Designing Convergent Overlapping Genes with Transformer Encoder Models and Lightweight Structural Proxies

Jason K. Morgan [1,*]

[1] Independent Researcher

* Corresponding author at: jkm783@gmail.com

Abstract

Overlapping genes allow multiple proteins to be encoded from a single DNA sequence, including convergent (antisense; tail-to-tail) orientations across three reading frames (phases 0, 1, and 2), with phase 1 most frequently observed in nature. Designing such overlaps is challenging due to codon degeneracy, phase-specific biases, and the need to preserve structural integrity for both proteins. Here, a purpose-built transformer encoder is introduced, trained on a balanced synthetic dataset of convergent overlaps spanning diverse prokaryotic genomes and GC contents. Controlled amino acid substitutions were incorporated during training to enhance model generalization, particularly for phase 1 overlaps. At inference, Monte Carlo dropout enabled uncertainty-aware sampling of synonymous codon solutions, which were iteratively refined using a windowed, multi-objective optimization framework. Candidate overlaps were scored using composite weighting across secondary structure preservation, substitution similarity, alignment identity, and ESM-2 contact map similarity, with the structural similarity index measure (SSIM) applied as a rapid proxy for structural fidelity. This approach generated convergent overlaps across all phases, with phase 1 showing the highest success rates.

Optimization trajectories revealed distinct dynamics, with secondary structure preservation steadily increasing despite its lower weight. External validation using SwissProt proteins stratified by AlphaFold2 (AF2) predicted local distance difference test (pLDDT) confidence supported generalization to proteins with differing rigidity, yielding high secondary structure preservation in silico. These results demonstrate that transformer models trained directly at the nucleotide level, when coupled with uncertainty-aware inference and lightweight structural proxies, can support the computational design of synthetic overlapping genes without requiring full structural prediction. This framework offers a scalable path for phase-specific, codon-aware overlap design under realistic constraints.

Introduction

The study of overlapping genes is an important area of interest in prokaryotic genetics, attributed to their ability to encode multiple functional products from a single DNA sequence (1–4). For example, a single gene sequence might code for two distinct proteins, each playing a different role in cellular function. Advancements in identifying and designing these overlapping sequences promise the creation of genetic constructs that are both more efficient and compact, potentially boosting the functionality and stability of engineered organisms (5,6). This phenomenon not only poses challenges but also opens new avenues in genetic engineering and synthetic biology (3).

Overlapping genes have been described using different approaches, but can be broadly grouped into unidirectional (sense; encoded on the same strand, → →) or convergent (antisense; encoded on the opposite strand) orientations (**Fig 1A**); convergent overlaps can be further refined into "head-to-head" (← →) or "tail-to-tail" (→ ←) orientations (7). Convergent overlapping genes occur in the same reading frame (phase 0), or frameshifted by one (phase 1) or two (phase 2)

nucleotides. For example, in phase 2, convergent overlaps share their third (degenerate) codon positions. Research into convergent (tail-to-tail; antisense) overlapping gene sequences in prokaryotes has included analyses on the length distribution among overlapping and near-overlapping genes (8), investigations into their prevalence (and misannotations) within prokaryotic genomes (9,10), as well as the characterization of convergent gene pairs (11–18). Notably, both convergent and unidirectional overlaps exhibit a non-uniform distribution with a significant phase bias, predominantly favoring phase 1 overlaps (8,19,20). While the majority of convergent overlaps fall into phase 2, when excluding 4 nucleotide overlaps, nearly half are instead found in phase 1 (8). These studies collectively highlight the phase preference and distribution patterns of naturally occurring gene overlaps.

The successful prediction of overlapping nucleotide sequences encoding two amino acid sequences poses unique computational challenges, and has been previously explored (6,21–25); the landscape of available methods has been summarized (22) with additional information pertaining to more recently released programs provided in **Supplementary Table 1**. An early solution for the design of unidirectional overlapping protein sequences included design of an algorithm to identify the shortest DNA sequence which could encode two amino acid sequences (26). More recently, examples included a dynamic programming algorithm (6) and a subsequent method that combines dynamic programming with a hidden Markov model (HMM) (5,27). The strategy was further extended to arbitrary pairs of natural protein domains (28). These approaches have collectively shown that it is possible to design fully embedded overlapping genes that encode functional proteins with high homology to starting sequences, underscoring the potential for engineered overlapping genes in prokaryotic genetics. Another approach, termed overlapping, alternate-frame insertion (OAFI) demonstrated the ability to design unidirectional

overlaps through insertion of one gene into an alternate reading frame of another by targeting sites that may tolerate insertions (29). Recently, Byeon et al. (2025) demonstrated that deep generative models can be used to design synthetic overlapping genes spanning distinct protein families, achieving high in silico and experimental success rates even under the constraints of the standard genetic code (30). Their results indicate that sequence space may contain many viable overlapping solutions, suggesting broader applicability of computational approaches to both unidirectional and convergent overlaps.

Transformer models have demonstrated superior performance in various sequence modeling tasks due to their ability to capture long-range dependencies and contextual information (31–33). This study leverages a novel application of the transformer encoder-based model architecture to predict synthetic convergent overlapping genes in prokaryotes. Unlike dynamic programming approaches, transformers utilize self-attention mechanisms to process the entire input sequence simultaneously (32), enabling the model to capture patterns and dependencies that span across long sequences. In this study, Monte Carlo (MC) dropout is applied during inference to enhance prediction performance. Dropout, more generally, is a method of reducing overfitting in deep neural networks during training (34). MC dropout retains dropout activity during inference, effectively sampling from a distribution of possible model predictions (34–36). This provides a practical Bayesian approximation of model uncertainty (37), which is especially advantageous for predicting overlapping genes where the degeneracy of the genetic code permits many distinct but functionally viable nucleotide encodings. By sampling multiple plausible solutions, MC dropout enables exploration of the many synonymous coding possibilities inherent to the genetic code, increasing the likelihood of identifying overlaps that preserve the structural and functional integrity of both proteins.

The objectives of this study are to (i) evaluate the ability of transformer encoder models to recover convergent overlaps across all three phases, (ii) quantify the impact of overlap length and sequence variability on prediction success, and (iii) assess multi-objective optimization as an alternative to full structural prediction in improving overlap fidelity.

Methods

Synthetic Overlapping Gene Dataset Construction for Model Training

While naturally occurring convergent overlaps have been systematically assessed and documented (9), there is no readily available database of convergent overlapping genes with equally distributed overlap lengths. Given this limitation, a database of synthetic convergent overlaps was generated using a diverse set of open reading frames (ORFs) from prokaryotic genomes. Synthetic convergent overlapping genes were created using ORF sequences from 24 different genomes from prokaryotes (**Supplementary Table 2**), using the R (version 4.3.1) and Python programming languages (version 3.8.8). Genomes were selected to have a wide range of genomic GC content (ranging from 23.3% to 74.9%, with a mean of 52.41%), with the aim of controlling for potential amino acid compositional bias driven by GC content (38). The between-strain coding GC content (calculated using the set of ORF sequences from each strain) ranged from 20.7% to 72.7%, with a mean of 50.5%.

Briefly, for each strain, sequences were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/) and each FASTA file was processed to extract ORFs, which were stored in separate list for each strain (**Fig 1B**). ORF sequences were filtered to remove sequences shorter than 450 nucleotides, to exclude sequences with insufficient opportunity for varied random sampling during overlap construction. A function was defined to

112 randomly select in-frame segments from the ORF sequence lists, to include a primary and
113 secondary sequence either with or without a convergent overlapping region. First, a contiguous
114 in-frame 312-nucleotide length section from each ORF sequence was extracted and used as the
115 basis for the primary sequence. The primary sequence segments were modified to ensure they
116 contained start and randomly selected stop codons, as these were removed during initial
117 sequence processing.

118 The reverse complement of this synthetic gene was selected to generate the secondary sequence
119 with a defined convergent overlap length. A uniform set of overlap lengths was generated for
120 each length by extracting the reverse complement, adding a stop codon, and removing any in-
121 frame stop codons with a random codon. Both the added stop codon and in-frame replacements
122 were selected such that stop codons in the forward direction were not introduced. However,
123 owing to the methodology employed, a portion of generated secondary sequences with
124 overlapping sequences contained more than one stop codon and were subsequently filtered from
125 the dataset to ensure all sequences contained only one stop codon and encoded a single amino
126 acid sequence.

127 The amino acid sequence pairs with known overlaps were modified to incorporate up to 4 of
128 BLOSUM62 weighted amino acid changes. The changes included altering amino acid sequences
129 such that 1) the internal reverse stop codon was removed from both amino acid sequences
130 (primary and secondary), while keeping the known overlap for the original amino acid
131 sequences, or 2) the internal reverse stop codon was removed from both amino acid sequences
132 (primary and secondary), while also introducing additional amino acid modifications per
133 sequence. This yielded modified primary and secondary amino acid sequences, with an

unmodified known convergent overlap nucleotide sequence capable of encoding amino acid sequences with high amino acid sequence identity to the primary and secondary amino acid sequences.

The length of the input amino acid sequences was limited to 104 amino acids to manage computational complexity during training and inference phases. Each amino acid was separated by a space and the overall included a terminal asterisk (*) and the primary and secondary sequences were concatenated to generate a single input sequence for training.

To balance the representation of different overlap lengths, data were stratified based on the length of the overlap. The overlapping gene training data were balanced, ensuring equal representation for each strain and each length of nucleotide overlap (specifically, convergent overlaps of length 199 to 312 nucleotides). The final dataset was composed of synthetic gene overlaps, each represented by a pair of modified amino acids sequences with a known target convergent overlap sequence. This dataset served as the training set for the transformer encoder-based models.

Initial Model Description and Training

The synthetic overlapping gene dataset was used to train transformer-encoder based models with the aim of generating convergent (tail-to-tail) overlaps for any two specified amino acid sequences. A total of 8,892,000 combined pairs, corresponding to 78,000 pairs per convergent overlap length.

Transformers have demonstrated superior performance in various sequence modeling tasks, particularly where understanding long-range dependencies is important (31–33). As such, this

155 study utilizes a PyTorch based transformer encoder model for overlap prediction (39).
156 Tokenization and text vectorization were performed using the TensorFlow Keras API (40), and
157 data were split into training and validation sets. During training, the models each utilized 156
158 embedding dimensions, 13 attention heads, 192 feed-forward network dimensions, 5 blocks, a
159 dropout rate of 0.1, and a total of 802,983 trainable parameters (**Fig. 2D**). Models were compiled
160 using the Adam optimizer with cross-entropy as the loss function.

161 Each model was trained for 3 epochs with a batch size of 32 on a single NVIDIA GeForce GTX
162 5070 Ti GPU with 16GB of memory, and reached a validation accuracy of ≥90% by the third
163 and final epoch.

164 **Test Dataset Generation and Model Inference**

165 Amino acid pairs were generated as described above for preparation of the training dataset, but
166 without the final modification step; however, a different set of strains (n = 12) were selected than
167 those used for training, with a wide range of genomic GC content (ranging from 23.6% to 74.1%,
168 with a mean of 50.2%) (**Supplementary Table 3**). Given that inference was performed using
169 only coding sequences, the coding sequence GC content was calculated for each strain, and the
170 corresponding mean GC value was used for subsequent analyses (varying from 24.2% to 72.1%,
171 and a mean of 50.0%).

172 It is noted that inference was performed using the trained models along with the same tokenizer
173 to vectorize the input sequence. Using a "greedy decoder" (41), the token with the highest
174 probability at each step of the sequence generation process was selected.

175 MC dropout was implemented during inference to allow exploration of the uncertainty in model
176 predictions (eg, variation in predicted overlap sequences), leveraging the benefits of Bayesian
177 inference, as previously described (34,36,42). During inference, depending on the objective,
178 feedforward dropout was set to 0.275 for initial predictions.

179 An algorithm was developed to generate overlapping gene sequences from input amino acid
180 sequences using the trained models. The algorithm first back-translated the input amino acid
181 sequences into a nucleotide sequence, leveraging *E. coli* codon usage frequencies to reflect the
182 probabilistic distribution of codons for each amino acid (43). After back-translation, the models
183 were used to predict potential overlaps within these nucleotide sequences. Predicted overlaps
184 were then integrated into the original nucleotide sequences, resulting in a collection of modified
185 sequences that incorporated the predicted overlapping regions downstream of the remaining
186 back-translated nucleotide sequence. These modified sequences (ie, the back-translated and
187 predicted regions) were then translated into amino acid sequences to assess the accuracy of the
188 predictions. In instances where the correct amino acid sequence was not identified, additional
189 rounds of prediction and back-translation were performed. This iterative process continued until
190 either an overlap yielding amino acid sequences meeting the pre-specified criteria was identified,
191 or the specified number of inference rounds were completed without a predicted overlap (defined
192 as an unsuccessful prediction).

193 An alignment score was calculated for each sequence prediction. The predicted sequence was
194 aligned against the known overlap using the Needleman-Wunsch algorithm (implemented via
195 pairwise2 from Biopython (44)). For sequences with a known overlap and a predicted overlap

196 sequence, the calculated alignment score was based on the length of the overlap, normalizing
197 scores to a scale of 0 to 1 to facilitate comparison across predictions.

198 **Secondary Structure Prediction, Substitution Score, and Structure Prediction**

199 The prediction of secondary structures was performed using S4PRED, which enabled predictions
200 based on an input amino acid sequence alone without the need for multiple sequence alignments
201 or known homologous sequences (45). S4PRED outputs were formatted as "ss2" files, from
202 which the predicted secondary structure sequence was extracted for downstream analyses. The
203 model employs a simplified three-state classification—α-helix (H), β-sheet (E), and coil/loop
204 (C), which corresponds to a coarse-grained version of the DSSP (Dictionary of Secondary
205 Structure of Proteins) annotations (46). Minor code modifications were implemented to enable
206 efficient batch processing, substantially reducing runtime. These predicted structures were used
207 to evaluate the structural compatibility of convergent overlapping gene candidates. Substitution
208 matrices utilized were BLOSUM62 (47,48) and ProtSub (49). Protein structures were predicted
209 either with AlphaFold3 or ColabFold (50–52).

210 **Optimization of predicted sequences**

211 An iterative, windowed optimization procedure was used to convert paired amino-acid inputs
212 into convergent (tail-to-tail) overlapping nucleotide sequences while preserving fold-relevant
213 features of both proteins and maintaining sequence similarity. Input amino-acid sequences were
214 first back-translated to an initial nucleotide scaffold using *E. coli* codon usage frequencies. The
215 region targeted for overlap design was tiled into partially overlapping amino-acid windows; in
216 the experiments reported here windows were 10 amino acids with an 8-amino-acid stride (10 AA

window, 8 AA stride; model receptive field used = 105 AA). During optimization, only nucleotides mapping to the current window were permitted to change; flanking nucleotides were held fixed except at user-designated residues enforced by token-level biases (for example to preserve active-site residues or motifs). Bracketed constraints on the forward strand were mirrored on the reverse-complement strand so that enforced bases remained consistent with the convergent orientation.

Within each window a transformer-encoder was executed in an uncertainty-aware mode: Monte-Carlo (MC) dropout was retained at inference to generate diverse synonymous codon proposals. Token-level fixed-logit (base) biases were applied at specified nucleotide positions to enforce required bases during sampling; combined forward/reverse fixed-logit maps ensured constraints were respected on both strands. Each sampled nucleotide pair was translated to amino-acid sequences and duplicate amino-acid solutions were removed prior to scoring.

Structural compatibility was assessed using single-sequence secondary structure prediction (S4PRED; three-state H/E/C). Secondary structure preservation was reported as the percentage of residues for which a candidate's predicted state matched the original sequence prediction. Substitution similarity was computed using BLOSUM62 and normalized to the original sequence's self-score to yield a percent similarity; global alignment identity was also measured. ESM-2 (53) (model: esm2_t12_35M_UR50D) was used to generate residue-residue contact probability matrices, having demonstrated utility as a fast and effective single-sequence prediction feature (54). Contact map similarity was assessed using the structural similarity index (SSIM) (55), computed between ESM-2-predicted contact maps of candidate overlap sequences and their respective originals. SSIM quantifies local and global structural pattern agreement

239 (range 0-1, with 1 indicating identical maps). Raw values were linearly scaled to a percentage (0-
240 100) to yield the ESM-2 contact map score. This metric served as a rapid, computationally
241 efficient proxy for structural preservation, enabling large-scale overlap design iterations while
242 retaining sensitivity to perturbations in contact map topology.

243 Candidates were ranked by a fixed weighted composite score across all experiments: 0.15
244 (secondary structure preservation), 0.15 (substitution similarity), 0.1 (alignment identity), and 0.6
245 (ESM2 contact map similarity). Although this weighting scheme was not exhaustively
246 optimized, empirical testing with randomly selected SwissProt proteins indicated that it yielded
247 stable rankings across experiments and supported overlap generation while balancing structural
248 fidelity with sequence similarity (**Figs. S3, S4, and S5**). The top-ranked candidate for the
249 window was integrated, and optimization proceeded to the next window. Multiple full passes
250 across windows were permitted for refinement. After the first structurally successful candidate,
251 global feedforward and attention dropout levels were reduced to bias sampling from exploratory
252 search toward fine-tuning, though the weighting scheme itself remained unchanged.

253 Efficiency measures implemented for reproducibility and speed included: (i) caching of
254 secondary structure predictions so identical amino-acid sequences were scored only once, and
255 (ii) caching of ESM-2 embeddings and contact maps. All candidate metadata (attempt, window,
256 per-candidate metrics, and secondary structure predictions) were retained in memory and
257 exported at the end of each input row to permit analysis.

258 **Mutual information, PCA, and robustness testing**

Terminal regions (104 amino acids) from 250 randomly selected SwissProt-derived proteins were used as seeds to generate ensembles of mutated sequences. For each seed, trajectories were created by stepwise amino acid substitutions sampled from the 20-letter alphabet, introducing one mutation per step for up to 75 iterations. This yielded a progressive set of variant sequences per seed protein. Each sequence in the ensemble was then scored for global alignment identity, ESM-2 contact map similarity, and secondary structure match relative to the reference. These metric vectors formed the input for subsequent mutual information (MI) and principal component analyses.

MI was estimated after quantile discretization of continuous scores into 20 bins to evaluate pairwise dependence among objectives. To reduce sensitivity to binning, MI was also estimated using a k-nearest-neighbor approach. Statistical significance of discrete MI values was assessed using permutation testing with 1,000 random shuffles. Principal component analysis (PCA) was applied to z-scored metric vectors (alignment, ESM-2 contact map similarity, secondary structure match) to examine the dimensional structure of variation. Robustness checks included varying the number of bins (5-40) for MI and bootstrap resampling of PCA explained variance.

Statistical Analyses and Figure Preparation

The R programming language (version 4.3.1) and the ggplot2 package (56,57) were used to perform statistical analyses and generate figures. Prediction success rates are reported as percentages with exact 95% confidence intervals calculated using the Clopper–Pearson method. Correlations between similarity metrics were assessed using Pearson's correlation coefficient with associated *p*-values. Relationships between secondary structure composition (coil, helix, sheet) and prediction performance were examined using LOESS smoothing with 95–99%

confidence bands. For external validation experiments, performance metrics were stratified by AlphaFold2 pLDDT confidence brackets and overlap phase. Group differences were evaluated using nonparametric pairwise contrasts with adjusted p-values (Bonferroni correction). Effect sizes are reported as absolute percentage-point differences between groups, alongside confidence intervals. Data are presented as mean values with ranges or confidence intervals where appropriate. Figures were generated with ggplot2, with additional overlays and schematics prepared in Inkscape (version 1.3.2).

Declaration of Generative AI and AI-assisted technologies

During the preparation of this work, ChatGPT (5.0; OpenAI) was used to assess readability and language, and as a tool to generate code snippets and assist in debugging. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Code and data availability

The datasets and code used for constructing the synthetic overlap dataset, training the models, and performing all analyses are available at: 10.5281/zenodo.17626094. Code for running model inference (to predict convergent overlaps) is available at: https://github.com/protosome/convergent_overlaps_aa_change

Results

Prediction of convergent overlapping genes from amino acid sequence pairs from a dedicated transformer-encoder model set

301 While prior work has established that it is technically feasible to generate overlapping genes
302 from paired amino acid sequences, the performance of dedicated transformer-encoder models for
303 this task has not been systematically evaluated. In this study, transformer-based models were
304 trained to generate convergent overlapping nucleotide sequences from amino acid pairs, and their
305 ability to recover overlaps across all three phases was assessed. These models were designed to
306 identify convergent overlaps capable of encoding both input proteins, using recent advances in
307 transformer-based sequence modeling to address the large solution space introduced by the
308 degeneracy of the genetic code (31,58). Prediction quality was evaluated using alignment
309 identity scores and secondary structure preservation metrics, allowing comparison across overlap
310 lengths from 199 to 312 nucleotides.

311 Initial filtering was performed using a conservative alignment score cutoff of 34% (28,59).
312 Without Monte Carlo dropout (ie, dropout enabled during inference), successful predictions were
313 substantially reduced, and repeated inference runs converged to identical codon solutions. This
314 indicated that deterministic decoding restricted the search space to narrow, low-diversity
315 outcomes. By contrast, applying Monte Carlo dropout during inference expanded sequence
316 diversity, producing multiple distinct codon assignments that encoded the same amino acid
317 sequences. This stochastic sampling increased the likelihood of recovering overlaps that satisfied
318 both similarity and structural thresholds, underscoring the importance of uncertainty-aware
319 inference in navigating the combinatorial complexity of convergent overlap prediction.

320 To further assess training requirements, the effect of imperfect overlaps and amino acid
321 substitutions in the training data was examined. Models trained only on unaltered overlaps
322 exhibited markedly reduced success rates, particularly for phase 1 (**Figs. 3D, 3E**). In a shared test

323 set of 5,000 amino acid sequence pairs without a known convergent overlap, training with
324 modified sequences (stop codon removed plus 1–4 substitutions per sequence) substantially
325 improved prediction rates: phase 0, 95.3% (94.7-95.9%); phase 2, 86.7% (85.7-87.6%); and
326 phase 1, 81.6% (80.5-82.6%). In contrast, models trained on unaltered overlaps performed
327 significantly worse: phase 0, 73.7% (72.4-74.9%); phase 2, 49.1% (47.7-50.5%); and phase 1,
328 9.7% (8.9-10.6%). These findings are consistent with the hypothesis that incorporating
329 controlled sequence variation during training increased model flexibility and suggest improved
330 generalization, particularly when coupled with Monte Carlo dropout during inference. On this
331 basis, models trained with amino acid substitutions were selected for all subsequent analyses.

332 The outputs of these models were further evaluated using BLOSUM62 and ProtSub substitution
333 matrices. For overlap lengths of 310-312 nucleotides, similarity scores were strongly correlated
334 with amino acid identity ($r = 0.88$; $p < 0.0001$), and the matrices themselves were highly
335 correlated with each other ($r = 0.98$; $p < 0.0001$).

336 In addition to substitution metrics, alignment scores and secondary structure scores were
337 compared across phases. Phase 1 overlaps exhibited the highest average alignment and
338 secondary structure scores, phase 0 displayed intermediate values, and phase 2 was generally the
339 lowest. However, score distributions overlapped considerably, suggesting that phase-dependent
340 differences were not absolute. The near-identical secondary structure score distributions for
341 phase 0 and phase 2 further indicated that structural outcomes may be more sensitive to
342 sequence-specific context than to phase alone.

343 To broaden the analysis, overlap lengths from 199 to 312 nucleotides were examined across a
344 larger set of amino acid pairs. Phase-dependent patterns persisted: phase 1 overlaps were

345 increasingly recoverable at shorter lengths, and the relative rank order of secondary structure and
346 alignment scores across phases remained consistent. These findings suggest that phase 1 overlaps
347 may be predicted to be structurally and codon-compatible favorable, and that performance of the
348 transformer encoder model set is influenced by overlap length and the diversity of sequences
349 available for training.

350 **High Secondary Structure Scores Are Linked to Coil-Dominated Precursor Sequences**

351 To assess whether the secondary structure composition of precursor amino acid sequences
352 influences the predicted quality of convergent overlaps, the proportion of coil (C), β-sheet (E),
353 and α-helix (H) structures in each input was compared against the model's predicted secondary
354 structure score. For each amino acid pair, structural proportions were calculated separately and
355 then averaged across both sequences. This approach reflects the biological premise that overlap
356 feasibility arises not from a single sequence in isolation, but from the combined structural
357 constraints and flexibilities of the two proteins that must be co-encoded (60,61).

358 Across all three overlap phases, sequences with higher coil content exhibited elevated secondary
359 structure scores (**Fig. 3C**). This trend was most apparent in phase 0 and phase 1, where LOESS-
360 smoothed curves showed a modest increase in secondary structure score with increasing coil
361 content, followed by a sharper rise beyond approximately 70% coil. In phase 2, a similar increase
362 was observed at high coil levels, but the trend was less consistent, with greater variability at
363 intermediate values. By contrast, α-helix content was inversely associated with secondary
364 structure score. The most pronounced effect occurred in phase 0, where increasing helix content
365 corresponded to a marked decline in secondary structure scores. A similar, though less steep,
366 trend was observed in the other phases. For β-sheet content, no consistent relationship with

367 secondary structure score was observed; across all three phases, smoothed curves were largely
368 flat, indicating that sheet content neither strongly promoted nor hindered predicted structural
369 quality.

370 To evaluate whether these relationships extended to sequence-level similarity, mean alignment
371 scores were also plotted against precursor structural compositions. No meaningful trends were
372 observed for any secondary structure class (**Fig. S1**). Alignment scores remained relatively
373 constant regardless of coil, β-sheet, or α-helix content, suggesting that precursor structure did not
374 systematically influence the model's ability to recover the reference overlap sequence.

375 **Multi-objective optimization and computational design of EGFP/AmpR convergent**
376 **overlaps**

377 Although the transformer-encoder models were capable of generating convergent overlaps for
378 paired amino acid sequences, the secondary structure preservation achieved in initial predictions
379 was variable and generally lower than values reported in prior work (60). To address this
380 limitation, an iterative sequence optimization algorithm was implemented to refine model
381 outputs using a multi-objective framework that did not rely on MSA availability (**Fig. 4A**).

382 The optimization procedure employed a windowed feedforward Monte Carlo (MC) dropout
383 strategy to generate multiple candidate solutions for each sequence window (**Fig. 4B**). Candidate
384 overlaps were then evaluated against two complementary structural criteria: 1) predicted
385 secondary structure states derived from S4PRED (62), providing local fold-relevant constraints,
386 and 2) ESM-2 contact map similarity (63), which encodes sequence-level context that may
387 reflect long-range residue dependencies. A key feature of the framework was the ability to

enforce selective preservation of amino acid residues when required. This was achieved by applying a feedforward mask to specific codon positions, constraining the model to maintain defined residues (for example, active-site motifs) while permitting synonymous variation elsewhere. In practice, this allowed critical sequence features to be retained while enabling exploration of codon-level degeneracy to optimize overlap formation.

Relationships among the chosen objectives were next quantified using mutual information (MI) and principal component analysis (PCA). Both approaches indicated that alignment identity, ESM-2 contact map similarity, and predicted secondary structure capture overlapping but distinct constraints. Discrete MI (20 quantile bins) showed that alignment carried moderate information about both ESM-2 contact map similarity (MI = 0.393) and secondary structure match (MI = 0.381), whereas the MI between ESM-2 contact map similarity and secondary structure similarity was lower (MI = 0.240), consistent with partial independence. Continuous MI estimates (kNN regression) produced similar results (0.509, 0.508, and 0.340, respectively). All associations were highly significant in permutation tests ($p = 0.001$). PCA of z-scored metrics supported this interpretation (**Fig. S2**). A single axis (PC1) explained 72.4% of variance (95% CI: 71.8-73.0) and loaded positively on all three objectives, reflecting a general similarity dimension. PC2 (18.4%; 95% CI: 18.0-18.9) contrasted ESM-2 contact map similarity with secondary structure similarity, while PC3 (9.2%) primarily captured sequence-only variance. Thus, although alignment dominates the shared signal, secondary structure preservation and ESM-2 contact map-based similarity provide orthogonal information that cannot be reduced to alignment alone. This justifies their inclusion as separate objectives during overlap optimization.

409 Metric behavior under progressive sequence divergence was also assessed by subjecting 250
410 randomly sampled SwissProt proteins to iterative single-residue substitution trajectories (75 steps
411 per sequence). Across trajectories, all four similarity measures declined with increasing
412 mutational distance, but the rate and pattern of decay differed by metric (**Fig. S3**). Alignment
413 identity and BLOSUM62 similarity decreased rapidly and near-linearly, consistent with their
414 direct dependence on residue-level identity. By contrast, ESM-2 contact map similarity and
415 secondary structure similarity showed more gradual declines. These distinct decay profiles
416 support the interpretation that the objectives capture complementary aspects of overlap fidelity.

417 Weighting experiments further clarified the contributions of each objective. When the relative
418 weights assigned to secondary structure similarity, ESM-2 contact map similarity, alignment
419 identity, and BLOSUM62 substitution similarity were varied across randomly selected SwissProt
420 pairs with very high AlphaFold2 pLDDT values (≥90), distinct trade-off dynamics were
421 observed (**Fig. S4**). Increased weighting on secondary structure similarity consistently improved
422 secondary structure preservation but reduced alignment and substitution similarity, whereas
423 embedding-focused weightings enhanced ESM-2 contact map similarity at the expense of
424 secondary structure similarity. Intermediate combinations (eg, 0.15-0.4 secondary structure with
425 0.6 ESM-2 contact map similarity) yielded more balanced outcomes across the four metrics.
426 Variance patterns indicated that secondary structure weighting stabilized optimization
427 trajectories, while ESM-2 contact map similarity remained comparatively variable across
428 candidate sequences.

429 A convergent overlap between enhanced green fluorescence protein (EGFP) and a TEM-1 β-
430 lactamase (ampicillin resistance marker from pCVD004; hereafter AmpR) was designed and

431 refined under the same objectives as a biologically relevant test case (64,65). The combination of
432 MC dropout sampling, secondary structure prediction, and embedding-based scoring consistently
433 improved fold preservation of computationally designed EGFP/AmpR 311 nucleotide (phase 1)
434 overlaps (**Fig. 4C; Fig. 5**). Compared to unoptimized outputs (ie, sequences generated from the
435 first window), optimized sequences displayed higher secondary structure similarity rates and
436 increased ESM-2 contact map similarity, demonstrating that overlap designs can be improved
437 using lightweight, MSA-independent objectives. During windowed optimization, alignment and
438 BLOSUM62 metrics remained largely constant or decreased across windows, reflecting their
439 role in constraining candidate sequences to a biologically plausible subset of sequence space. In
440 contrast, secondary structure similarity generally increased throughout optimization, despite
441 being weighted at only 0.15 in the composite score. ESM-2 contact map similarity increased
442 modestly in the early windows before plateauing. These findings suggest that fold-relevant
443 secondary structure metrics may act as the principal discriminative driver of optimization
444 progress, with alignment- and embedding-based metrics providing a stabilizing baseline.
445 Following optimization, a pair of amino acids with the highest combined score was selected for
446 structural comparison using predictions from AlphaFold3 (50), while selecting amino acid
447 sequences in both pairs for preservation based on predicted importance (eg, active site). The
448 AlphaFold3-predicted models of designed overlaps were highly similar to reference predictions
449 (**Fig. 4C**), with TM-scores of 0.98 (EGFP) and 0.90 (AmpR), the latter with deviations generally
450 localized to a disordered coil not located in the overlapping terminal region (66).

451 **Performance of this algorithm in predicting overlapping genes using pLDDT as an**
452 **orthogonal proxy for intrinsic structure rigidity**

453 Analyses on prokaryote-derived amino acid sequence pairs indicated that secondary structure
454 preservation increased with coil content and decreased with helix content, with sheet content
455 showing no consistent effect. To determine whether this pattern reflected a generalizable
456 property of overlap design for convergent overlap derived from terminal amino acid sequences,
457 an external validation set was assembled from SwissProt proteins stratified by AlphaFold2 (AF2)
458 per-residue confidence (pLDDT), an orthogonal proxy for intrinsic structural rigidity (52,67–69).
459 SwissProt sequences were filtered to 105–200 amino acids to align with the 105-AA receptive
460 field and to bound computational cost, then binned into Low-intermediate (50-69), Confident
461 (70-89), and Very high (90-100) pLDDT groups. For each bracket, 100 non-redundant amino-
462 acid pairs were formed, and convergent overlaps were designed across phases 0, 1, and 2 at 310,
463 311, and 312 nt using the same transformer-encoder, windowed Monte-Carlo (MC) dropout
464 inference, objective weights, and constraint handling described previously.

465 Quantitative analysis of secondary structure composition (determined using S4PRED) across
466 pLDDT confidence bins revealed clear trends. Looking specifically at the terminal AA regions,
467 as pLDDT confidence increased from the Low-intermediate to the Very high bins, coil content
468 declined substantially (from ~64% to ~41%), while both helix and sheet content rose (helix from
469 ~29% to ~39%; sheet from ~6% to ~20%) (**Figs. 6A, 6B**). These shifts suggest that AlphaFold's
470 higher-confidence regions correspond generally to more ordered structural states, and reinforce
471 the interpretation of pLDDT as an approximate proxy for structural order and rigidity (50,70,71).

472 These AF2-stratified findings were consistent with the earlier composition-based analysis using
473 prokaryotic ORFs in this study. Sequence pairs with lower pLDDT (coil-enriched) were
474 associated with slightly higher secondary structure preservation, whereas higher-confidence

475 inputs tended to show reduced variation (**Fig. 6C**). Nevertheless, the overall high secondary
476 structure scores across all three brackets emphasize that the algorithm was able to recover
477 structurally faithful overlaps even for proteins predicted by AlphaFold2 to adopt stable, well-
478 defined folds. Alignment identity and substitution similarity remained generally stable across
479 brackets, indicating that secondary structure preservation was the primary dimension along
480 which flexibility exerted its effect.

481 Bracket- and phase-specific effects were evident across all three metrics, consistent with the
482 secondary structure composition of the pLDDT bins. Across overlap phases, sequences in the
483 Low-intermediate (50-69) bracket generally outperformed those in the Confident (70-89) and
484 Very high (90-100) brackets, with the gap being most pronounced in phase 2 (**Fig. 6D**). In phase
485 1, the bracket effect was clear for secondary structure preservation and for the combined score
486 relative to Very high (90-100), whereas pairwise differences were not evident for ESM-2 contact
487 map similarity; Low-intermediate (50-69) also did not differ from Confident (70-89) for the
488 combined score. Overall, phase 1 maintained the highest preservation, while phase 2 exhibited
489 the largest bracket separations. Although group differences were statistically significant, effect
490 sizes varied by metric: within a phase, combined and ESM-2 contact map similarity differences
491 were typically ~1-4 percentage points, whereas secondary structure differences reached ~5-8
492 points in some phase 2 contrasts. All groups maintained mean secondary structure preservation
493 above 85%. Full pairwise estimates with confidence intervals and adjusted p-values are provided
494 in **Supplementary Tables 4 and 5**.

495 While statistically significant differences were observed between pLDDT brackets, these
496 differences were relatively small, and overall secondary structure preservation remained

consistently high across all conditions (**Fig. 6E**). These results reinforce the generalizability of the earlier coil-associated trend, but also highlight the robustness of the multi-objective, uncertainty-aware optimization framework in achieving high predicted structural preservation regardless of precursor rigidity.

Discussion

Overlapping gene architectures have long been recognized as both a source of evolutionary constraint and an avenue for novel functionality (72–80). In addition to their evolutionary importance, overlapping structures can be purposefully engineered under realistic codon and structural constraints. The present study shows that phase-specific convergent overlaps can be computationally generated and optimized using a transformer encoder-based multi-objective framework without requiring full 3D prediction at the optimization stage; targeted 3D checks were applied post hoc. This framework offers a complementary tool that could expand the range of strategies available for synthetic overlap design, particularly where phase, length, and nucleotide-level control are critical.

Recent work has demonstrated that deep generative protein models can discover viable overlapping solutions under genetic code constraints. Byeon et al. (2025) showed that pretrained amino acid-space generative models can design synthetic overlapping genes across multiple reading frames, validating expression experimentally using Gibbs sampling with codon-compatibility constraints (30). Xu et al. (2025) further demonstrated that pretrained generative protein language models (ESM-3) can yield viable entangled protein pairs through structure-conditioned inverse folding and CAMEOS-based entanglements, filtered by cross-entropy and Potts energy scores (81). Importantly, Xu et al. established that pretrained generative models are

519 capable of identifying functional overlap-compatible solutions, albeit in a limited experimental
520 context (the InfA/AroB system) and without explicit codon-level optimization.

521 The framework presented here builds on this momentum but takes a different route. Rather than
522 relying on pretrained amino acid-space sampling, it employs a purpose-built, phase-generalizable
523 transformer encoder model set trained on a balanced synthetic dataset of convergent overlaps
524 spanning diverse GC contexts and overlap lengths. Multi-objective optimization is integrated
525 directly into the inference loop, with candidate overlaps iteratively refined through a windowed,
526 uncertainty-aware (MC dropout) search. Evaluation criteria span predicted secondary structure
527 preservation (S4PRED), amino acid substitution similarity (BLOSUM62), pairwise alignment
528 identity, and ESM-2 contact map similarity. This approach emphasizes codon-level control while
529 maintaining structural fidelity, enabling overlap recovery without requiring full 3D structural
530 prediction at each iteration.

531 MC dropout was retained during inference to enable stochastic sampling of synonymous codon
532 solutions, thereby broadening the effective search space (35). Candidate overlaps were ranked
533 using a fixed composite score that balanced structural preservation, substitution similarity,
534 alignment identity, and embedding-based metrics. This weighting scheme served as a practical
535 approach for navigating competing design constraints without requiring full structural prediction
536 at each iteration, and was most closely aligned with prior work arguing that synthetic overlap
537 design requires explicit balancing of competing biological constraints and objectives (5,60).
538 While other multi-objective optimization strategies could be considered in future work (such as
539 ε-constraint methods (82,83), adaptive scalarization (84,85), or evolutionary multi-objective

optimizers (86)), the fixed weighted framework proved sufficient to achieve high overlap fidelity under the conditions evaluated.

The synthetic convergent overlapping gene dataset construction method developed as part of this study provides a structured framework for exploring overlapping gene design in silico, as it permits generation of diverse and uniformly distributed overlaps. However, one potential limitation of this approach is the reliance on synthetic, rather than natural, overlapping genes for model training, which were generated using ORFs from a diverse set of prokaryotic strains. This may introduce a bias affecting the predictive ability and variability in sequence output. Despite this concern, the method employed to generate synthetic overlapping genes is in principle analogous to the formation of new overlaps through the process of gene extension following the loss of a stop codon (8). This results in synthetic overlapping genes where the second (extended) strand's encoded amino acid composition is derived from the reverse complement of the first strand, without undergoing further selection. As recently reported, there is extensive evidence suggesting functionality in prokaryotic convergent (antisense) proteins, indicating that non-coding RNA regions could also encode functional proteins (87). For same-strand overlaps, a significant difference in overall composition compared to non-overlapping genes has been previously observed (88), along with a bias towards disorder-promoting amino acids (89). The extent to which this affects convergent overlapping genes remains unclear. Future work will therefore aim to (i) characterize compositional differences between natural and synthetic overlaps, (ii) extend design to longer proteins beyond the 199-312 nucleotide range considered here, and (iii) experimentally validate whether in silico-designed overlaps from this framework maintain expression, folding, and function in vivo.

In summary, this study establishes that phase-aware transformer models trained directly at the nucleotide level, when coupled with uncertainty-aware inference and multi-objective optimization, can recover computationally designed convergent overlapping genes under realistic genomic constraints. By integrating codon-level sampling with structural and substitution-based objectives, this framework complements recent amino acid-space generative approaches while offering length- and phase-specific control. These findings provide a computational foundation for experimentally testable designs and point toward a scalable strategy for compact, synthetic gene architectures.

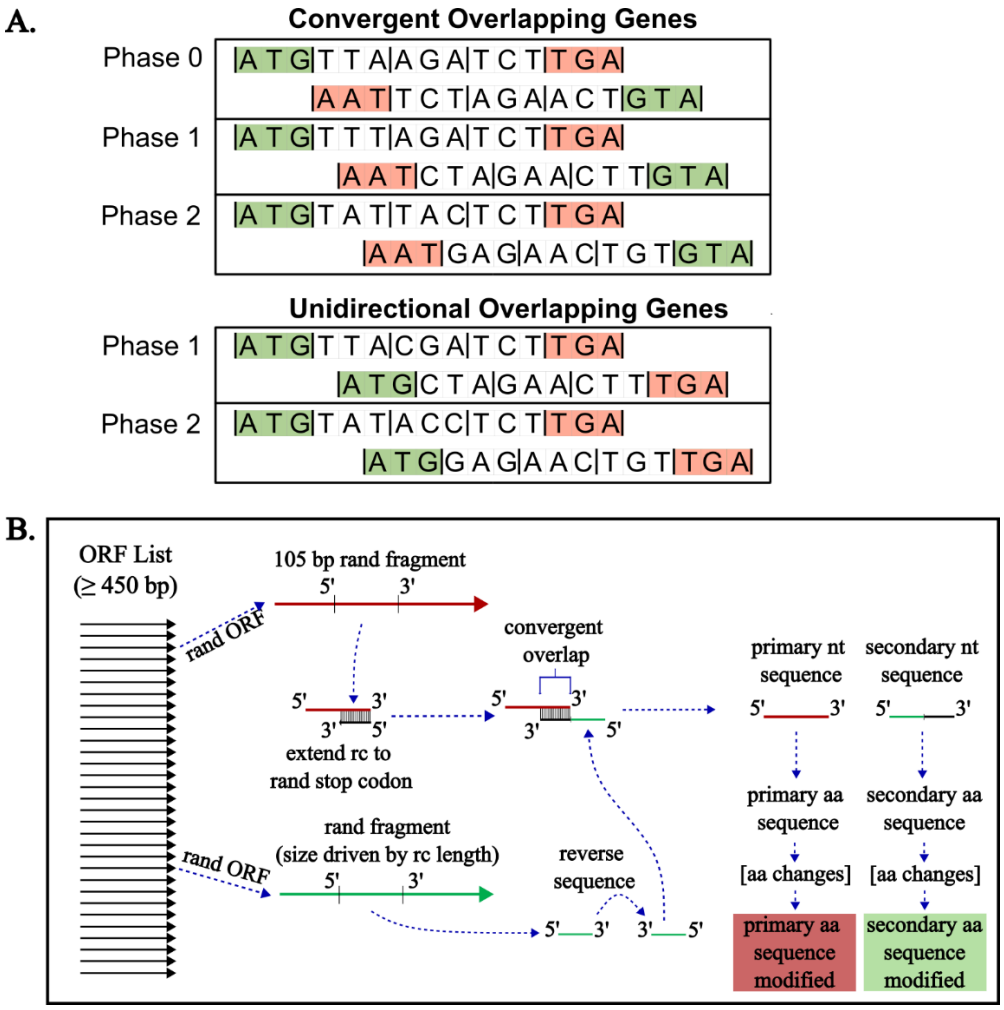

Fig 1. Depiction of convergent overlapping genes in three reading frames and the process developed to computationally design synthetic convergent overlaps from coding sequences.

(A) Sequences are represented for convergent (tail-to-tail) and unidirectional (same-strand) overlapping genes in each of the three reading frames (phase 0, phase 1, and phase 2). Note, phase 0 unidirectional overlaps are not represented as these are considered as an alternative start site for the same gene (90). In each phase, the reading frame was shifted by one nucleotide. Start and stop codons are indicated with green and red shading, respectively. (B) The general process for creating synthetic convergent overlapping genes is presented, starting with a pool of ORF sequences filtered by length (≥450 base pair [bp]). The primary sequence and in-frame 105-bp fragment were randomly selected. A randomly selected stop codon from the reverse complement (rc), and the fragment reverse complement was extended to that stop codon to serve as the downstream sequence of the secondary sequence. Given the downstream secondary sequence length, another sequence was randomly selected from the same gene list. A size appropriate fragment was selected from that gene with length equal to the difference between the desired gene length and the upstream overlap fragment (ie, the reverse complement of the primary sequence to the selected stop codon). The upstream fragment was then combined with the downstream fragment to generate the complete secondary sequence. The resultant primary and

secondary nucleotide sequences, sharing a known convergent overlapping region, were then translated to amino acid (aa) sequences (either modified or unmodified) for subsequent use in model training or inference.

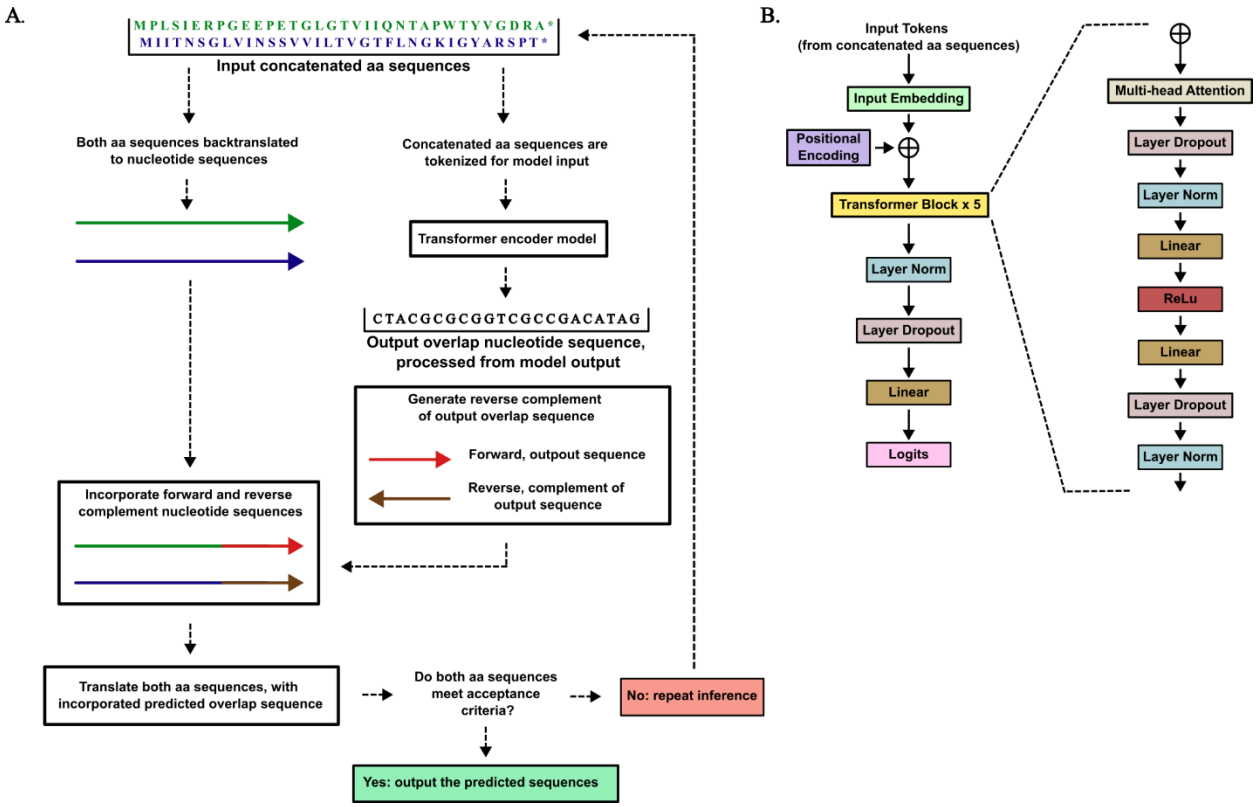

Fig 2. Illustration of the overlap transformer-encoder architecture and general structure of the data flow

(A) Depiction of the general structure of the data flow, for two input concatenated amino acid sequences (primary and secondary) used during inference. The primary and secondary amino acid sequences are represented with green and blue text, respectively, and the flow proceeds from top to bottom. (B) Illustration of the transformer encoder model architecture. The model input is tokenized concatenated paired amino acid sequences, and the model outputs raw logits. The output logits are processed using a greedy decoder to generate a predicted overlap sequence.

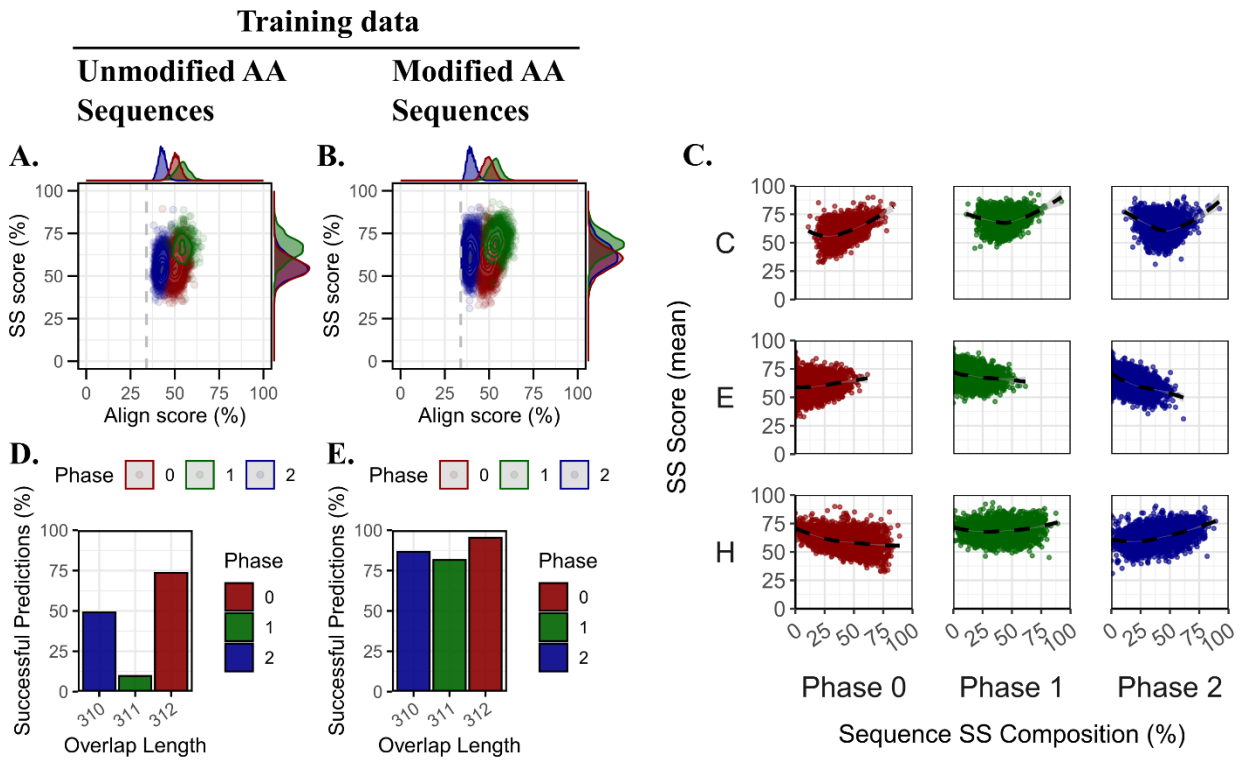

Fig 3. Training data evaluation of secondary structure preservation and convergent overlap-phase performance.

(A, D) Results for unmodified amino acid sequences. (A) Scatter density plot of secondary structure (SS) score versus alignment score, with contours and marginal density plots showing the distribution by overlap phase (0 = red, 1 = green, 2 = blue). (D) Fraction of successful predictions across overlap lengths (310, 311, and 312 nucleotides), stratified by phase. Success rates are reported with exact 95% Clopper–Pearson confidence intervals. (B, E) Results for modified amino acid sequences. (B) Scatter density plot of SS score versus alignment score with the same phase-coloring scheme. (E) Fraction of successful predictions across overlap lengths for modified sequences, stratified by phase. Success rates are reported with exact 95% Clopper–Pearson confidence intervals. (C) Relationship between sequence-level SS composition (coil [C], sheet [E], helix [H]) and mean SS score across phases 0, 1, and 2. Dashed lines indicate LOESS regression fits with 99% confidence bands.

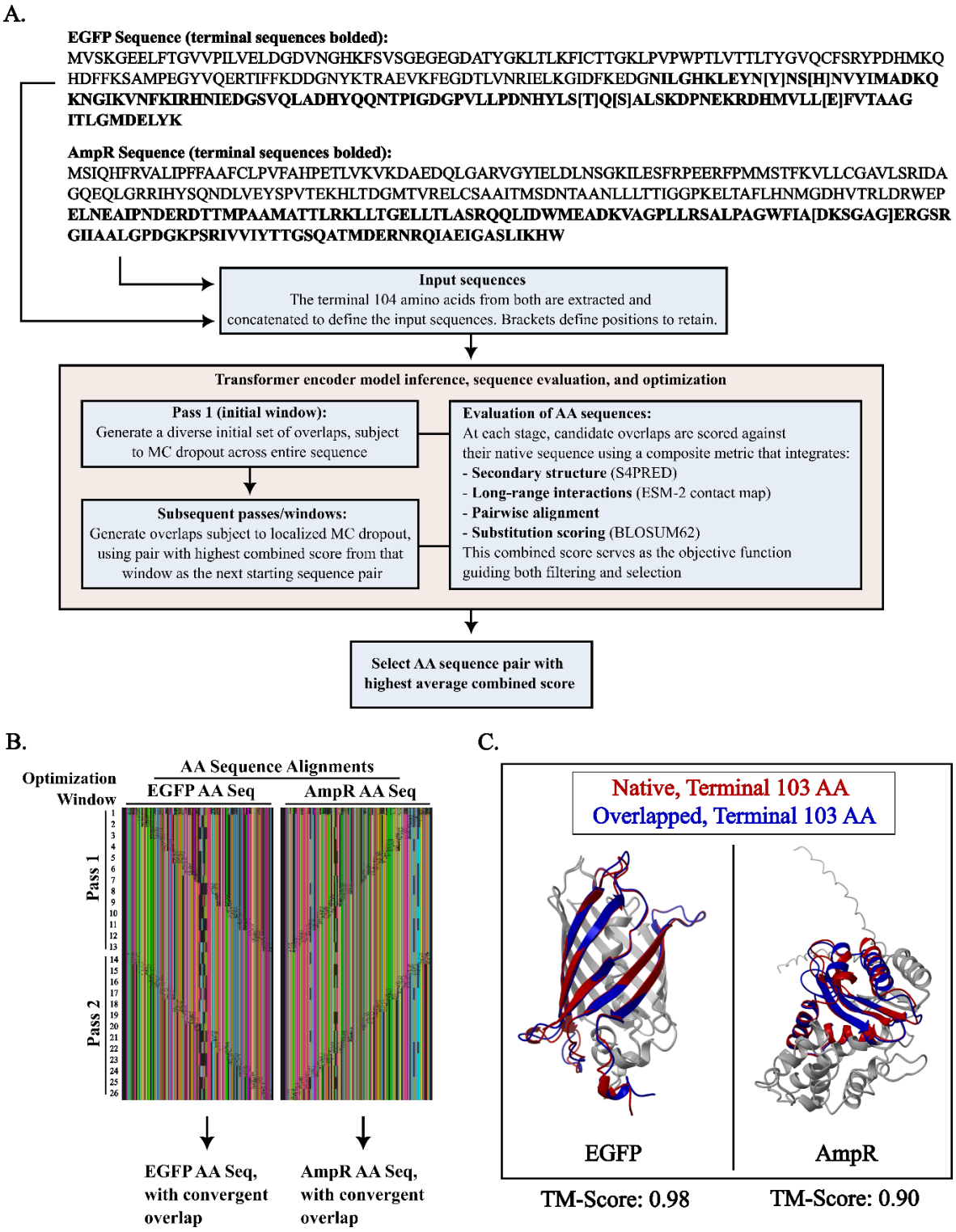

Fig 4. In silico design of convergent overlap with EGFP and AmpR

Overview of the multi-objective optimization process for EGFP and AmpR proteins using the transformer encoder framework. (A) Schematic of the transformer encoder–based workflow. The terminal 104 amino acids of EGFP and AmpR (across all tested overlap lengths) are concatenated and passed through the model. Outputs are scored by a composite metric integrating alignment, substitution, secondary structure preservation, and long-range interactions. Bracketed amino acids from the input are preserved to enforce positional constraints. The final sequence pair is selected from the output table by the highest average combined score. (B) Multiple-sequence alignment of all output pairs (n=3,624) for EGFP and AmpR across both optimization passes and windows (top to bottom). Diagonal dropout patterns correspond to localized MC dropout applied to the overlap region. Variability reflects bracket-constrained codon degeneracy introduced on the non-fixed partner sequence. (C) Predicted tertiary structures generated with AlphaFold3 for EGFP (left) and AmpR (right). Native and overlapping terminal regions relevant to the 311-nt overlap are highlighted (native in red, overlapped in blue); non-overlapping regions are shown in gray.

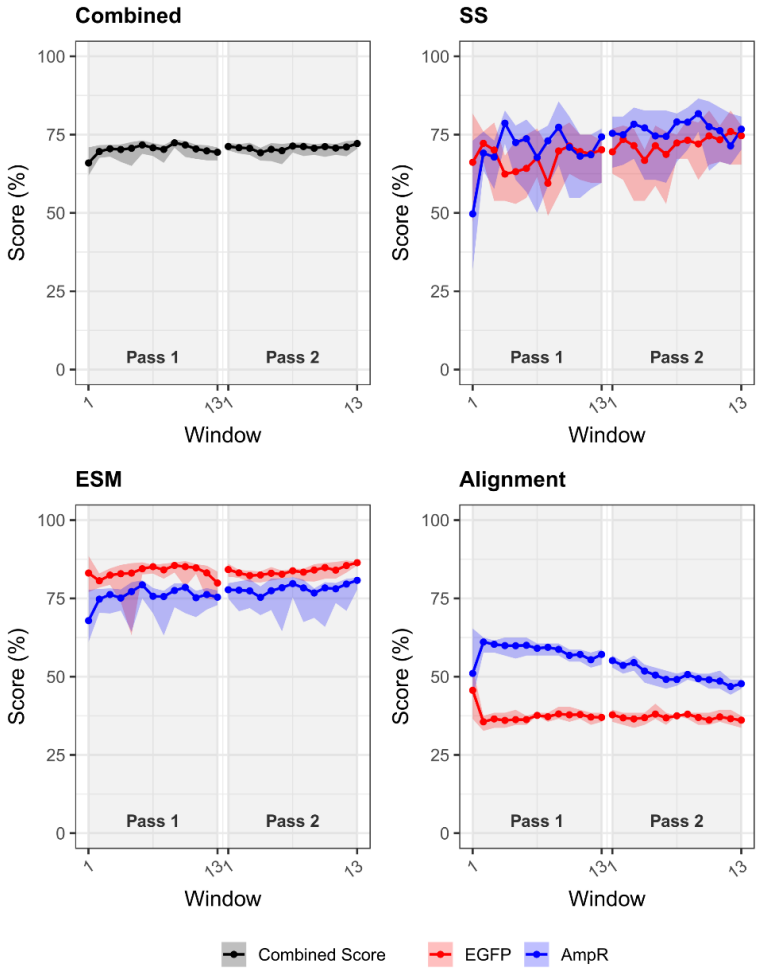

Fig 5. Trajectories of composite and component metrics during computational optimization of EGFP/AmpR convergent overlaps

Optimization performance for EGFP and AmpR across two passes, each consisting of 13 windows. Metrics shown include the combined score (top left), which integrates secondary structure (SS), ESM-2 contact map similarity, and alignment components, as well as the individual SS (top right), ESM-2 (bottom left), and amino acid alignment (bottom right) scores. Solid red (EGFP) and blue (AmpR) lines denote mean values, and shaded regions represent the minimum-maximum range across sequences within each window.

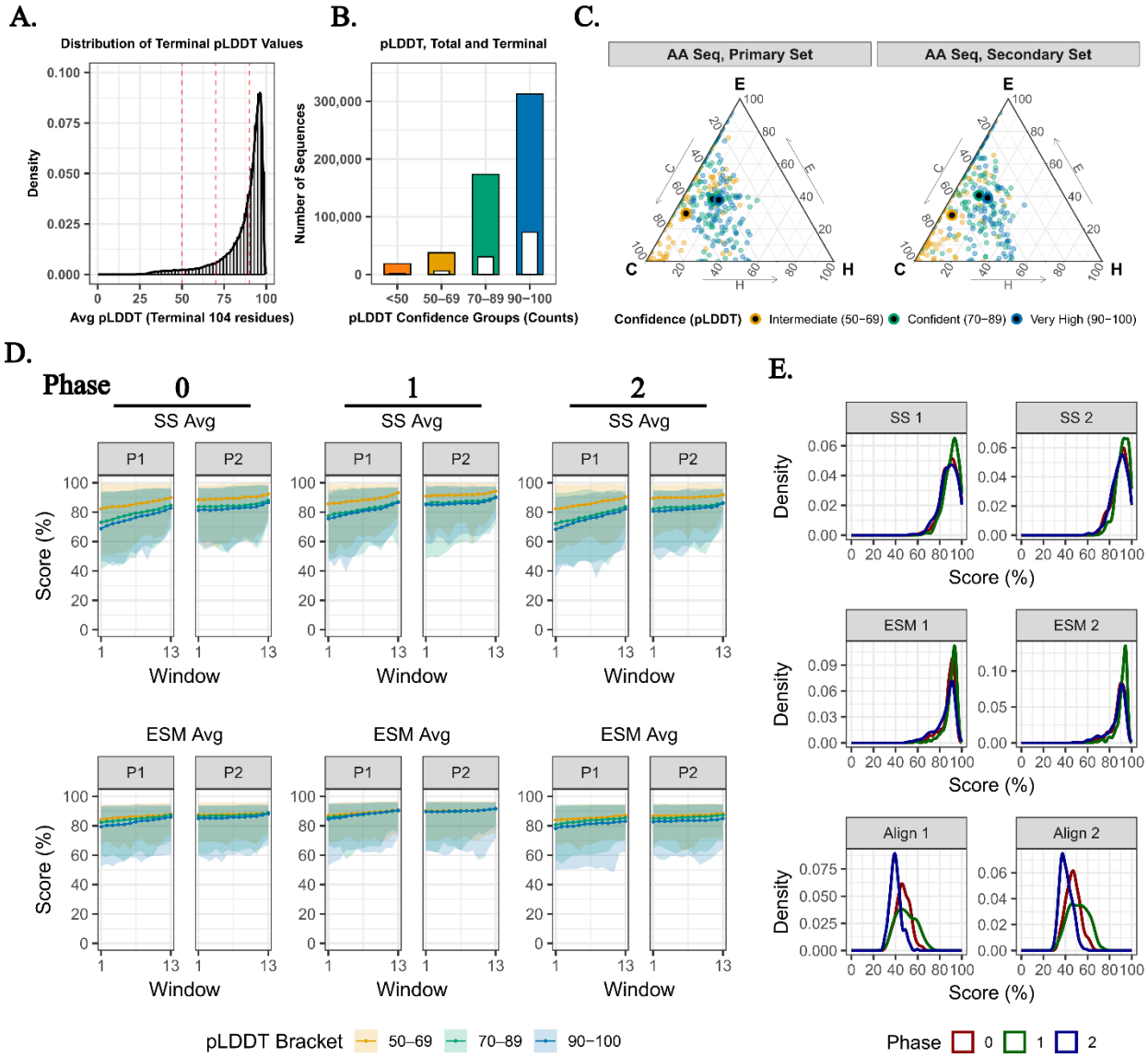

Fig 6. Comparison of computational overlap predictions across pLDDT brackets, with structural composition and optimization outcomes

Analyses are based on random sampling of SwissProt protein pairs (n=100 for each bracket) from the AlphaFold2 Protein Structure Database (AF2). (A) Distribution of average pLDDT values for the terminal 104 residues across sampled sequences, with red dashed lines marking the confidence thresholds. (B) Counts of sequences assigned to pLDDT confidence brackets, shown for both total and terminal regions (internal white bars denote terminal regions). (C) Ternary plots of amino acid composition (helix [H], sheet [E], and coil [C]) for the primary and secondary sequence sets, colored by pLDDT bracket (low-intermediate, confident, and very high). Circles with black centers denote the centroid (mean position of all sequences within a given pLDDT bracket) in ternary space. (D) Optimization performance across convergent overlap phases (0, 1, 2) for Pass 1 (P1) and Pass 2 (P2). Plots show average secondary structure (SS Avg) and ESM-2 contact map similarity (ESM Avg) scores by optimization window, with shaded regions representing the range of values stratified by pLDDT bracket. (E) Density distributions of individual metric scores (SS, ESM, and Alignment) for the highest combined score sequences (primary and secondary) across phases 0, 1, and 2.

References

1. Barrell BG, Air GM, Hutchison CA. Overlapping genes in bacteriophage φX174. Nature. 1976 Nov;264(5581):34–41.

2. Keese PK, Gibbs A. Origins of genes: “big bang” or continuous creation? Proc Natl Acad Sci USA. 1992 Oct 15;89(20):9489–93.

3. Wright BW, Molloy MP, Jaschke PR. Overlapping genes in natural and engineered genomes. Nat Rev Genet. 2022 Mar;23(3):154–68.

4. Grainger DC. The unexpected complexity of bacterial genomes. Microbiology. 2016 July 1;162(7):1167–72.

5. Blazejewski T, Ho HI, Wang HH. Synthetic sequence entanglement augments stability and containment of genetic information in cells. Science. 2019 Aug 9;365(6453):595–8.

6. Opuu V, Silvert M, Simonson T. Computational design of fully overlapping coding schemes for protein pairs and triplets. Sci Rep. 2017 Nov 20;7(1):15873.

7. Sabath N, Graur D, Landan G. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. Biology Direct. 2008;3(1):36.

682 8. Kingsford C, Delcher AL, Salzberg SL. A Unified Model Explaining the Offsets of
683 Overlapping and Near-Overlapping Prokaryotic Genes. Molecular Biology and Evolution.
684 2007 Sept;24(9):2091–8.

685 9. Pallejà A, Harrington ED, Bork P. Large gene overlaps in prokaryotic genomes: result of
686 functional constraints or mispredictions? BMC Genomics. 2008;9(1):335.

687 10. Zehentner B, Ardern Z, Kreitmeier M, Scherer S, Neuhaus K. Evidence for Numerous
688 Embedded Antisense Overlapping Genes in Diverse *E. coli* Strains [Internet]. 2020 [cited
689 2024 July 12]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.11.18.388249

690 11. Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Scherer S, Neuhaus K. The Novel
691 Anaerobiosis-Responsive Overlapping Gene ano Is Overlapping Antisense to the Annotated
692 Gene ECs2385 of Escherichia coli O157:H7 Sakai. Front Microbiol. 2018 May 14;9:931.

693 12. Vanderhaeghen S, Zehentner B, Scherer S, Neuhaus K, Ardern Z. The novel EHEC gene asa
694 overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase.
695 Sci Rep. 2018 Dec 14;8(1):17875.

696 13. Zehentner B, Ardern Z, Kreitmeier M, Scherer S, Neuhaus K. A Novel pH-Regulated,
697 Unusual 603 bp Overlapping Protein Coding Gene pop Is Encoded Antisense to ompA in
698 Escherichia coli O157:H7 (EHEC). Front Microbiol. 2020 Mar 20;11:377.

699 14. Behrens M, Sheikh J, Nataro JP. Regulation of the Overlapping *pic/set* Locus in *Shigella*
700 *flexneri* and Enteroaggregative *Escherichia coli*. Infect Immun. 2002 June;70(6):2915–25.

701 15. Balabanov VP, Kotova VYu, Kholodii GY, Mindlin SZ, Zavilgelsky GB. A novel gene, *ardD*
702 , determines antirestriction activity of the non-conjugative transposon Tn *5053* and is located
703 antisense within the *tniA* gene. FEMS Microbiol Lett. 2012 Dec;337(1):55–60.

704 16. Delaye L, DeLuna A, Lazcano A, Becerra A. The origin of a novel gene through overprinting
705 in Escherichia coli. BMC Evol Biol. 2008;8(1):31.

706 17. Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, et al. Phenotype of
707 *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella* , completely
708 overlapping in antisense to *yaaW*. FEMS Microbiol Lett. 2014 Jan;350(1):57–64.

709 18. Graf F, Zehentner B, Fellner L, Scherer S, Neuhaus K. Three Novel Antisense Overlapping
710 Genes in E. coli O157:H7 EDL933. Gilk SD, editor. Microbiol Spectr. 2023 Feb
711 14;11(1):e02351-22.

712 19. Cock PJA, Whitworth DE. Evolution of Gene Overlaps: Relative Reading Frame Bias in
713 Prokaryotic Two-Component System Genes. J Mol Evol. 2007 Apr;64(4):457–62.

714 20. Cock PJA, Whitworth DE. Evolution of Relative Reading Frame Bias in Unidirectional
715 Prokaryotic Gene Overlaps. Molecular Biology and Evolution. 2010 Apr 1;27(4):753–6.

21. Pavesi A. Prediction of two novel overlapping ORFs in the genome of SARS-CoV-2. Virology. 2021 Oct;562:149–57.

22. Nelson CW, Ardern Z, Wei X. OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes. Tamura K, editor. Molecular Biology and Evolution. 2020 Apr 3;msaa087.

23. Schlub TE, Buchmann JP, Holmes EC. A Simple Method to Detect Candidate Overlapping Genes in Viruses Using Single Genome Sequences. Malik H, editor. Molecular Biology and Evolution. 2018 Oct 1;35(10):2572–81.

24. Chlebek JL, Leonard SP, Kang-Yun C, Yung MC, Ricci DP, Jiao Y, et al. Prolonging genetic circuit stability through adaptive evolution of overlapping genes. Nucleic Acids Research. 2023 July 21;51(13):7094–108.

25. Decrulle AL, Frénoy A, Meiller-Legrand TA, Bernheim A, Lotton C, Gutierrez A, et al. Engineering gene overlaps to sustain genetic constructs in vivo. Braun EL, editor. PLoS Comput Biol. 2021 Oct 8;17(10):e1009475.

26. Wang B, Papamichail D, Mueller S, Skiena S. Two Proteins for the Price of One: The Design of Maximally Compressed Coding Sequences. In: Carbone A, Pierce NA, editors. DNA Computing [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006 [cited 2024 July 10]. p. 387–98. (Lecture Notes in Computer Science; vol. 3892). Available from: http://link.springer.com/10.1007/11753681_31

27. Manuel Martí J, Hsu C, Rochereau C, Xu C, Blazejewski T, Nisonoff H, et al. GENTANGLE: integrated computational design of gene entanglements. Elofsson A, editor. Bioinformatics. 2024 June 21;btae380.

28. Wichmann S, Scherer S, Ardern Z. Biological factors in the synthetic construction of overlapping genes. BMC Genomics. 2021 Dec;22(1):888.

29. Leonard SP, Halvorsen T, Lim B, Park DM, Jiao Y, Yung M, et al. Creating overlapping genes by alternate-frame insertion [Internet]. 2024 [cited 2024 Nov 29]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.11.07.622342

30. Byeon GW, Expòsit M, Baker D, Seelig G. Design of overlapping genes using deep generative models of protein sequences [Internet]. Synthetic Biology; 2025 [cited 2025 Aug 8]. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.05.06.652464

31. Choi SR, Lee M. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology. 2023 July 22;12(7):1033.

32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arXiv; 2023 [cited 2023 Oct 18]. Available from: http://arxiv.org/abs/1706.03762

33. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. AI Open. 2022;3:111–32.

34. Srivastava, N, Hinton, G, Krizhevsky, A, Sutskever, I, Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. JMLR. 2014;15(56):1929−1958.

35. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of The 33rd International Conference on Machine Learning. 2016;(48):pp 1050-1059.

36. Lemay A, Hoebel K, Bridge CP, Befano B, De Sanjosé S, Egemen D, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. npj Digit Med. 2022 Nov 18;5(1):174.

37. Seoh R. Qualitative Analysis of Monte Carlo Dropout [Internet]. arXiv; 2020 [cited 2024 July 13]. Available from: https://arxiv.org/abs/2007.01720

38. Singer GAC, Hickey DA. Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. Molecular Biology and Evolution. 2000 Nov 1;17(11):1581–8.

39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Internet]. arXiv; 2019 [cited 2024 July 13]. Available from: https://arxiv.org/abs/1912.01703

40. TensorFlow Developers. TensorFlow [Internet]. Zenodo; 2023 [cited 2024 Feb 24]. Available from: https://zenodo.org/doi/10.5281/zenodo.10126399

41. Zarrieß S, Voigt H, Schüz S. Decoding Methods in Neural Language Generation: A Survey. Information. 2021 Aug 30;12(9):355.

42. Graczyk KM, Pawłowski J, Majchrowska S, Golan T. Self-normalized density map (SNDM) for counting microbiological objects. Sci Rep. 2022 June 22;12(1):10583.

43. Subramanian K, Payne B, Feyertag F, Alvarez-Ponce D. The Codon Statistics Database: A Database of Codon Usage Bias. Saitou N, editor. Molecular Biology and Evolution. 2022 Aug 3;39(8):msac157.

44. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 June 1;25(11):1422–3.

45. Moffat L, Jones DT. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. Xu J, editor. Bioinformatics. 2021 Nov 5;37(21):3744–51.

46. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577–637.

47. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992 Nov 15;89(22):10915–9.

48. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004 Aug;22(8):1035–6.

49. Jia K, Jernigan RL. New amino acid substitution matrix brings sequence alignments into agreement with structure matches. Proteins. 2021 June;89(6):671–82.

50. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 June 13;630(8016):493–500.

51. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 June;19(6):679–82.

52. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

53. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123–30.

54. Singh J, Litfin T, Singh J, Paliwal K, Zhou Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. Martelli PL, editor. Bioinformatics. 2022 Mar 28;38(7):1888–94.

55. Zhou Wang, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans on Image Process. 2004 Apr;13(4):600–12.

56. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York, NY: Springer New York; 2009 [cited 2024 Mar 22]. Available from: https://link.springer.com/10.1007/978-0-387-98141-3

57. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS. 2019 Nov 21;4(43):1686.

58. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bateman A, editor. Bioinformatics Advances. 2023 Jan 5;3(1):vbad001.

59. Rost B. Twilight zone of protein sequence alignments. Protein Engineering, Design and Selection. 1999 Feb;12(2):85–94.

60. Wichmann S, Scherer S, Ardern Z. Biological factors in the synthetic construction of overlapping genes. BMC Genomics. 2021 Dec;22(1):888.

61. Lebre S, Gascuel O. The combinatorics of overlapping genes. 2016 [cited 2023 Dec 29]; Available from: https://arxiv.org/abs/1602.04971

62. Moffat L, Jones DT. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. Xu J, editor. Bioinformatics. 2021 Nov 5;37(21):3744–51.

63. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123–30.

64. Zhang G, Gurtu V, Kain SR. An Enhanced Green Fluorescent Protein Allows Sensitive Detection of Gene Transfer in Mammalian Cells. Biochemical and Biophysical Research Communications. 1996 Oct;227(3):707–11.

65. Taton A, Unglaub F, Wright NE, Zeng WY, Paz-Yepes J, Brahamsha B, et al. Broad-host-range vector system for synthetic biology and biotechnology in cyanobacteria. Nucleic Acids Research. 2014 Sept 29;42(17):e136–e136.

66. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004 Dec;57(4):702–10.

67. Alderson TR, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. Proc Natl Acad Sci USA. 2023 Oct 31;120(44):e2304302120.

68. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021 Aug 26;596(7873):590–6.

69. Binder JL, Berendzen J, Stevens AO, He Y, Wang J, Dokholyan NV, et al. AlphaFold illuminates half of the dark human proteins. Current Opinion in Structural Biology. 2022 June;74:102372.

70. Abbas U, Chen J, Shao Q. Assessing Fairness of AlphaFold2 Prediction of Protein 3D Structures [Internet]. Bioinformatics; 2023 [cited 2025 Sept 14]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.05.23.542006

71. Guo HB, Perminov A, Bekele S, Kedziora G, Farajollahi S, Varaljay V, et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. Sci Rep. 2022 June 23;12(1):10696.

72. Ardern Z. Alternative Reading Frames are an Underappreciated Source of Protein Sequence Novelty. J Mol Evol. 2023 Oct;91(5):570–80.

73. Cassan E, Arigon-Chifolleau AM, Mesnard JM, Gross A, Gascuel O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. Proc Natl Acad Sci USA. 2016 Oct 11;113(41):11537–42.

74. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. Proc R Soc B. 2010 Dec 22;277(1701):3809–17.

75. Pavesi A. Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review. Genes. 2021 May 26;12(6):809.

76. Pavesi A. Asymmetric evolution in viral overlapping genes is a source of selective protein adaptation. Virology. 2019 June;532:39–47.

77. Pavesi A, Romerio F. Extending the Coding Potential of Viral Genomes with Overlapping Antisense ORFs: A Case for the De Novo Creation of the Gene Encoding the Antisense Protein ASP of HIV-1. Viruses. 2022 Jan 14;14(1):146.

78. Safari M, Jayaraman B, Yang S, Smith C, Fernandes JD, Frankel AD. Functional and structural segregation of overlapping helices in HIV-1. eLife. 2022 May 5;11:e72482.

79. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, et al. Functional Segregation of Overlapping Genes in HIV. Cell. 2016 Dec;167(7):1762-1773.e12.

80. Krakauer DC. Stability and Evolution of Overlapping Genes. Evolution. 2000 June;54(3):731–9.

81. Xu C, Chlebek JL, Allen JE, Nisonoff H, Park DM. Generative Protein Design for Overlapping Genes. In OpenReview; 2025. Available from: https://openreview.net/forum?id=35

82. On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. IEEE Trans Syst, Man, Cybern. 1971 July;SMC-1(3):296–7.

83. Miettinen K. Nonlinear Multiobjective Optimization [Internet]. Boston, MA: Springer US; 1998 [cited 2025 Sept 9]. (Hillier FS, editor. International Series in Operations Research & Management Science; vol. 12). Available from: http://link.springer.com/10.1007/978-1-4615-5563-6

84. Marler RT, Arora JS. Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization. 2004 Apr 1;26(6):369–95.

85. Das I, Dennis JE. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems. SIAM J Optim. 1998 Aug;8(3):631–57.

86. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Computat. 2002 Apr;6(2):182–97.

87. Ardern Z, Neuhaus K, Scherer S. Are Antisense Proteins in Prokaryotes Functional? Front Mol Biosci. 2020 Aug 14;7:187.

88. Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, et al. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. Jan E, editor. PLoS ONE. 2018 Oct 19;13(10):e0202513.

886 89. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping Genes Produce
887 Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation.
888 J Virol. 2009 Oct 15;83(20):10719–36.

889 90. Fonseca MM, Harris DJ, Posada D. Origin and Length Distribution of Unidirectional
890 Prokaryotic Overlapping Genes. G3 Genes|Genomes|Genetics. 2014 Jan 1;4(1):19–27.

891