustness of the per-descriptor rankings to individual influential molecules, a nonparametric case-resampled bootstrap was performed with 5,000 iterations using NumPy's PCG64 generator (seed = 9); on each iteration, molecules were sampled with replacement (n = 23) and Spearman's correlations between each descriptor and $\Delta\log_2$(titer) were recomputed, with iterations discarded when either variable lacked variance in the resample. Two-sided 95% confidence intervals were derived using the percentile method (2.5th and 97.5th percentiles across valid resamples), and a minimum of 100 valid resamples was required for a CI to be reported. For the viral evolution analyses in **Fig. 2**, Spearman's and Pearson correlations between per-sequence Evo score and sampling year were computed independently for each Influenza A H1N1 protein (nucleoprotein, hemagglutinin, neuraminidase), with R^2 reported as the square of the Pearson r. Because sample sizes were large (N = 855–6,710), p-values below the numerical precision floor of the underlying SciPy implementation are reported as $p < 1 \times 10^{-3}$.

Analytical-SEC to measure aggregation and protein purity

Antibodies were analyzed by analytical size exclusion chromatography (aSEC) to assess monomer purity and aggregation. Analyses were performed using an Agilent 1290 Infinity II system equipped with an ACQUITY UPLC Protein BEH SEC column (200 Å, 1.7 µm, 4.6 × 150 mm; Waters Corp., Milford, MA).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The separation was conducted at a flow rate of 0.5 mL/min over a 5 min runtime at 37 °C. Samples were eluted using a mobile phase consisting of 50 mM sodium phosphate, 200 mM L-arginine hydrochloride, and 0.05% (w/v) sodium azide at pH 6.8. The mobile phase was prepared fresh, filtered, and degassed prior to use.

Elution was monitored by UV absorbance at 280 nm. Monomer and aggregate species, including low molecular weight (LMW) and high molecular weight (HMW) components, were quantified based on peak area integration via Agilent OpenLab software.

Dynamic light scattering to provide further measurements on aggregation and protein purity

Dynamic light scattering (DLS) was performed to further evaluate protein aggregation and sample homogeneity. Measurements were conducted using a Prometheus Panta instrument (NanoTemper Technologies, Munich, Germany). Samples were analyzed under native conditions without dilution unless otherwise specified.

Measurements were carried out at 25 °C with automatic laser intensity optimization. The cumulant radius, polydispersity index, and scattering intensity were determined from the scattered light intensity using instrument-provided analysis software.

Size distributions were calculated based on intensity-weighted analysis to detect the presence of monomeric species and higher-order aggregates. Data were used to assess sample purity, aggregation propensity, and colloidal stability, complementing chromatographic analysis obtained by aSEC.

Viral evolution dataset curation

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We curated a viral evolution benchmark from the NCBI Virus database [87] by querying for Alphainfluenzavirus influenzae sequences isolated from human hosts and corresponding to genome segment 5, which encodes nucleoprotein. Coding sequences and associated metadata were downloaded and restricted to H1N1 isolates. Sequences were then filtered to retain only complete, in-frame coding DNA sequences: records containing non-ATGC characters, internal stop codons, lengths not divisible by three, or missing terminal stop codons were excluded.

To enable a controlled comparison between codon- and protein-level language models, coding sequences were collapsed by translated amino-acid sequence. Specifically, when multiple synonymous coding sequences encoded the same nucleoprotein sequence, only the earliest available coding-sequence representative for that protein sequence was retained. This deduplication strategy preserved a one-to-one mapping between each protein sequence and a single representative coding sequence, ensuring that SynCodonLM and ESM-2 were evaluated on matched evolutionary sequence instances rather than on unequal codon-level redundancy.

Viral Evolution Evo Score Calculations

Evolutionary compatibility scores were computed separately for matched coding and protein sequences using SynCodonLM and ESM-2, respectively. SynCodonLM was applied to the full viral coding sequence, including the initiating ATG and terminal stop codon. ESM-2 was applied to the corresponding translated protein sequence from the initiating methionine through the final amino acid, excluding the stop codon because stop symbols are not represented as standard amino-acid tokens in the protein model vocabulary. ESM-2 is a transformer-based protein language model distributed as part of the Evolutionary Scale Modeling family of pretrained protein models, and ESM-2-style implementations support masked-token prediction and per-token logits suitable for sequence scoring.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

For each model, an Evo score was calculated using a single-position masked pseudo-log-likelihood framework. For a sequence x = $(x_1, \ldots, x_L)$, each position i was masked independently, and the model-assigned log probability of the observed token $x_i$was computed conditional on the remaining unmasked sequence context $x_{\setminus i}$. The sequence-level Evo score was then defined as the mean pseudo-log-likelihood across all scored positions:

Equation 7:

$$Evo\ score(x) = \frac{1}{L}\sum_{i=1}^{L} \log p\left(x_i \mid x_{\setminus i}\right),$$

where x_i denotes either a codon token for SynCodonLM or an amino-acid token for ESM-2. Higher Evo scores indicate greater compatibility between the observed sequence and the evolutionary/statistical constraints learned by the corresponding language model. Because SynCodonLM and ESM-2 operate over different token spaces and were trained under different modeling objectives, Evo scores were interpreted within each model rather than as directly interchangeable absolute quantities across models. Comparisons between codon- and protein-level scoring were therefore based on relative trends across matched viral sequence instances.

AI Usage

Microsoft Copilot was used to assist with drafting portions of the manuscript and with code development. All content was reviewed, edited, and validated by the authors.

Code Availability

A repository to reproduce the work in this article, as well as utilize the model for future purposes, is available at https://github.com/Boehringer-Ingelheim/Directed-Evolution-in-Codon-Space. The software is fully open source and built for ease of use and implementation.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Data Availability

Data is available attached to this article in **Supplemental Data**.

Acknowledgements

We thank Boehringer Ingelheim for providing the high-performance computing resources as well as wet lab, high throughput molecular biology facilities utilized in this work. J.H. gratefully acknowledges funding support from S.C. through his institute professor's endowed chair, as well as additional support from Boehringer Ingelheim.

References

1. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).

2. Liu, Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal.* **18**, 145 (2020).

3. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **59**, 744–754 (2015).

4. Chen, S. *et al.* Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. *Mol. Biol. Evol.* **34**, 2944–2958 (2017).

5. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).

6. Kimchi-Sarfaty, C. *et al.* A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science* **315**, 525–528 (2007).

7. Sun, L. *et al.* Codon optimization regulates IgG3 and IgM expression and glycosylation in N. benthamiana. *Front. Bioeng. Biotechnol.* **11**, 1320586 (2023).

8. Núñez-Manchón, E. *et al.* Transgene codon usage drives viral fitness and therapeutic efficacy in oncolytic adenoviruses. *NAR Cancer* **3**, zcab015 (2021).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

9. Ronzitti, G. *et al.* A translationally optimized AAV-UGT1A1 vector drives safe and long-lasting correction of Crigler-Najjar syndrome. *Mol. Ther. - Methods Clin. Dev.* **3**, 16049 (2016).

10. Bell, P. *et al.* Effects of Self-Complementarity, Codon Optimization, Transgene, and Dose on Liver Transduction with AAV8. *Hum. Gene Ther. Methods* **27**, 228–237 (2016).

11. Cabanes-Creus, M. *et al.* Codon-Optimization of Wild-Type Adeno-Associated Virus Capsid Sequences Enhances DNA Family Shuffling while Conserving Functionality. *Mol. Ther. - Methods Clin. Dev.* **12**, 71–84 (2019).

12. Magistrelli, G. *et al.* Optimizing assembly and production of native bispecific antibodies by codon de-optimization. *mAbs* **9**, 231–239 (2017).

13. Srila, Oli, W. and, Yamabhai, D. and & Montarop. Codon and Signal Peptide Optimization to Enhance Therapeutic Antibody Production from CHO Cells. 33–48 (2025) doi:10.1007/978-1-0716-4104-0_4.

14. Lin, Y. & Wang, T. Enhancement of adalimumab expression in CHO cells by antioxidant small molecule additives and the underlying mechanism. https://doi.org/10.21203/rs.3.rs-6788918/v1 (2025) doi:10.21203/rs.3.rs-6788918/v1.

15. Mauro, V. P. & Chappell, S. A. A critical analysis of codon optimization in human therapeutics. *Trends Mol. Med.* **20**, 604–613 (2014).

16. Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**, 1281–1295 (1987).

17. Goulet, D. R. *et al.* Codon Optimization Using a Recurrent Neural Network. *J. Comput. Biol.* **30**, 70–81 (2023).

18. Fu, H. *et al.* Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* **10**, 17617 (2020).

19. Rajbanshi, B. & Guruacharya, A. codonGPT: Reinforcement learning on a generative language model optimizes RNA sequences under biological constraints. *bioRxiv* https://doi.org/10.1101/2025.06.25.661500 (2025) doi:10.1101/2025.06.25.661500.

20. Fallahpour, A., Gureghian, V., Filion, G. J., Lindner, A. B. & Pandi, A. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat. Commun.* **16**, 3205 (2025).

21. Chen, Y., Zhang, Y., Li, J., Tian, B. & Huang, H. CodonTranslator: a conditional codon language model for codon optimization across life domains. *bioRxiv* 2025.11.24.690310 (2025) doi:10.1101/2025.11.24.690310.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

22. Zulkower, V. & Rosser, S. DNA Chisel, a versatile sequence optimizer. *Bioinformatics* **36**, 4508–4509 (2020).

23. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024).

24. Heuschkel, J., Kingsley, L., Pefaur, N., Nixon, A. & Cramer, S. Advancing codon language modeling with synonymous codon constrained masking. *Nucleic Acids Res.* **54**, gkag166 (2026).

25. Hallee, L., Rafailidis, N. & Gleghorn, J. P. cdsBERT - Extending Protein Language Models with Codon Awareness. *bioRxiv* 2023.09.15.558027 (2023) doi:10.1101/2023.09.15.558027.

26. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intell.* **6**, 170–179 (2024).

27. Li, S. *et al.* CodonBERT large language model for mRNA vaccines. *Genome Res.* **34**, 1027–1035 (2024).

28. Ren, Z. *et al.* CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics* **40**, btae330 (2024).

29. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

30. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.

31. Durocher, Y., Perret, S. & Kamen, A. High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res.* **30**, e9–e9 (2002).

32. Stuible, M. *et al.* Optimization of a high-cell-density polyethylenimine transfection method for rapid protein production in CHO-EBNA1 cells. *J. Biotechnol.* **281**, 39–47 (2018).

33. Fan, K., Li, Y., Chen, Z. & Fan, L. GenRCA: a user-friendly rare codon analysis tool for comprehensive evaluation of codon usage preferences based on coding sequences in genomes. *BMC Bioinform.* **25**, 309 (2024).

34. Sharp, P. M. & Li, W.-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).

35. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

36. Satapathy, S. S., Sahoo, A. K., Ray, S. K. & Ghosh, T. C. Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc ($\hat{N}c$) and ENCprime ($\hat{N}'c$) measures. *Genes Cells* **22**, 277–283 (2017).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

37. Alexaki, A. *et al.* Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. *J. Mol. Biol.* **431**, 2434–2441 (2019).

38. Stenico, M., Lloyd, A. T. & Sharp, P. M. Codon usage in Caenorhabditis elegans : delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22**, 2437–2446 (1994).

39. Bauer, A. P. *et al.* The impact of intragenic CpG content on gene expression. *Nucleic Acids Res.* **38**, 3891–3908 (2010).

40. Simmonds, P., Xia, W., Baillie, J. K. & McKinnon, K. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genom.* **14**, 610 (2013).

41. Bennetzen, J. L. & Hall, B. D. Codon selection in yeast. *J. Biol. Chem.* **257**, 3026–3031 (1982).

42. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**, 573–597 (1982).

43. Reis, M. dos, Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).

44. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258 (2009).

45. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

46. Verma, M. *et al.* A short translational ramp determines the efficiency of protein synthesis. *Nat. Commun.* **10**, 5774 (2019).

47. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**, 344–354 (2010).

48. Moss, M. J., Chamness, L. M. & Clark, P. L. The Effects of Codon Usage on Protein Structure and Folding. *Annu. Rev. Biophys.* **53**, 87–108 (2023).

49. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

50. Diepenbruck, C., Klinger, M., Urbig, T., Baeuerle, P. & Neef, R. Productivity and Quality of Recombinant Proteins Produced by Stable CHO Cell Clones can be Predicted by Transient Expression in HEK Cells. *Mol. Biotechnol.* **54**, 497–503 (2013).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

51. Raybould, M. I. J. *et al.* Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* **48**, D383–D388 (2019).

52. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).

53. Peryea, T. *et al.* Global Substance Registration System: consistent scientific descriptions for substances related to health. *Nucleic Acids Res.* **49**, D1179–D1185 (2020).

54. Francisco, J. A. *et al.* cAC10-vcMMAE, an anti-CD30–monomethyl auristatin E conjugate with potent and selective antitumor activity. *Blood* **102**, 1458–1465 (2003).

55. Labrijn, A. F. *et al.* Efficient generation of stable bispecific IgG1 by controlled Fab-arm exchange. *Proc. Natl. Acad. Sci.* **110**, 5145–5150 (2013).

56. Smith, E. J. *et al.* A novel, native-format bispecific antibody triggering T-cell killing of B-cells is robustly active in mouse tumor models and cynomolgus monkeys. *Sci. Rep.* **5**, 17943 (2015).

57. Ridgway, J. B. B., Presta, L. G. & Carter, P. 'Knobs-into-holes' engineering of antibody CH3 domains for heavy chain heterodimerization. *Protein Eng., Des. Sel.* **9**, 617–621 (1996).

58. Dall'Acqua, W. F., Kiener, P. A. & Wu, H. Properties of Human IgG1s Engineered for Enhanced Binding to the Neonatal Fc Receptor (FcRn)*. *J. Biol. Chem.* **281**, 23514–23524 (2006).

59. Angal, S. *et al.* A single amino acid substitution abolishes the heterogeneity of chimeric mouse/human (IgG4) antibody. *Mol. Immunol.* **30**, 105–108 (1993).

60. Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* https://doi.org/10.48550/arxiv.1910.03771 (2019) doi:10.48550/arxiv.1910.03771.

61. Ho, S. C., Tong, Y. W. & Yang, Y. Generation of monoclonal antibody-producing mammalian cell lines. *Pharm. Bioprocess.* **1**, 71–87 (2013).

62. Goldfarb, T. *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res.* **53**, D243–D257 (2024).

63. Anwar, A. M. *et al.* gtAI: an improved species-specific tRNA adaptation index using the genetic algorithm. *Front. Mol. Biosci.* **10**, 1218518 (2023).

64. Fox, J. M. & Erill, I. Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression. *DNA Res.* **17**, 185–196 (2010).

65. Clarke, T. F. & Clark, P. L. Rare Codons Cluster. *PLoS ONE* **3**, e3412 (2008).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

66. Kunec, D. & Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep.* **14**, 55–67 (2016).

67. Coleman, J. R. *et al.* Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science* **320**, 1784–1787 (2008).

68. Diament, A., Pinter, R. Y. & Tuller, T. Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. *Nat. Commun.* **5**, 5876 (2014).

69. Bourret, J., Alizon, S. & Bravo, I. G. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biol. Evol.* **11**, 3523–3528 (2019).

70. Freire-Picos, M. A. *et al.* Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. *Gene* **139**, 43–49 (1994).

71. Urrutia, A. O. & Hurst, L. D. Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection. *Genetics* **159**, 1191–1199 (2001).

72. Supek, F. & Vlahoviček, K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinform.* **6**, 182 (2005).

73. Sabi, R. & Tuller, T. Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. *DNA Res.* **21**, 511–526 (2014).

74. Roymondal, U., Das, S. & Sahoo, S. Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia coli Genome. *DNA Res.* **16**, 13–30 (2009).

75. Wan, X.-F., Xu, D., Kleinhofs, A. & Zhou, J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* **4**, 19 (2004).

76. Wan, X.-F., Zhou, J. & Xu, D. CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. *Int. J. Gen. Syst.* **35**, 109–125 (2006).

77. Karlin, S., Mrázek, J. & Campbell, A. M. Codon usages in different gene classes of the Escherichia coli genome. *Mol. Microbiol.* **29**, 1341–1355 (1998).

78. Suzuki, H., Saito, R. & Tomita, M. Measure of synonymous codon usage diversity among genes in bacteria. *BMC Bioinform.* **10**, 167 (2009).

79. Suzuki, H., Saito, R. & Tomita, M. The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. *Gene* **335**, 19–23 (2004).

80. Plotkin, J. B., Dushoff, J. & Fraser, H. B. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature* **428**, 942–945 (2004).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.03.742557; this version posted August 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

81. Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**, 7055–7074 (1982).

82. Gribskov, M., Devereux, J. & Burgess, R. R. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* **12**, 539–549 (1984).

83. Zhang, Z. *et al.* Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinform.* **13**, 43 (2012).

84. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

85. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE* **8**, e80635 (2013).

86. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).

87. Brister, J. R., Ako-adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).