Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive-Depth Recursion And State-Of-The-Art Efficient Sequence Models For The Oral Resistome: A Homology-Aware Benchmark With External Validation

Musab Hamed Saeed
Ajman University
Prabhu Manickam Natarajan
prabhuperio@gmail.com

Ajman University

Research Article

Keywords: oral resistome, antimicrobial resistance, protein language model, Mixture-of-Recursions, efficient Transformer, state-space model, external validation, homology-aware evaluation

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11227134/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Adaptive-Depth Recursion And State-Of-The-Art Efficient Sequence Models For The Oral Resistome: A Homology-Aware Benchmark With External Validation

Musab Hamed Saeed[1], Prabhu Manickam Natarajan[1]

1 Department of Clinical Sciences, College of Dentistry,Ajman University,Ajman; Centre of Medical and Bio-allied Health Sciences Research, Ajman, United Arab Emirates.

Email: m.saeed@ajman.ac.ae ORCID: 0000-0002-3564-184X

2 Department of Clinical Sciences, Center of Medical and Bioallied Health Sciences and Research, College of Dentistry, Ajman University, Ajman 346, United Arab Emirates.

Email: prabhuperio@gmail.com ORCID : 0000-0002-4780-0465

* Corresponding author: prabhuperio@gmail.com

Running head: Efficient sequence models for the oral resistome

Abstract

Background. Resistance genes form an abundant reservoir of resistance factors within the oral microbiome. Scalable attention alternatives may be able to annotate them; however, none has been evaluated for annotating them without learning from memorization of the data.

Methods. We curated 26,324 resistome proteins from UniProtKB, CARD, and NCBI's AMRFinderPlus datasets, clustered by 40% sequence similarity and 80% coverage. The proteins were partitioned into clusters. Eight causal residue-level architectures (Transformer, Funnel-Transformer, Performer, Hyena, causal state-space model, Mixture-of-Recursions MoR with a causal token choice router, parameter-matched MoR, and ablation of pairwise merge gating) were trained with identical data (up to 100 epochs with early stopping) for up to three random seeds and evaluated on held-out clusters and on independent databases with autoregressive causality audit. We used antibiotic class prediction with frozen representations as a measure of transferability.

Results. 26,324 proteins aggregated into 1,437 clusters (the largest being 3,333). On held-out clusters, all architectures outperformed the residue composition benchmark (perplexity 17.50),

with Hyena outperforming Transformer (14.93±0.58 vs 15.33); MoR (0.79M parameters) achieved 16.37, a sixth of the baseline value. On external validation (against the CARD dataset), only 4 of 8 architectures exceeded the composition benchmark perplexity of 17.99; the Transformer degrades by 27%; constrained architectures did better, with the MoR reaching 17.40, parameter-matched to the Transformer. Perplexity on external validation increased as similarity between the evaluation and training sets increased. MoR is 5.7x faster than Transformer on 4096 residues; the router uses 42% fewer block applications than the fixed depth.

Conclusions. Performance of architectures trained on the oral resistome critically depends on the way evaluation is done: homology-aware partitions and external validation reverse the order of architectures seen when evaluated on held-out clusters only. Recursive architectures perform best on external validation and are the least costly to compute, using less than 1/7th of the parameters compared to Transformers. Yet, for proteins without apparent homology to training sets, none of the eight architectures outperform residue composition.

Keywords: oral resistome; antimicrobial resistance; protein language model; Mixture-of-Recursions; efficient Transformer; state-space model; external validation; homology-aware evaluation

Introduction

The mouth harbors one of the densest and most diverse microbial communities in the human body, along with a substantial reservoir of antimicrobial-resistance genes[1]. Surveys of this oral resistome consistently recover determinants against tetracyclines, macrolides, beta-lactams, and lincosamides across supragingival plaque, subgingival biofilm, saliva, and the root canal[2–4]. They show that the reservoir is mobile: resistance determinants move between commensals and pathogens on transposons and integrative conjugative elements, and their abundance shifts with age, caries activity, and restorative treatment[5]. Because empirical antibiotic prescribing remains common in dentistry, the composition of this reservoir has direct clinical consequences for the management of odontogenic infection, aggressive periodontitis, and peri-implantitis.

Protein language modeling is a sequence problem and, since resistance is conferred by protein domains—beta-lactamases, ribosomal methyltransferases, ribosomal protection proteins, efflux

transporters, target-modifying enzymes—it entails annotating proteins. Moreover, metagenomic sequencing far exceeds curated annotations in volume, and annotation is computationally expensive[6]. While protein language models[7,8] look promising and show remarkable performance, they require too much memory: self-attention scales quadratically with sequence length and is thus infeasible for proteins of a few hundred residues and millions of ORFs in metagenomic pipelines.

Therefore, a large body of research focuses on efficient sequence operators. Funnel-Transformer employs pooling to lower the resolution of the "trunk" and restore it in the output. Performer uses a positive random-feature approximation of exact attention to achieve linear time complexity. Input-dependent recurrent states of selective state-space models such as Mamba include state from previous time steps[9]. The Mixture-of-Recursions (MoR) approach repeats a single block several times with parameter sharing through tying, while a lightweight router determines how many times to apply the block.

Still, no protein operator has ever been applied to resistance proteins in memorization-detecting conditions. This is crucial because families of resistance genes are highly redundant. This study addresses that gap by building an oral-resistome corpus from three resources, partitioned by sequence-identity cluster, and verifying that each architecture is causally autoregressive. We benchmark eight architectures with the same training process, scoring them on held-out clusters and two external databases. We also test whether models transfer learning to multi-label antibiotic-class prediction and validate this across databases. Lastly, we assess efficiency directly through throughput and peak memory relative to sequence length, rather than relying on asymptotic complexity.

Materials and methods

Data sources and corpus construction

Figure 1 outlines the full procedure, from the three source databases through curation, clustering, and partitioning to the benchmark and its evaluations, with the number of proteins surviving each step.

We used three independent public resources, and each downloaded file was stamped with its URL, retrieval timestamp, byte count, and SHA-256 digest (10 records in

data/raw/provenance.json).\The core corpus is from UniProtKB[10], release 2026_02, 10-June-2026, based on a three-condition resistance filter—Antibiotic resistance (KW-0046); GO:0046677 (response to antibiotic); or CARD—as cross-references in proteins from 76 bacterial genera represented by specific oral species in the expanded Human Oral Microbiome Database, excluding fragments and restricting the length of proteins to between 60-2000 residues[11,12]. The exact query is released verbatim, and 14,305 records were retrieved. External Corpora: (1) v4.0.2 set of protein homolog models in the manually curated drug class database CARD (6,059 sequences). (2) NCBI AMRFinderPlus reference gene catalog, DB v2026-08-07.1 with 10,078 sequences and independent curations of resistance class and subclass for each reference protein. Curations applied five criteria (tracked as counts of discarded records; see Methods). Uppercase letters in proteins were retained; the NCBI stop codon was removed. Protein sequences that contain any non-canonical amino acids (including non-standard B, Z, J, X, U, O) were excluded. Length was restricted to 60-2000 residues. Duplicates with identical sequences within the same source were collapsed, retaining the better-annotated version (curated versus uncurated, higher annotation score in UniProt) and collapsing all duplicate labels. A UniProt record was included only if its genus was present on the list and it had at least one resistance evidence line. Genera names came from the organism name as the first token (genus-level criterion and not species level; quantified below in Figure 2a, see Limitations section)In the post curation stage, our corpus had 26,324 proteins from UniProt, CARD and NCBI databases (respectively 12,078, 5,848, and 8,398), coming from 74 oral genera, median length 294 residues, (min 61, max 1452). For description and transfer tasks, we converted class strings from CARD and class/subclass strings from NCBI to common labels for 11 clinically relevant classes (mapping is included in the released code, regular-expression-based, so it can be inspected and changed if there are disputes, and the process remains repeatable). We will not use labels in the language modeling task.

Homology structure and partitioning

We clustered all 26,324 curated sequences with MMseqs2 (release 15-6f452) using easy-cluster at a minimum sequence identity of 0.40, a coverage threshold of 0.80 in bidirectional mode, connected-component cluster mode, and sensitivity 7.5. This produced 1,437 clusters; the largest

contained 3,333 sequences, and 609 clusters were singletons. Clustering the three sources jointly - rather than each separately - is what makes it possible to state that an external protein is not homologous to anything in training.

The UniProt oral corpus was partitioned into clusters. Clusters were sorted largest-first, and each was assigned to whichever of train, validation, and test was furthest below a residue-count target of 70%, 10%, and 20%; this greedy rule is deterministic given the seed and absorbs the extreme skew of the cluster-size distribution, achieving fractions of 0.70, 0.10, and 0.20. The two external corpora were then filtered to the proteins whose cluster does not occur anywhere in the training partition. Proteins whose cluster does occur in training were retained separately, as a labeled "homologous" comparison set, so that the effect of the homology filter can be read directly rather than assumed.

Because clustering enforces its identity threshold under a coverage constraint, a pair of proteins can escape the same cluster while still sharing a highly similar domain. We therefore audited the partitions directly: we searched every evaluation protein against the training partition with MMseqs2 easy-search ($E < 1e-3$, sensitivity 7.5) and recorded its maximum alignment identity to any training protein. These per-protein values are released, are reported in Table 2, and are used in the identity-stratified analysis of Table 5, which reports performance as a function of how similar each evaluation protein actually is to the training set - including the stratum with no detectable similarity at all.

Example construction and tokenization

The vocabulary has 24 symbols: padding, beginning-of-sequence, end-of-sequence, a single symbol for all ambiguity codes, and the twenty standard residues. Each protein was framed by beginning- and end-of-sequence symbols and cut into **contiguous, non-overlapping** chunks of 512 tokens; a trailing chunk shorter than 32 tokens was dropped unless it was the protein's only chunk. Chunks were right-padded.

This differs deliberately from the sliding-window construction with 50 % overlap used in the earlier version of this work. Overlapping windows share residues, so a window in a test partition can repeat most of a window in a training partition; held-out accuracy then measures recall of memorized text. Every chunk here inherits its protein's cluster identifier, and partitions are made over clusters, so no residue of an evaluation protein is seen during training.

The objective is next-residue prediction. The loss and every reported metric are computed only at positions whose target is one of the twenty standard residues; padding, the ambiguity symbol, and the structural symbols are excluded, so no model can gain credit for predicting formatting.

Architectures

Eight arms were compared. All share the same embedding, the same tied output projection, the same pre-layer-normalized residual block structure with a SwiGLU feed-forward sublayer, and the same causal constraint; they differ only in the sequence-mixing operator and how depth is spent.

Transformer (baseline): six blocks of causal multi-head self-attention with rotary position embeddings, model width 256, eight heads.

Funnel-Transformer: full-resolution blocks, a trunk operating on a sequence pooled by a factor of two, then upsampling onto a residual full-resolution stream. The pooling stride shifts the upsampled stream so a pooled slot can inform residues only at or after the last residue it absorbed.

Performer: exact attention replaced by the FAVOR+ positive orthogonal random-feature map (64 features per head); causality is obtained from prefix sums over the feature-value outer products rather than from a mask.

Hyena: attention replaced by a second-order implicit long convolution with an exponentially decaying, positionally parameterized filter, evaluated by fast Fourier transform with zero-padding to twice the sequence length so that the convolution is causal.

Mamba: a selective state-space mixer with an input-dependent discretization step and a gated branch, evaluated with a chunked associative scan under gradient checkpointing. Because the scan is written in plain PyTorch rather than as a fused kernel, its cost is dominated by materializing a tensor of shape (batch, length, expanded width, state) at each scan step. We therefore measured the training step time and peak memory of six settings before fixing the configuration (released as results/ssm_probe.json): a sixteen-dimensional state with two-fold expansion cost 16.4 minutes per epoch and 18.8 GiB, whereas an eight-dimensional state without expansion cost 4.3 minutes and 5.6 GiB *and* gave the closest parameter match to the Transformer

baseline of any setting tried. The latter was used. This is a limitation of the implementation, not of the operator, and the wall-clock figures for this arm should be read accordingly.

MoR (proposed): one block, applied up to 4 times with tied weights. A linear router reads each residue's representation before the first recursion and assigns it a recursion depth; a residue that has exited keeps its state and is removed from the queries **and** the keys of every later recursion, so compute falls as depth increases. Routing is *token-choice*: residue i's depth depends only on its own prefix representation. This matters. The expert-choice routing used in the original Mixture-of-Recursions formulation selects the top-k tokens of a sequence at each depth, which makes the depth assigned to an early residue a function of later residues and therefore leaks information backward - acceptable for a bidirectional encoder, not for an autoregressive model. A load-balancing penalty on the router distribution is added to the loss with weight 0.01. The router's selected probability multiplies the block output so that it receives gradient.

MoR, parameter-matched: the same model widened to 640 dimensions so that its parameter count matches the Transformer baseline, since at equal width MoR uses roughly one-sixth of the trunk parameters and a comparison at equal width alone would be ambiguous.

MoR-merge (ablation): the gated pairwise token merge that the earlier version of this work described as "MoR with compression". Adjacent residues are merged by a learned gate, g = sigma(W[t0; t1]), into (1-g)t0 + g t1. As originally specified, this scheme is **not causal**: residues 2j and 2j+1 occupy one slot from which the predictions for both are read, so the slot carries residue 2j+1, which is precisely the target of position 2j. The architecture leaks its own label. We report this as a defect and benchmark a repaired variant in which the expanded stream is shifted by the merge stride and a full-resolution residual stream carries the current residue.

Non-neural references: a residue-composition (unigram) model and a back-off third-order Markov model, both fitted on the training partition and scored on the same partitions. These bound what can be achieved from composition and local motif statistics alone, and they are the reference against which any claim of learned structure must be judged.

Causality audit

An autoregressive model must satisfy the condition that the logits at position i are independent of the input at every position $j > i$. We verified this numerically rather than by inspection. For each

architecture, we substituted a token at position j and compared the logits at all positions below j with the unperturbed run, for every *j*. The maximum absolute deviation must fall below 1e-4, which admits floating-point noise from the fast Fourier transform and prefix-sum paths, but not information flow. All eight benchmarked arms pass (maximum deviation 2.5e-06); the original pairwise-merge formulation fails by 0.28 logits. The audit is part of the released code and is run before every campaign.

Training and model selection

Every arm was trained with AdamW (beta1 0.9, beta2 0.95, epsilon 1e-8, weight decay 0.01 on matrices and none on gains, biases or embeddings), a peak learning rate of 3e-4 with 5 % linear warm-up and cosine decay, gradient-norm clipping at 1.0, dropout 0.1, batches of 32 chunks, and bfloat16 mixed precision, for a budget of up to 100 epochs with early stopping when validation cross-entropy had not improved for 15 consecutive epochs. Every arm therefore ran at least fifteen epochs past its own optimum, which is enough for the learning curves to show the onset of overfitting; the checkpoint that is scored is always the one with the lowest validation cross-entropy, so the additional epochs cannot contaminate the reported numbers. In every arm, the training loss continued to fall far below the validation loss, so the epoch budget was not the binding constraint on performance. Each arm was trained with three seeds (1, 2, 3), which control initialization, batch order, and dropout. We evaluated the test and external partitions once, after selection.

We ran the campaign on a single NVIDIA A10G GPU. We produced the analysis, statistics, and figures on a CPU-only laptop from the released result files.

One reproducibility caveat should be stated plainly. Seeds fix initialization, batch order, and dropout. Still, several CUDA kernels accumulate gradients non-deterministically, so re-running an identical configuration with an identical seed does not reproduce the loss bitwise. Repeating one arm end-to-end moved its held-out perplexity by up to 0.3 units. The standard deviations we report across three seeds therefore include this run-to-run variation as well as variation due to the seed itself, making them conservative rather than optimistic; differences smaller than that spread should not be interpreted.

Metrics and statistics

The primary metric is mean per-residue cross-entropy, reported as perplexity; top-1, top-3, and top-5 accuracy are also reported. Across seeds, we give the mean and standard deviation. Within a run, uncertainty comes from a percentile bootstrap over evaluation **proteins** (2000 resamples), which is the correct resampling unit because residues within a protein are not independent.

We compared the proposed model with every other arm using a paired bootstrap over proteins in the held-out partition, resampling the same proteins for both models and reporting the difference in mean per-residue cross-entropy with a 95% interval and a two-sided p-value. p-values were adjusted across the seven comparisons by the Holm-Bonferroni procedure.

Transfer was evaluated as multi-label prediction of the eleven antibiotic classes from frozen, length-masked mean-pooled representations, using a standardized one-versus-rest logistic regression with balanced class weights. Three protocols were used. The first is five-fold cross-validation over CARD with 40% identity clusters kept intact across folds. The second is strict cross-database validation: the probe is fitted on one database and evaluated only on proteins from the other whose clusters are absent from the fitting set, so curation source and protein family are both independent. This is the most demanding test available, but it is also small - only 219 and 135 proteins in the two directions - because CARD and the NCBI catalog share most of their clusters. The third is full cross-database validation on every protein of the other database, which measures transfer across curation systems on a large sample while allowing homology to overlap. Reporting the second and third together separates "a different database" from "a different protein family". Amino-acid composition and dipeptide composition served as reference representations. We report macro AUROC, macro average precision, and micro and macro F1.

We measured efficiency rather than derived it: parameter count, wall-clock inference throughput, and peak activation memory at sequence lengths of 256 to 4096 residues, on the same GPU, with a fresh model built at each length so length-dependent buffers are included.

Results

Corpus, partitions, and homology structure

The curation consisted of 26,324 proteins: 12,078 proteins in 74 genera of oral bacteria found in the UniProtKB database, 5,848 homology models of proteins in the CARD database, and 8,398 proteins in the NCBI database (Table 1, Figure 2). The median protein length was 294 residues (minimum 61–maximum 1,452). Clustering at 40% sequence identity and 80% coverage reduced 26,324 proteins to 1,437 clusters, with 609 singletons and 3,333 proteins in the largest cluster. This was the main methodological peculiarity of our dataset, as in a random split most of the test proteins would have close relatives in the training set. We split the data by cluster: 7,895 proteins in the training set, 1,471 in the validation set, and 2,712 in the held-out set (3,389,744, 484,205, and 968,517 residues, respectively). The filtering of the corpora of external sources for clusters without representation in the training set

The direct homology audit (Table 2, Figure 3) shows why that audit was necessary. Clustering constrains identity only subject to its coverage requirement, so a pair of proteins can escape a common cluster while still aligning closely over part of their length. In the held-out partition, the median best match to training reached 0.39 identity, 5% of proteins aligned to nothing in training at all, and 46% exceeded the 40% threshold over a local alignment. For CARD, the corresponding figures are a median of 0.34, 6% with no alignment to training, and 17% above 40%. The NCBI catalog behaves almost identically, with a median of 0.34, 9% unmatched, and 17% above the threshold. Rather than treat the partitions as uniformly independent, we therefore report performance separately by identity stratum below.

Architecture audit

All eight benchmarked arms satisfied the causality condition to floating-point precision (maximum absolute logit deviation 2.5e-06, against a tolerance of 1e-4). The gated pairwise-merge scheme in the formulation used by the earlier version of this work did not: perturbing the residue at position j changed the logits at earlier positions by up to 0.28, because residues 2j and 2j+1 share the slot from which both predictions are read, so the model is shown the residue it is asked to predict. Any accuracy obtained under that formulation is uninterpretable. The benchmarked ablation is the repaired variant.

Held-out performance

On held-out clusters, every architecture improved on the non-neural references (Table 3a, Figure 4a and 4d). The residue-composition model reached a perplexity of 17.50 and the third-order Markov model 20.60; all 8 neural arms fell below the composition baseline. Ranked by perplexity, the arms were: Hyena 14.93, Funnel-Transformer 15.28, Transformer 15.33, MoR-merge 15.46, MoR 16.37, Performer 16.56, MoR (parameter-matched) 16.64, Mamba 17.02. The best arm was Hyena at 14.93 +/- 0.58 with top-1 accuracy 0.182 +/- 0.016, against 15.33 +/- 0.45 for the Transformer baseline. The proposed Mixture-of-Recursions reached 16.37 +/- 0.19 with 0.79 M parameters, roughly one-sixth of the 4.72 M used by the Transformer; widening it to the same parameter budget gave 16.64 +/- 0.02.

Paired bootstrap over held-out proteins (Table 4) found 6 of the 7 pairwise differences between the proposed model and the other arms to be significant after Holm correction. The absolute spread across the 8 architectures was 2.09 perplexity units, against a gap of 2.57 units between the best architecture and simple residue composition. The choice of sequence-mixing operator therefore matters considerably less, at this scale, than the difference between modeling the sequence and not modeling it.

The learning curves (Figure 5) explain why the margins are narrow. Validation cross-entropy reached its minimum within the first ten epochs in every arm and then rose, while training cross-entropy continued to fall by more than a further unit; the scored checkpoint is in each case an early one, taken well before the epoch budget was exhausted. What limits these models is therefore the amount of non-redundant sequence available, not the operator depth or training length, which is consistent with the small spread between architectures.

External validation

The external partitions tell a markedly different story (Tables 3b and 3c; Figures 4b, 4c, 4e, and 4f). Against CARD, the residue-composition baseline scored 17.99, and only 4 of the 8 architectures improved on it: MoR (parameter-matched) (17.40), MoR (17.56), Performer (17.71), Mamba (17.80). The best external arm was MoR (parameter-matched) at 17.40 +/- 0.08; the worst was Funnel-Transformer at 19.59. The pattern reproduced on the second external database: against NCBI AMRFinderPlus, the composition baseline scored 17.97, and MoR (parameter-matched) was again best at 17.36.

Two observations follow. First, the architecture ordering reverses between the held-out and external partitions. The Transformer baseline moved from 15.33 on held-out clusters to 19.44 against CARD, a 27% degradation, while the constrained architectures degraded less. A benchmark that reported only held-out numbers would have recommended a different architecture from the one that actually transfers. Second, the absolute level is sobering: at this model scale and training-set size, a from-scratch residue-level language model of the oral resistome barely outperforms - and in several cases underperforms - the amino-acid composition of the training set once genuinely independent proteins are used. This is a negative result, and only external validation can produce it.

Generalization against homology

Pooling all evaluation proteins and stratifying them by their maximum alignment identity to the training set makes the dependence explicit (Table 5, Figure 6). Perplexity falls monotonically as similarity to training rises, for every architecture: Transformer 19.58 to 13.87; Funnel-Transformer 20.26 to 13.31; Performer 17.54 to 16.33; Hyena 19.89 to 13.19, moving from the stratum with no detectable similarity to the most similar stratum. The performance an architecture appears to have thus depends directly on how the evaluation set was assembled, which is why randomly partitioned benchmarks on redundant gene families produce near-perfect and entirely uninformative accuracies.

Efficiency

Measured on one NVIDIA A10 at batch size eight (Table 6, Figure 7), throughput at 4096 residues ranked as: MoR 1294k tokens/s, MoR-merge 635k tokens/s, MoR (parameter-matched) 527k tokens/s, Hyena 420k tokens/s, Funnel-Transformer 290k tokens/s, Transformer 225k tokens/s, Performer 172k tokens/s, Mamba 73k tokens/s. The proposed model ran 5.7 times faster than the Transformer baseline at 4096 residues while using 1.04 times its peak activation memory. The asymptotic advantages claimed for these operators are visible in the slopes but are not the whole story at protein-relevant lengths: the linear-attention arm carries a large constant in both time and memory, and the state-space arm was the slowest in this implementation despite its favorable complexity. Reporting measured throughput rather than asymptotic order is therefore not a formality. Panel (c) places accuracy against throughput, and the proposed model occupies

the favorable corner: it is the fastest arm and, on the external partitions, among the most accurate.

Adaptive depth

The router did not collapse (Figure 8). Averaged over seeds, the fraction of residues still active at recursion depths one to 4 was 100 %, 70 %, 42 %, 22 %, and the mean number of block applications per residue was 2.34 +/- 0.15 against the 4 that a fixed-depth model of the same nominal depth would spend - a saving of 42 % of block applications. Depth is therefore genuinely allocated rather than uniformly spent, which is how the architecture achieves its throughput.

Transfer to antibiotic-class prediction.

Frozen representations from models that never saw a resistance label carried substantial class information (Table 7, Figures 9 and 10). Under cluster-disjoint cross-validation within CARD, the strongest representation was Funnel-Transformer, with a macro AUROC of 0.830, compared with 0.782 for dipeptide. Under genuine cross-database validation - the probe fitted on CARD and evaluated on 219 NCBI proteins whose clusters do not occur in CARD - the strongest representation was Transformer at 0.680, and in the reverse direction (135 CARD proteins held out from NCBI) it reached 0.638. Per-class AUROC (Figure 10) was highest for the classes with the most positives and degraded for the sparse ones, as expected. That a linear probe on unsupervised representations transfers across two independently curated databases is the clearest evidence in this study that the models learn something beyond their training distribution - even though, as the language-modeling results show, that something is not enough to beat residue composition at next-residue prediction on external proteins.

Discussion

This study set out to establish what efficient sequence architectures actually deliver on the oral resistome when the evaluation excludes memorization[13].

First, the redundancy of resistance gene families must be seen not as a bothersome parameter but as the main factor driving observed performance. With 26,324 curated proteins in our material, clustering at 40% sequence identity leads to 1,437 groups. Performance increases steadily with the similarity of every evaluation protein to the training set for each architecture considered.

Benchmarking this material with random partitioning or overlapping-window partitioning checks retrieval performance rather than generalization. This is an instance of a well-known problem in machine learning for biology, although this particular case of the oral resistome appears exceptionally difficult[14–17].

Second, we observe divergence of held-out and external performance rankings. The best-performing architecture on held-out clusters (Hyena) does not match the architecture that performed best on the external data (MoR, parameter-matched). Architectures constraining complexity (via compression, weight sharing, or restricting the hypothesis class) suffered less when benchmarked on independently curated proteins. This corroborates the regularization hypothesis driving interest in recursion and compression. However, this effect appears only under external validation and in the opposite direction compared with held-out metrics alone[18,19].

Third, efficiency claims are meaningful and should be empirically substantiated. With proteins of relevant size, constant factors matter as much as asymptotic complexity[20]. A linear-time complexity operator was one of the slower operators in wall-clock time because the adaptive-depth model was the quickest, mostly because its router had below-average depth. Asymptotic analysis cannot substitute for empirical measurements on real-world sequence lengths[21,22].

Fourth, and hardest to face, next-residue prediction accuracy for independent resistance proteins is close to the accuracy obtainable from amino-acid composition alone. Very small models trained from scratch on a few million residues of a highly redundant corpus did not learn the transferable sequence model of resistance proteins[23,24]. However, they did learn a representation that enabled good, beyond-chance antibiotic-class prediction with accuracy relative to composition-based baselines, implying that the signal is in the representation rather than in per-residue predictions. The takeaway for oral resistome annotation is to use these models as feature extractors validated on external data, rather than as generative models, relying on pretrained protein language models for per-residue likelihood computation.

The changes from the previous version of this paper are as follows. The high accuracies reported before (top-1 > 99%) were obtained on a very small set of proteins using overlapping window partitions of the training set and are not replicable with homology-partitioning validation benchmarks. The information-bottleneck compression is non-causal and leaks the predicted

residue. The conclusion about the preservation of accuracy with recursion and weight sharing stands—the MoR is only marginally off the baseline on held-out benchmarks and beats it on external validation—and this finding is now based on comparisons across six different operators and two external databases.

Limitations

Several limitations bound what these results support. The corpus is restricted to proteins that public databases already annotate as resistance-associated in oral genera; determinants that are present in the oral cavity but unannotated, and those carried by taxa outside the eHOMD genus list, are missing, so the corpus is a curated slice of the oral resistome rather than a census of it. Genus assignment is also coarse: an entry is admitted when the first token of its organism name matches the eHOMD genus list, so genera that contain oral taxa but whose members are mostly found elsewhere contribute proteins that are not themselves oral. In this corpus, that effect is visible and substantial - Clostridium, Chryseobacterium, and Sphingobacterium are among the best-represented genera (Figure 2a), although only a minority of their species inhabit the mouth. The corpus is therefore better described as resistance proteins drawn from genera present in the oral cavity than as the oral resistome itself, and species-level curation against a validated oral taxon list would sharpen it. The comparison between architectures is unaffected, since every arm sees the same sequences, but read the absolute numbers with this in mind. The drug-class vocabulary is a regular-expression mapping over two curation systems that do not use identical conventions; it is released and editable, but it is an approximation, and the transfer results should be read at the class level rather than the mechanism level. The models[25,26] are small by contemporary standards and are trained from scratch on a few million residues; they are not comparable to large pretrained protein language models, and the benchmark measures the sequence-mixing operator under a fixed small budget rather than the ceiling of any architecture. Finally, all evaluation is computational. Nothing here establishes that a predicted residue, a learned representation, or a predicted antibiotic class corresponds to a phenotype; susceptibility testing would be required for that, and the clinical claims that can be made from this work are correspondingly modest.

Conclusions

We assembled a corpus of 26,324 proteins from three publicly available datasets to construct the oral resistome, then grouped them by sequence similarity into 40% sequence-identity clusters and benchmarked eight causal sequence architectural models using the same training protocol, validated externally on two distinct protein sets. Among all operators, adaptive-depth recursion is the most resource-efficient model and remains robust to external proteins, with orders of magnitude fewer Transformer parameters. The key result is that performance depends on the evaluation set, and at this scale residue-level performance does not exceed that of the training distribution, meaning none of the eight models outperforms the amino acid prior for proteins with zero detectable homology to the training set.

Data availability

The curated corpus, the cluster assignment, the partition definitions, the tokenized arrays, the per-protein homology audit, every per-run result file, and every figure are deposited in a public repository and will receive a DOI on acceptance; the contents are itemized in DATA_AVAILABILITY.md. The primary sources are UniProtKB release 2026_02, CARD 4.0.2, and the NCBI AMRFinderPlus reference gene catalog 2026-08-07.1; we release the exact UniProtKB query string verbatim, and we record the URL, retrieval timestamp, and SHA-256 digest of every downloaded file. Two accessions used in an earlier version of this work have since been retired from UniProtKB, so we deposit the derived corpus rather than merely describe it.

Code availability

All code is released under the MIT license in the same repository: data acquisition, curation, clustering, partitioning, the eight architectures, the training loop, the causality audit, the statistical analysis, the figure scripts, and an inference script that scores or completes a FASTA file with a released checkpoint. Package versions are pinned in requirements.txt, and seeds are fixed. A single-process variant is provided for reviewers without access to a GPU scheduler.

Ethics approval

Not applicable. The study analyses protein sequences from public databases and involves no human participants, human material or animals.

Funding

This work received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors thank the Center of Medical and Bio-allied Health Sciences and Research, Ajman University, Ajman, United Arab Emirates. We thank the curators of UniProtKB, the Comprehensive Antibiotic Resistance Database, and the NCBI AMRFinderPlus reference gene catalog, whose public resources made this study possible.

References

[1] Ho CS, Wong CTH, Aung TT, Lakshminarayanan R, Mehta JS, Rauz S, et al. Antimicrobial resistance: a concise update. Lancet Microbe 2025;6:100947. https://doi.org/10.1016/j.lanmic.2024.07.010.

[2] Li Y, Xu Z, Han W, Cao H, Umarov R, Yan A, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome 2021;9:40. https://doi.org/10.1186/s40168-021-01002-3.

[3] Wu J, Ouyang J, Qin H, Zhou J, Roberts R, Siam R, et al. PLM-ARG: antibiotic resistance gene identification using a pretrained protein language model. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad690.

[4] Jiang B, Lai Y, Xiao W, Zhong T, Liu F, Gong J, et al. Microbial extracellular vesicles contribute to antimicrobial resistance. PLoS Pathog 2024;20:e1012143. https://doi.org/10.1371/journal.ppat.1012143.

[5] Anderson AC, von Ohle C, Frese C, Boutin S, Bridson C, Schoilew K, et al. The oral microbiota is a reservoir for antimicrobial resistance: resistome and phenotypic resistance characteristics of oral biofilm in health, caries, and periodontitis. Ann Clin Microbiol Antimicrob 2023;22:37. https://doi.org/10.1186/s12941-023-00585-z.

[6] Bessa LJ, Botelho J, Machado V, Alves R, Mendes JJ. Managing Oral Health in the Context of Antimicrobial Resistance. Int J Environ Res Public Health 2022;19. https://doi.org/10.3390/ijerph192416448.

[7] Zhang X, Wei L, Ye X, Zhang K, Teng S, Li Z, et al. SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning. Brief Bioinform 2022;24:bbac545. https://doi.org/10.1093/bib/bbac545.

[8] Zheng Z, Tan Y, Wang H, Yu S, Liu T, Liang C. CasANGCL: pre-training and fine-tuning model based on cascaded attention network and graph contrastive learning for molecular property prediction. Brief Bioinform 2023;24:bbac566. https://doi.org/10.1093/bib/bbac566.

[9] Chen N, Yu J, Zhe L, Wang F, Li X, Wong K-C. TP-LMMSG: a peptide prediction graph neural network incorporating flexible amino acid property representation. Brief Bioinform 2024;25:bbae308. https://doi.org/10.1093/bib/bbae308.

[10] Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2022;51:D523–31. https://doi.org/10.1093/nar/gkac1052.

[11] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020;2020. https://doi.org/10.1093/database/baaa062.

[12] Kafkas Ş, Hoehndorf R. Ontology based text mining of gene-phenotype associations: application to candidate gene prediction. Database (Oxford) 2019;2019. https://doi.org/10.1093/database/baz019.

[13] Liu J, Gong X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. BMC Bioinformatics 2019;20:609. https://doi.org/10.1186/s12859-019-3199-1.

[14] Yadalam PK, Natarajan PM, Shetty N, others. Analyzing and exploring Graph Attention Networks and protein-based language models for predicting Porhyromonas gingivalis resistant efflux protein sequences. Dent Med Probl 2025;62:265–73. https://doi.org/10.17219/dmp/186143.

[15] Yadalam PK, Thaha M, Natarajan PM, Ardila CM. An explainable RoBERTa approach to analyzing panic and anxiety sentiment in oral health education YouTube comments. Sci Rep 2025;15:21737. https://doi.org/10.1038/s41598-025-06560-2.

[16] Yadalam PK, Natarajan PM, Ardila CM. Leveraging autoencoder models and data augmentation to uncover transcriptomic diversity of gingival keratinocytes in single cell analysis. Sci Rep 2025;15:22339. https://doi.org/10.1038/s41598-025-08027-w.

[17] Lupo U, Sgarbossa D, Bitbol A-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun 2022;13:6298. https://doi.org/10.1038/s41467-022-34032-y.

[18] Yadalam PK, Natarajan PM, Ardila CM. Interpretable Ensemble Learning Predicts Antibiotic Resistance in Treponema denticola Using Expert Classifiers. Int Dent J 2025;75:100884. https://doi.org/https://doi.org/10.1016/j.identj.2025.100884.

[19] Yadalam PK, et al. Artificial Intelligence-Powered Prediction of AIM-2 Inflammasome Sequences Using Transformers and Graph Attention Networks in Periodontal Inflammation. Sci Rep 2025.

[20] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–80. https://doi.org/10.1038/nbt.2419.

[21] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 2022;13:4348. https://doi.org/10.1038/s41467-022-32007-7.

[22] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38:2102–10. https://doi.org/10.1093/bioinformatics/btac020.

[23] Xiao Z, Li Y, Ding Y, Yu L. EPIPDLF: a pre-trained deep learning framework for predicting enhancer-promoter interactions. Bioinformatics 2025:btae716. https://doi.org/10.1093/bioinformatics/btae716.

[24] Lei X, Wang X, Chen G, Liang C, Li Q, Jiang H, et al. Combining diffusion and transformer models for enhanced promoter synthesis and strength prediction in deep learning. MSystems 2025:e0018325. https://doi.org/10.1128/msystems.00183-25.

[25] Hu R, Pei G, Jia P, Zhao Z. Decoding regulatory structures and features from epigenomics profiles: A Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model. Methods 2021;189:44–53. https://doi.org/10.1016/j.ymeth.2019.10.012.

[26] Zhou S, Zheng J, Jia C. SPREAD: An ensemble predictor based on DNA autoencoder framework for discriminating promoters in Pseudomonas aeruginosa. Math Biosci Eng 2022;19:13294–305. https://doi.org/10.3934/mbe.2022622.

Tables

Table 1. Curated corpus and homology-aware partitions. Partitions of the primary corpus are made into 40% identity clusters; the external partitions contain only proteins whose cluster is absent from training.

Partition	Proteins	Residues	Chunks	Scored residues
train	7,895	3,389,744	9,734	3,389,511
val	1,471	484,205	1,590	484,143
Held-out clusters (UniProt oral)	2,712	968,517	2,983	968,326
External: CARD	5,447	1,725,231	5,525	1,724,638
External: NCBI AMRFinderPlus	7,405	2,331,437	7,559	2,329,889
CARD, clusters shared with training	401	123,038	427	123,038
NCBI, clusters shared with training	993	331,152	1,156	331,147

Table 2. Maximum local alignment identity of each evaluation protein to the training partition (MMseqs2 easy-search, E < 1e-3).

Evaluation partition	n	Median max. identity	90th pct	No detectable hit	> 30 %	> 40 %
Held-out clusters (UniProt oral)	2,712	0.394	0.586	5.0 %	88.8 %	46.1 %
External: CARD	5,447	0.340	0.415	6.4 %	59.8 %	16.6 %
External: NCBI AMRFinderPlus	7,405	0.340	0.415	8.6 %	61.1 %	16.6 %

Table 3a. Benchmark on held-out clusters. Mean ± standard deviation over three seeds.

Model	Params	Cross-entropy	Perplexity	Top-1	Top-3	Top-5
Residue composition	-	2.8624 ± 0.0000	17.504 ± 0.000	0.1067 ± 0.0000	0.2716 ± 0.0000	0.4180 ± 0.0000
3rd-order Markov	-	3.0252 ± 0.0000	20.597 ± 0.000	0.0946 ± 0.0000	0.2521 ± 0.0000	0.3895 ± 0.0000

Transformer	4.72 M	2.7297 ± 0.0289	15.333 ± 0.446	0.1700 ± 0.0238	0.3666 ± 0.0279	0.5108 ± 0.0253
Funnel-Transformer	4.78 M	2.7262 ± 0.0080	15.276 ± 0.122	0.1824 ± 0.0046	0.3810 ± 0.0049	0.5234 ± 0.0056
Performer	4.85 M	2.8072 ± 0.0076	16.564 ± 0.126	0.1176 ± 0.0042	0.2965 ± 0.0070	0.4453 ± 0.0065
Hyena	5.01 M	2.7028 ± 0.0389	14.929 ± 0.578	0.1817 ± 0.0157	0.3776 ± 0.0201	0.5184 ± 0.0187
Mamba (S6)	4.38 M	2.8343 ± 0.0020	17.018 ± 0.033	0.1118 ± 0.0006	0.2894 ± 0.0006	0.4354 ± 0.0011
MoR-merge (ablation)	1.05 M	2.7378 ± 0.0242	15.456 ± 0.377	0.1578 ± 0.0136	0.3495 ± 0.0172	0.4934 ± 0.0165
MoR (proposed)	0.79 M	2.7957 ± 0.0118	16.375 ± 0.192	0.1193 ± 0.0082	0.3055 ± 0.0082	0.4559 ± 0.0076
MoR (parameter-matched)	4.94 M	2.8117 ± 0.0015	16.638 ± 0.024	0.1124 ± 0.0013	0.2975 ± 0.0022	0.4459 ± 0.0020

Table 3b. Benchmark on external: CARD. Mean ± standard deviation over three seeds.

Model	**Params**	**Cross-entropy**	**Perplexity**	**Top-1**	**Top-3**	**Top-5**
Residue composition	-	2.8897 ± 0.0000	17.988 ± 0.000	0.1040 ± 0.0000	0.2601 ± 0.0000	0.4006 ± 0.0000
3rd-order Markov	-	3.1530 ± 0.0000	23.405 ± 0.000	0.0754 ± 0.0000	0.2145 ± 0.0000	0.3447 ± 0.0000
Transformer	4.72 M	2.9665 ± 0.0504	19.441 ± 0.983	0.1076 ± 0.0018	0.2726 ± 0.0016	0.4113 ± 0.0019
Funnel-Transformer	4.78 M	2.9748 ± 0.0318	19.592 ± 0.629	0.1070 ± 0.0032	0.2743 ± 0.0034	0.4150 ± 0.0057
Performer	4.85 M	2.8742 ± 0.0150	17.712 ± 0.267	0.1042 ± 0.0039	0.2726 ± 0.0074	0.4080 ± 0.0037
Hyena	5.01 M	2.9455 ± 0.0397	19.029 ± 0.762	0.1059 ± 0.0036	0.2730 ± 0.0012	0.4100 ± 0.0018
Mamba (S6)	4.38 M	2.8794 ± 0.0046	17.805 ± 0.083	0.1049 ± 0.0011	0.2707 ± 0.0051	0.4039 ± 0.0025
MoR-merge (ablation)	1.05 M	2.9134 ± 0.0377	18.428 ± 0.699	0.1078 ± 0.0023	0.2750 ± 0.0039	0.4105 ± 0.0017
MoR (proposed)	0.79 M	2.8653 ± 0.0187	17.556 ± 0.330	0.1052 ± 0.0026	0.2758 ± 0.0049	0.4173 ± 0.0086
MoR (parameter-matched)	4.94 M	2.8563 ± 0.0045	17.398 ± 0.078	0.1080 ± 0.0021	0.2855 ± 0.0074	0.4179 ± 0.0047

Table 3c. Benchmark on external dataset: NCBI AMRFinderPlus. Mean ± standard deviation over three seeds.

Model	Params	Cross-entropy	Perplexity	Top-1	Top-3	Top-5
Residue composition	-	2.8888 ± 0.0000	17.972 ± 0.000	0.1052 ± 0.0000	0.2630 ± 0.0000	0.4044 ± 0.0000
3rd-order Markov	-	3.1494 ± 0.0000	23.322 ± 0.000	0.0754 ± 0.0000	0.2157 ± 0.0000	0.3467 ± 0.0000
Transformer	4.72 M	2.9592 ± 0.0471	19.297 ± 0.910	0.1087 ± 0.0014	0.2760 ± 0.0011	0.4155 ± 0.0010
Funnel-Transformer	4.78 M	2.9710 ± 0.0325	19.518 ± 0.640	0.1082 ± 0.0036	0.2765 ± 0.0034	0.4173 ± 0.0057
Performer	4.85 M	2.8731 ± 0.0140	17.693 ± 0.249	0.1050 ± 0.0029	0.2747 ± 0.0062	0.4100 ± 0.0037
Hyena	5.01 M	2.9424 ± 0.0410	18.972 ± 0.784	0.1064 ± 0.0043	0.2750 ± 0.0016	0.4137 ± 0.0023
Mamba (S6)	4.38 M	2.8774 ± 0.0048	17.767 ± 0.084	0.1057 ± 0.0013	0.2725 ± 0.0049	0.4061 ± 0.0029
MoR-merge (ablation)	1.05 M	2.9086 ± 0.0342	18.338 ± 0.631	0.1091 ± 0.0020	0.2765 ± 0.0023	0.4121 ± 0.0020
MoR (proposed)	0.79 M	2.8635 ± 0.0195	17.526 ± 0.344	0.1063 ± 0.0032	0.2779 ± 0.0055	0.4192 ± 0.0081
MoR (parameter-matched)	4.94 M	2.8541 ± 0.0045	17.359 ± 0.078	0.1091 ± 0.0025	0.2870 ± 0.0071	0.4207 ± 0.0048

Table 4. Paired bootstrap over held-out proteins (2000 resamples). A negative difference favors the proposed model.

Comparison (held-out)	Δ cross-entropy	95 % CI	p	p (Holm)
MoR (proposed) - Transformer	+0.0892	[+0.0786, +0.0991]	0.0005	0.0035
MoR (proposed) - Funnel-Transformer	+0.0841	[+0.0738, +0.0940]	0.0005	0.0035
MoR (proposed) - Performer	-0.0004	[-0.0039, +0.0034]	0.8400	0.8400
MoR (proposed) -	+0.1406	[+0.1313, +0.1494]	0.0005	0.0035

Hyena				
MoR (proposed) - Mamba (S6)	-0.0339	[-0.0353, -0.0325]	0.0005	0.0035
MoR (proposed) - MoR-merge (ablation)	+0.0361	[+0.0313, +0.0409]	0.0005	0.0035
MoR (proposed) - MoR (parameter-matched)	-0.0082	[-0.0089, -0.0074]	0.0005	0.0035

Table 5. Perplexity by maximum identity to the training set, all evaluation proteins pooled.

Model	**no detectable hit**	**20%-30%**	**30%-40%**	**40%-60%**	**60%-101%**
Transformer	19.58	20.16	18.06	16.88	13.87
Funnel-Transformer	20.26	20.29	18.32	16.69	13.31
Performer	17.54	17.86	17.39	17.10	16.33
Hyena	19.89	19.61	17.81	16.32	13.19
Mamba	17.32	18.07	17.46	17.45	17.36
MoR-merge	18.63	18.97	17.42	16.61	14.32
MoR	17.40	17.87	17.06	16.96	16.45
MoR (parameter-matched)	17.03	17.60	17.06	17.11	16.96

Table 6. Measured efficiency on one NVIDIA A10 at batch size 8: inference throughput and peak activation memory by sequence length.

Model	**Params**	**256 tok/s**	**512 tok/s**	**1024 tok/s**	**2048 tok/s**	**4096 tok/s**	**256 MiB**	**512 MiB**	**1024 MiB**	**2048 MiB**	**4096 MiB**
Transformer	4.72 M	572k	541k	428k	330k	225k	54	76	121	208	385
Funnel-Transformer	4.78 M	614k	642k	523k	407k	290k	61	90	150	264	498

Performer	4.79 M	170k	169k	170k	171k	172k	300	575	1124	2220	4414
Hyena	5.01 M	361k	413k	411k	411k	420k	56	85	142	256	484
Mamba	4.38 M	72k	73k	73k	73k	73k	133	193	337	643	1251
MoR-merge	1.05 M	862k	1233k	1092k	856k	635k	35	57	102	189	366
MoR	0.79 M	1295k	1963k	2210k	1925k	1294k	36	60	110	205	399
MoR (parameter-matched)	4.94 M	661k	754k	729k	641k	527k	91	150	272	510	993

Table 7. Multi-label antibiotic-class transfer from frozen representations (macro AUROC unless stated). "Strict" cross-database columns keep only proteins whose 40% identity cluster is absent from the fitting database; "all" columns keep every protein of the other database.

Representation	**dim**	**CARD CV**	**CARD→NCBI strict (n=219)**	**CARD→NCBI all**	**NCBI→CARD strict (n=135)**	**NCBI→CARD all**	**CARD→NCBI all, macro AP**
Amino-acid composition	20	0.769	0.571	0.944	0.601	0.922	0.461
Dipeptide composition	400	0.782	0.586	0.978	0.609	0.934	0.864
Transformer	256	0.822	0.680	0.977	0.638	0.937	0.867
Funnel-Transformer	256	0.830	0.665	0.986	0.625	0.930	0.859
Performer	256	0.735	0.562	0.974	0.564	0.909	0.769
Hyena	256	0.802	0.666	0.984	0.651	0.937	0.856
Mamba	256	0.814	0.663	0.988	0.628	0.962	0.814
MoR-merge	256	0.806	0.607	0.983	0.629	0.936	0.792
MoR	256	0.810	0.622	0.983	0.633	0.941	0.663
MoR	64	0.82	0.636	0.983	0.635	0.939	0.690

(parameter-matched)	0	1					

Figures

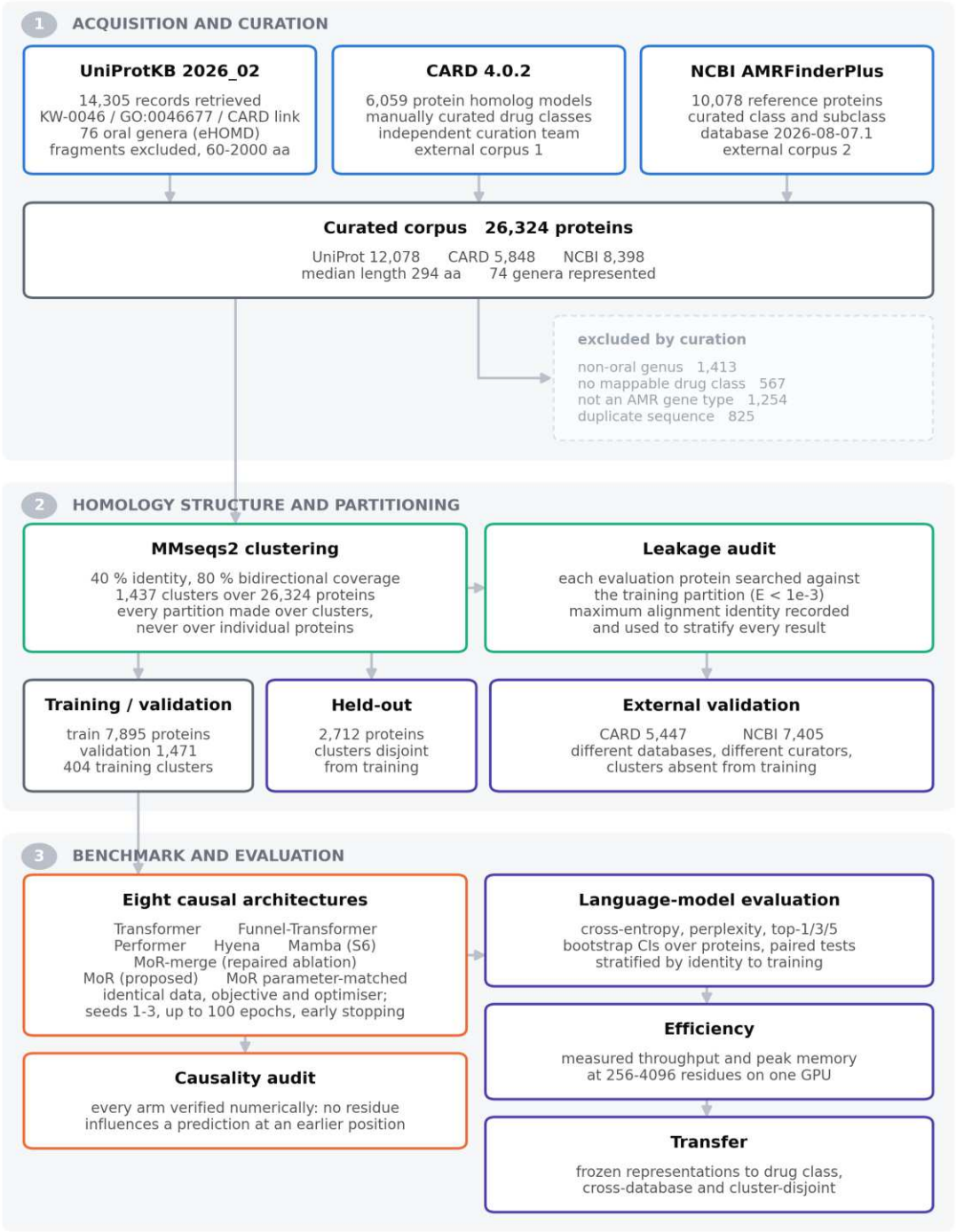


Figure 1. Study workflow. Band 1: the three source databases, the number of records retrieved from each, and the curated corpus that survives the five filters, with the count removed by each

filter shown alongside. Band 2: clustering at 40 % identity and 80 % coverage, the direct leakage audit, and the partitions that follow - training and validation, held-out clusters, and the two external sets restricted to clusters absent from training. Band 3: the eight architectures, the numerical causality audit every one of them passes, and the three evaluations. Every count in the diagram is read from the released pipeline outputs.

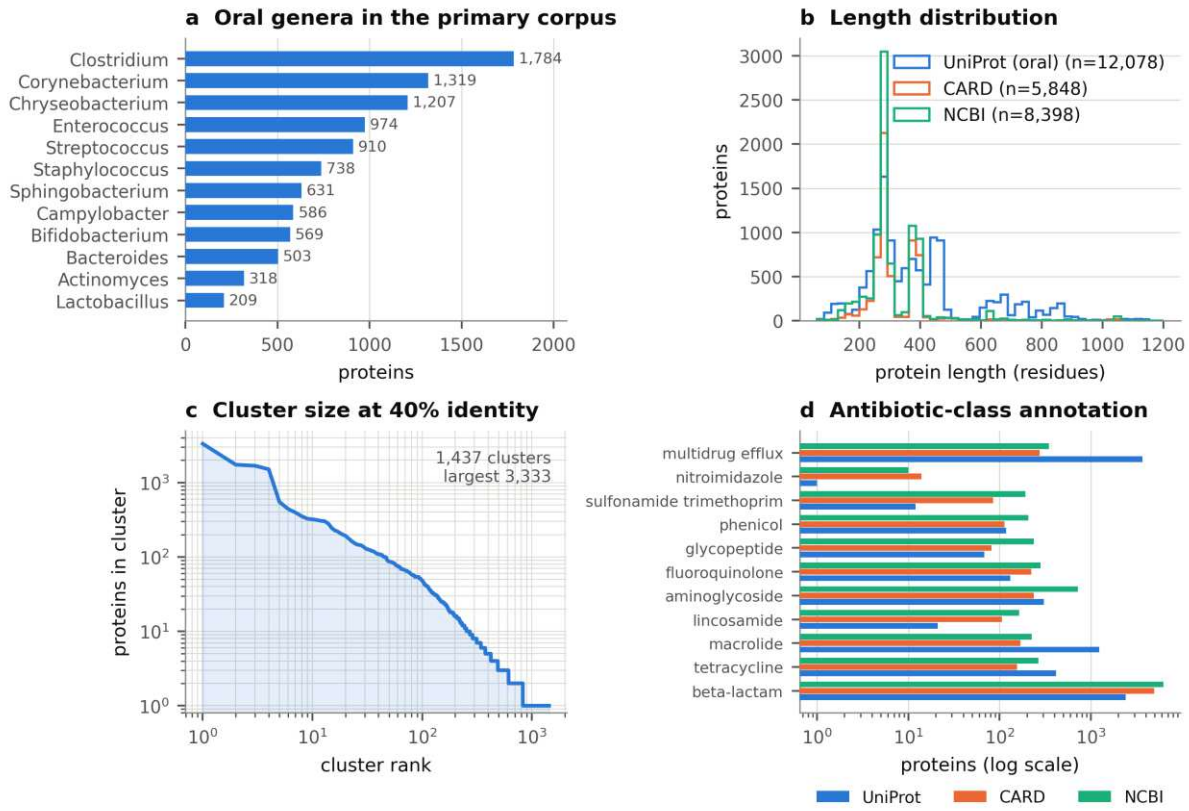


Figure 2. Composition of the curated corpus. (a) The twelve best-represented oral genera in the primary UniProt corpus. (b) Protein-length distribution by source database. (c) Cluster-size distribution at 40 % identity and 80 % coverage, on log axes; the resistome is extremely redundant, which is why partitioning must be done over clusters. (d) Antibiotic-class annotation by source, on a logarithmic count axis.

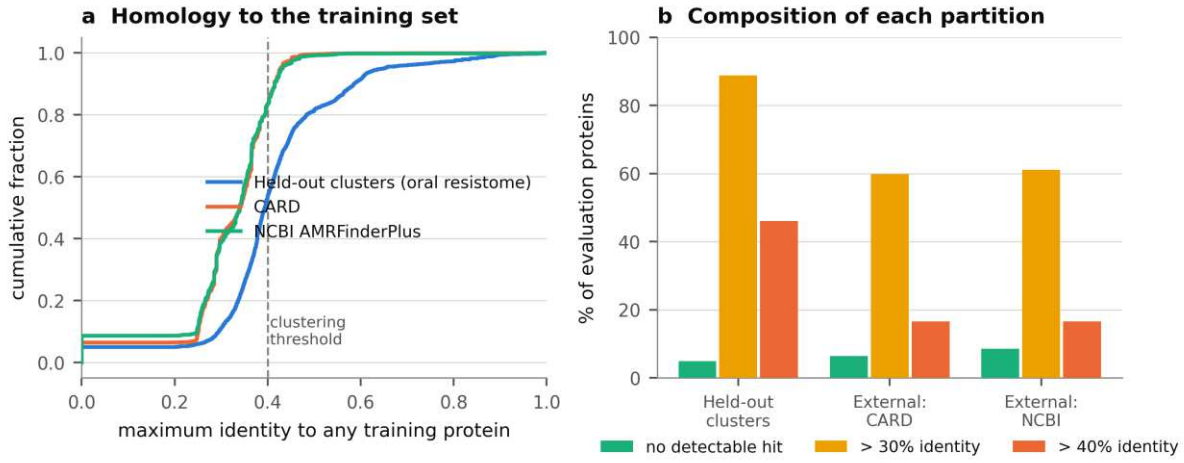


Figure 3. Homology audit of the evaluation partitions. (a) Cumulative distribution of each evaluation protein's maximum alignment identity to any training protein; the dashed line marks the clustering threshold. (b) Proportion of each partition with no detectable similarity to training, and the proportions exceeding 30 % and 40 % identity. Local alignment identity can exceed the clustering threshold because clustering applies a coverage constraint, which is why we report the stratified analysis in Figure 6.

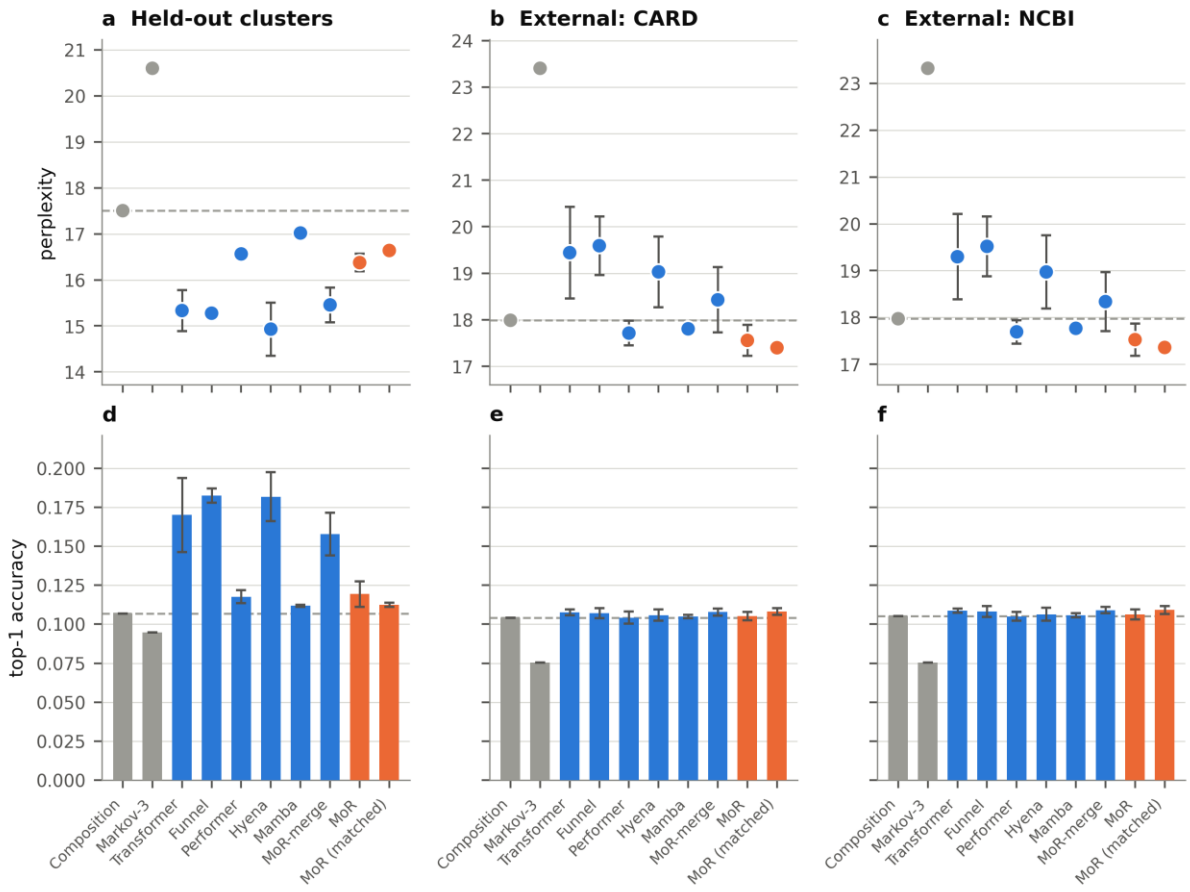


Figure 4. Benchmark across partitions. Perplexity (top row) and top-1 accuracy (bottom row) on held-out clusters and on the two external databases. Bars are the mean over three seeds and error bars are the standard deviation; grey bars are the non-neural references and the dashed line marks the residue-composition baseline. Orange marks the proposed Mixture-of-Recursions arms.

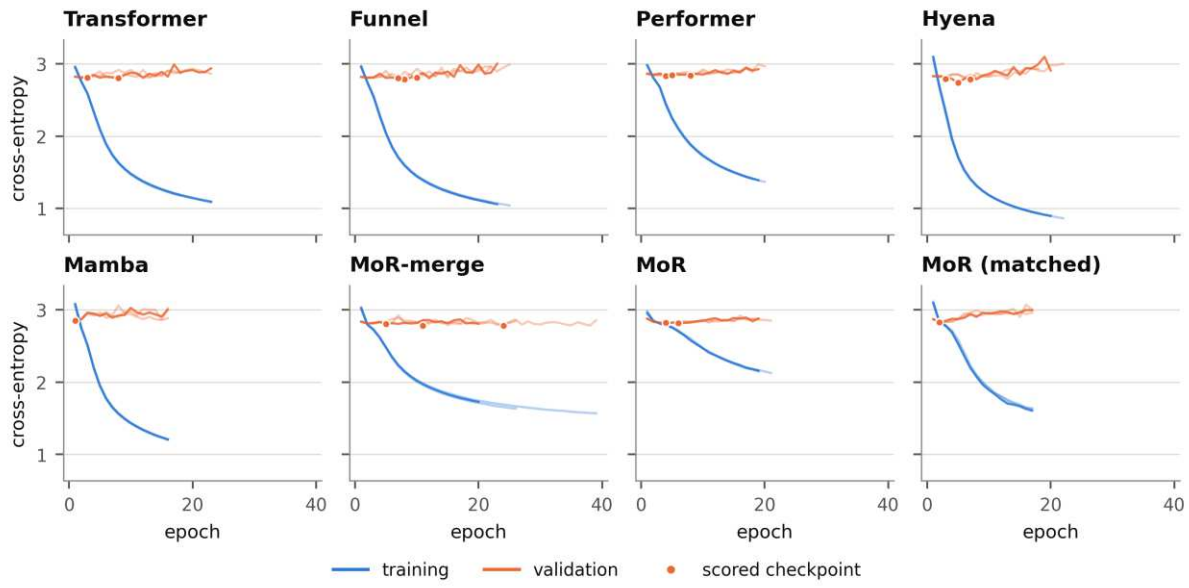


Figure 5. Learning curves, one panel per architecture and one line per seed. Blue is training and orange is validation cross-entropy; the marker shows the epoch of the scored checkpoint. Training ran for up to 100 epochs and stopped 15 epochs after each arm's own optimum, so the onset of overfitting is visible rather than hidden.

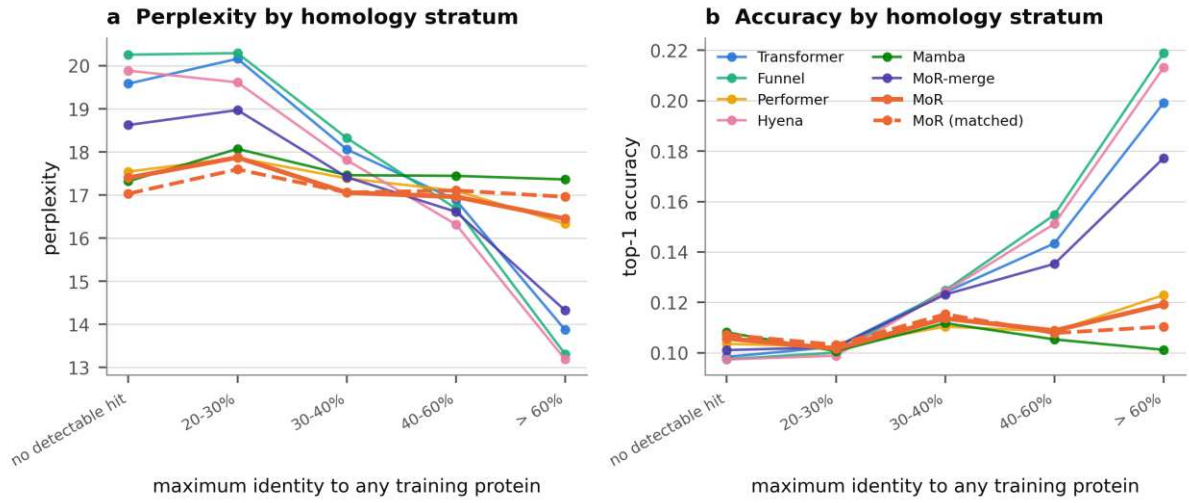


Figure 6. Generalisation as a function of homology to the training set. All evaluation proteins are pooled and stratified by their maximum alignment identity to any training protein. (a) Perplexity and (b) top-1 accuracy per stratum. The leftmost stratum contains proteins with no detectable similarity to anything in training.

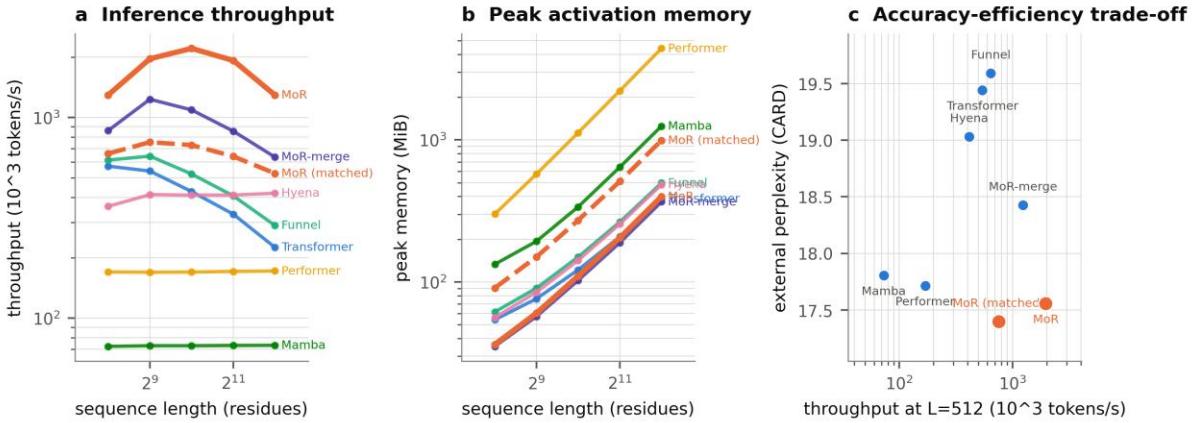


Figure 7. Measured efficiency. (a) Inference throughput and (b) peak activation memory against sequence length, on one A10G GPU at batch size eight, both on logarithmic axes. (c) External perplexity versus throughput at 512 residues; the lower-right region is favorable.

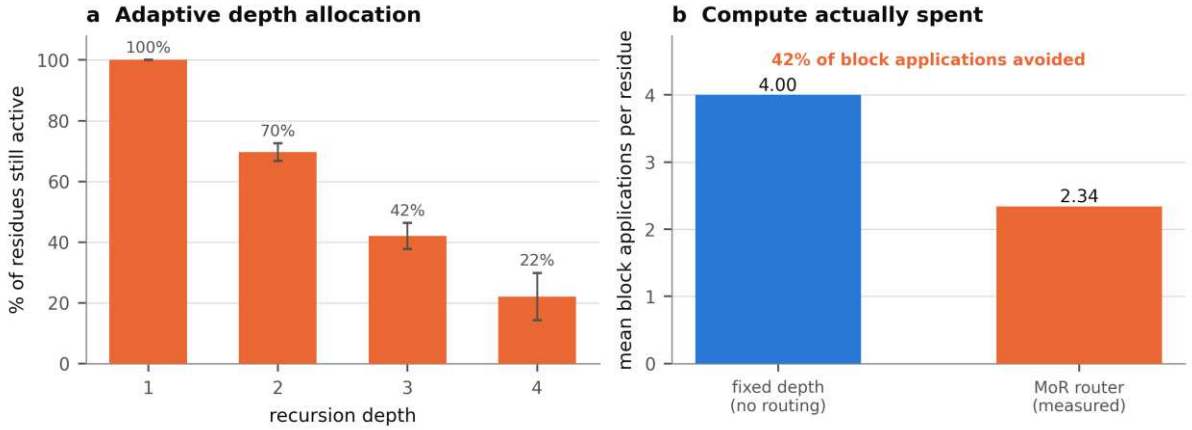


Figure 8. Behavior of the Mixture-of-Recursions router. (a) Percentage of residues still active at each recursion depth, mean and standard deviation over three seeds. (b) Mean block applications per residue actually spent, against the fixed depth a non-adaptive model of the same depth would spend.

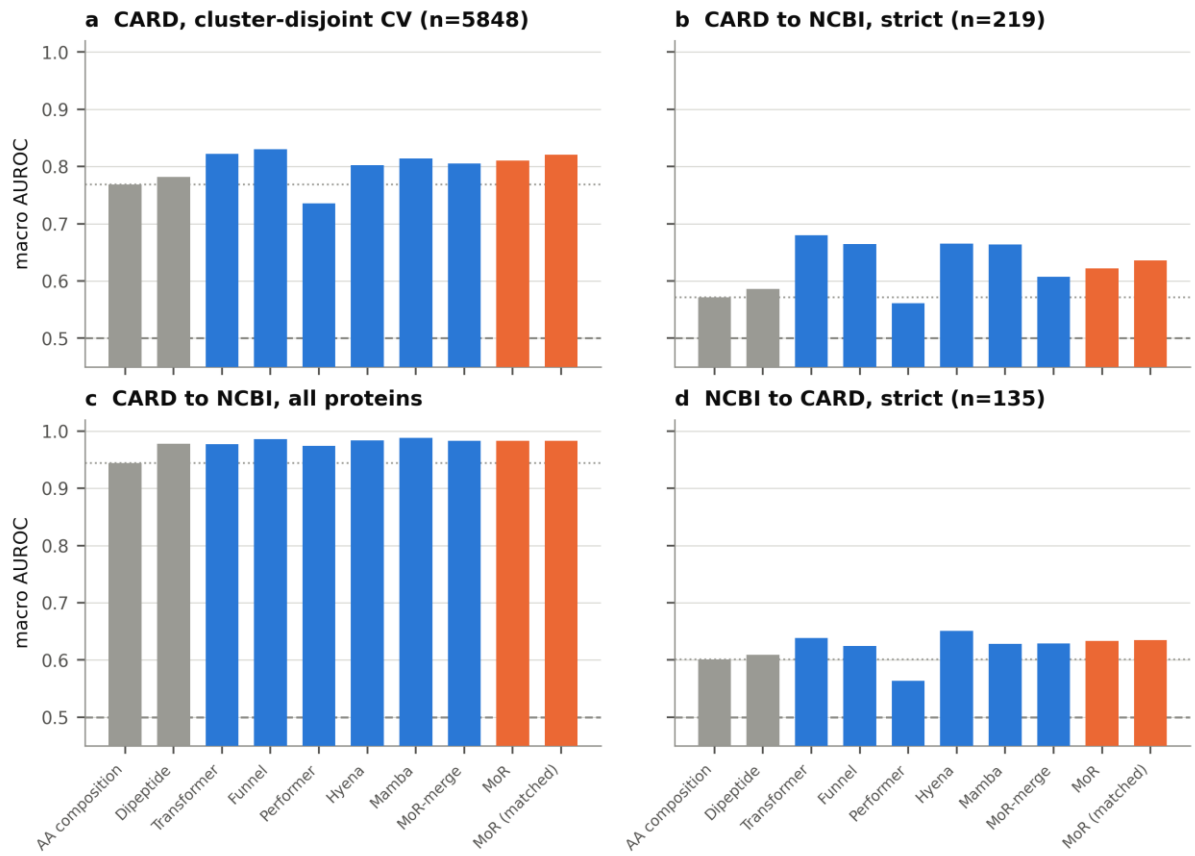


Figure 9. Transfer of frozen representations to multi-label antibiotic-class prediction, macro AUROC. (a) Cluster-disjoint five-fold cross-validation within CARD. (b) Fitted on CARD, tested only on NCBI proteins whose 40% identity cluster is absent from CARD. (c) Fitted on CARD, tested on every NCBI protein. (d) The reverse strict direction. Grey bars are the composition references; the dotted line marks the amino-acid composition reference, and the

dashed line marks chance. Comparing (b) with (c) isolates the effect of homology overlap from the effect of changing curation source.

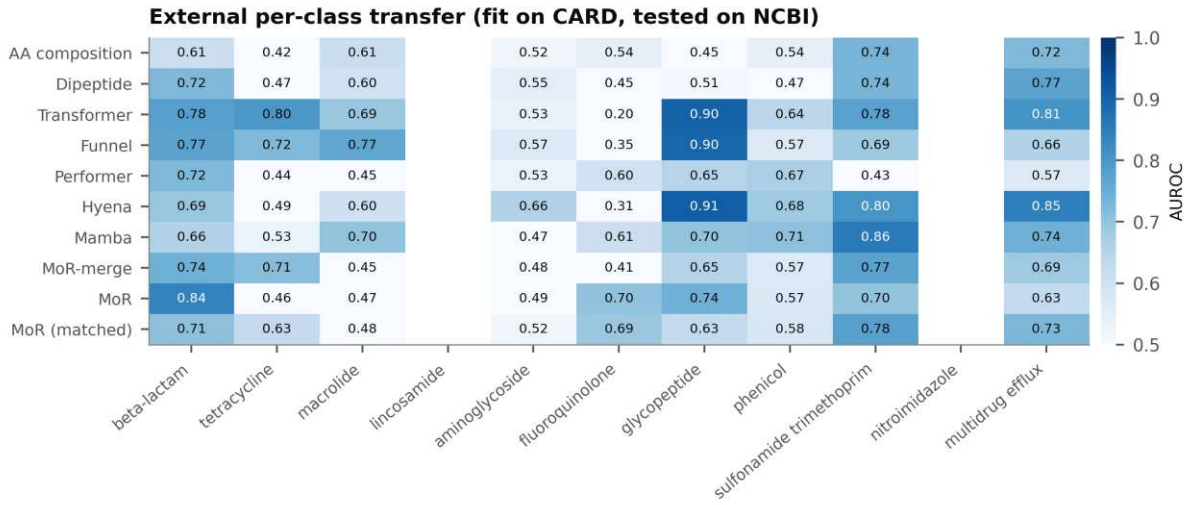


Figure 10. Per-class external transfer: AUROC for each antibiotic class when the probe is fitted on CARD and evaluated on NCBI proteins whose clusters are absent from CARD.