NaviKinase 1.0: evaluation of a domain-specific protein language model for the phosphotransferase superfamily

Jaydev Bhatt

Navitra Technologies Ltd, London, United Kingdom

Correspondence: Jaydev.bhatt@navitratech.com

ORCID: 0009-0002-7552-1516

Keywords: protein language model; protein kinases; EC 2.7; sequence generation; model evaluation; memorisation

Abstract

NaviKinase 1.0 is a 149.6 M-parameter decoder-only protein language model trained from scratch on 2.45 million sequences from the EC 2.7.- phosphotransferase superfamily, which includes the protein kinases. We evaluate it on sequences held out of training, on its own generated sequences, and against natural and shuffled controls. On held-out sequences it reaches a perplexity of about 4.1 and identifies the EC sub-subclass of a sequence with 95.6% balanced top-1 accuracy (chance 7.1%). Accuracy remains at 91% for held-out sequences below 50% identity to any training sequence (50 of 55), although perplexity rises as similarity falls. Conditioning on the Ser/Thr kinase tag raises the share of samples containing all three catalytic kinase motifs to 30%, against 2.5% for unconditioned samples (2 of 80 across two seeds). Generated sequences are mostly novel (median identity to training 34.5% to 38%), although two of 40 EC 2.7.11 samples reached 90% identity or more, and ESMFold predicts confident folds for most generated sequences (median pLDDT 78 to 88, against 82 for natural kinases). Function has not been tested, samples are small, and the novelty search is a heuristic lower bound.

1. Introduction

Protein language models learn the statistics of amino-acid sequences from large corpora and can score sequences, generate new ones and support structure prediction [1,2]. Most are large and general. NaviKinase 1.0 is a small, domain-narrow alternative: a 149.6 M-parameter decoder-only transformer [3,4] trained from scratch on a single Apple Silicon machine, on one enzyme superfamily (EC 2.7.-, the phosphotransferases) drawn from UniProt [5] and deduplicated at the UniRef90 level [6].

This report asks what that model has learned and what the available evidence supports. It tests the model on sequences held out of training, measures how similar those sequences and the model's own generations are to the training set, checks generated sequences with an independent structure predictor [1], and states the limits of each result. It does not compare NaviKinase with other models and does not test any generated protein in the laboratory.

2. Methods

2.1 Model and training

NaviKinase 1.0 is a 149.6 M-parameter decoder-only protein language model trained from scratch on Apple Silicon with MLX, covering the EC 2.7.- phosphotransferase superfamily (protein kinases plus sugar, lipid and nucleotide kinases). Architecture and training settings are given in Table 1.

Table 1. Architecture, training data and conditioning scheme of NaviKinase 1.0.

Property	Value
Architecture	18-layer transformer, pre-norm RMSNorm, RoPE, tied embeddings
Width	d_model 832, 13 heads (head_dim 64), FFN 3328
Context length	1,024 tokens
Vocabulary	64 reserved slots, 45 used: 4 special tokens, 20 standard residues, 6 ambiguity codes, and 15 EC tags (14 sub-subclass tags with IDs 30-43, plus EC_UNKNOWN at ID 44)
Training data	2,450,504 UniRef90-deduplicated sequences, 1.32 G tokens
Training run	40,289 steps (2 epochs), AdamW, peak LR 3×10^{-4}, no early stop triggered
Conditioning	Each sequence is prefixed with an EC sub-subclass tag, for example EC 2.7.11 for Ser/Thr kinases. The 14 tagged sub-subclasses are 2.7.1 to 2.7.4, 2.7.6 to 2.7.14 and 2.7.99; EC 2.7.5 is not among them. Any EC number outside these 14, and any empty or malformed field, maps to EC_UNKNOWN, which covers about 3.3% of the corpus

2.2 Data and splits

Training data were UniProtKB entries with EC 2.7.- [5], deduplicated at UniRef90 [6]. Train, validation and test splits were made at the UniRef90 cluster level, so near-identical sequences do not leak across splits. The held-out test split was never used for training or checkpoint selection.

2.3 Held-out perplexity and EC identification

The 24,997 held-out test sequences (data/processed/test.bin) were re-scored with the shipped weights. Sequences over 1,000 residues were excluded, leaving 22,507. Perplexity given the true EC tag was computed on 1,000 randomly drawn sequences, token-weighted, and compared with the same sequences with residue order shuffled.

For EC identification, 372 test sequences (up to 30 per class, seed 0) were each scored under all 14 EC sub-subclass tags (IDs 30 to 43); EC_UNKNOWN was not among the candidates, so test sequences carrying that tag are excluded from this analysis. The tag with the highest likelihood was taken as the prediction, with no tag prior, giving chance baselines of 1 in 14 and 3 in 14. Class 2.7.99 had no eligible test sequences and 2.7.14 only 12, so the balanced average covers 13 classes.

2.4 Similarity to training data

A held-out split is only as demanding as the sequences it contains, so similarity to training was measured directly. Each query sequence was compared with all 2,450,504 training sequences. A 5-mer scan over the 1.32 G training tokens collected the most similar candidates, and the best of the top eight per query was aligned with Smith-Waterman [7] using BLOSUM62 [8] in Biopython [9]. Identity is identical positions divided by query length, so a short matching fragment does not count as a near-copy. This denominator matters: computing identity over aligned columns instead changes the number of sequences at or above 90% from 2, 0 and 11 to 2, 1 and 14 for the generated EC 2.7.11, unconditioned and natural held-out groups respectively. All identities reported here use query length.

The scan is a heuristic prefilter rather than an exhaustive search, so every identity reported here is a lower bound on the true nearest-neighbour identity. This caveat applies to Sections 3.2 and 3.3 alike.

Two query sets were used. For Section 3.2, 120 sequences: 40 generated with the EC 2.7.11 tag, 40 generated without a tag, and 40 natural held-out EC 2.7.11 test sequences of 150 to 400 residues.

For Section 3.3, 800 test sequences of 100 to 600 residues with a known EC tag (seed 3), each matched to its most similar training sequence and then grouped by that identity.

2.5 Structure prediction

Generated, natural and shuffled sequences were folded with ESMFold [1], Meta's public API and a separate model, and scored by mean pLDDT [10], where above 70 is conventionally read as confident and above 90 as very high. The selection was fixed in advance: the first 10 EC 2.7.11 samples, the first 10 untagged samples, the first 6 natural held-out EC 2.7.11 sequences, and shuffled copies of those 6. Two of the 10 EC 2.7.11 folds failed after repeated gateway timeouts and are excluded, leaving 8 and 30 folds in total.

2.6 Motif scanning

Sequences were scanned with a regular-expression check for the P-loop (G.G..G), catalytic HRD and activation-loop DFG motifs, which are characteristic of protein kinases [11]. For each generation condition, 40 sequences were generated at temperature 0.8 with a maximum of 400 residues. A motif hit records local motif presence, not catalytic activity.

2.7 Known-protein controls

Six real proteins were fetched from UniProt on 27 October 2026, truncated to 1,000 residues, and scored teacher-forced, each with a shuffled copy of the same amino-acid composition as a control. A perplexity near 20 is what a model with no knowledge of amino-acid order would give over 20 residue types.

2.8 Software and reproducibility

Analyses were run locally with the released weights. Software versions, scripts, seeds and raw outputs are listed in Section 6.

3. Results

Six analyses are reported, ordered from the strongest evidence to the weakest, as seven figures and four tables. Two further figures are given as supplementary material (Appendix B). The model's own held-out test perplexity, reported in the release README, is 4.32 (loss 1.4633).

3.1 Held-out perplexity and EC identification

The model scores held-out sequences far better than shuffled controls, and identifies their EC sub-subclass almost always (Table 2). Perplexity given the true EC tag was 4.14 on 1,000 held-out sequences, close to the README's 4.32; the README figure uses packed windows and includes the tag token, so the two differ slightly in method. All 1,000 real sequences scored better than their shuffled copy.

Table 2. Held-out scoring and EC identification. Perplexity on 1,000 randomly drawn held-out sequences; identification on 372 sequences, up to 30 per class, seed 0.

Measure	Value	Reference
Perplexity, true EC tag (token-weighted)	4.14	README held-out figure 4.32
Perplexity, median	4.04	—
Perplexity, 10th–90th percentile	1.88–8.96	—

Measure	Value	Reference
Perplexity, residue order shuffled	18.19	Uniform guess over 20 residues ≈ 20
Balanced top-1 EC accuracy	95.6%	Chance 7.1%
Balanced top-3 EC accuracy	99.2%	Chance 21.4%
Per-class top-1 range (13 classes)	83.3–100%	See Supplementary Figure S1

Accuracy was uniformly high across classes and lowest for Ser/Thr kinases (2.7.11, 25 of 30 correct); the per-class breakdown is given in Supplementary Figure S1.

3.2 Similarity of held-out and generated sequences to the training data

Generated sequences are mostly far from the training set: median identity 34.5% for EC 2.7.11 and 38% without a tag (Figure 1). Two of 40 EC 2.7.11 samples reached 90% or more, one of them 97% over its full 400 residues, which is effectively a memorised training sequence. No untagged sample reached 90%. Natural held-out sequences had a median identity of 87%, and 11 of 40 were above 90%. The cluster-level split therefore removes exact clusters but not close relatives. Low identity is also not evidence of quality, since it would hold equally for sequences that do not fold.

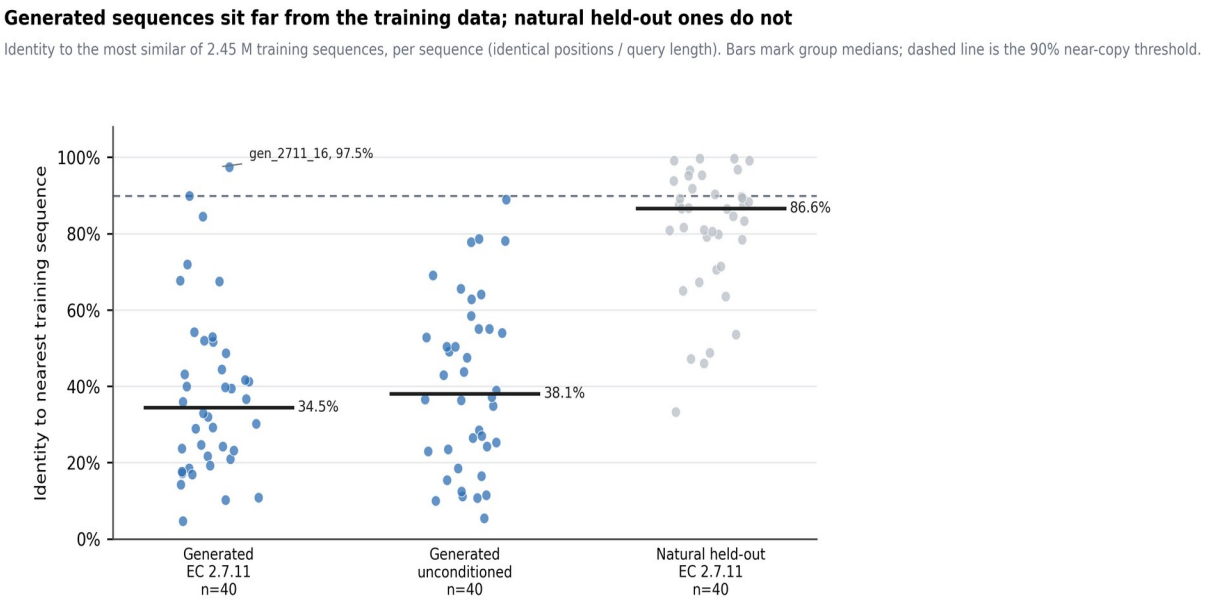


Figure 1. Generated sequences sit far from the training data; natural held-out ones do not. Identity to the most similar of 2.45 M training sequences, per sequence (identical positions / query length); near-copy threshold 90%. Medians 34.5% (generated, EC 2.7.11), 38.1% (generated, unconditioned) and 86.6% (natural held-out, EC 2.7.11), n = 40 per group. Per-sequence values are in identity.csv. Seed 1, temperature 0.8; scripts/gen_samples.py and scripts/novelty.py.

3.3 Performance versus similarity to training data

Median perplexity rose steadily as similarity fell, from 2.2 at 90% identity or more to 3.7 at 70–90%, 5.8 at 50–70%, 8.2 below 50%, and 10.6 for the 13 sequences below 30% (Figure 2). EC identification fell far less: 98% at 70% identity or more, 95% at 50 to 70%, and 91% below 50% (50 of 55; 95% confidence interval 80% to 96%). Of the 13 sequences below 30%, 12 were identified correctly (interval 67% to 99%). Always guessing the most common class would score 20.6%.

The EC signal therefore survives low similarity to training data, while the likelihoods degrade, as expected for a model that has not seen close relatives. The same lower-bound caveat applies: a sequence labelled low-identity may have a more distant relative in training that the search missed.

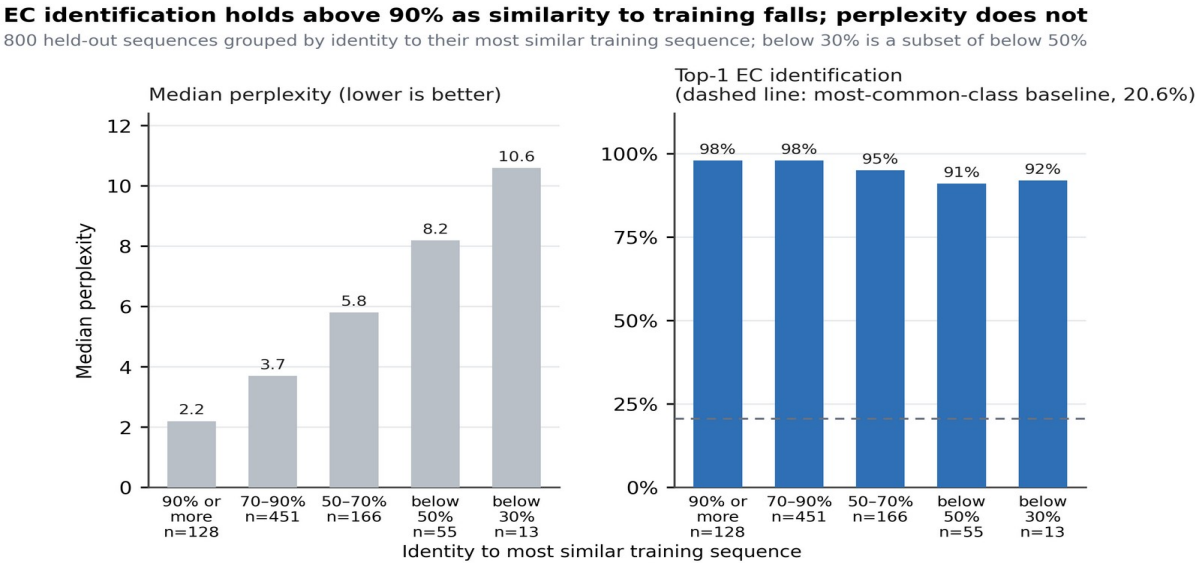


Figure 2. EC identification holds above 90% as similarity to training falls; perplexity does not. 800 held-out sequences grouped by identity to their most similar training sequence; below 30% is a subset of below 50%. Left, median perplexity (lower is better); right, top-1 EC identification, with the most-common-class baseline at 20.6%. Seed 3; scripts/similarity_eval.py and scripts/novelty.py; identity is a lower-bound estimate.

3.4 Structure confidence of generated sequences

Shuffled copies scored a median pLDDT of 28.8 (mean 29.0), so the metric does separate real sequence from noise (Figure 3A). Generated sequences reached a median of 78.1 (mean 77.1; EC 2.7.11, n = 8) and 88.2 (mean 84.9; untagged, n = 10), against 81.9 (mean 82.3) for natural held-out kinases (n = 6), so most generated sequences are predicted to fold about as confidently as natural ones. Medians are quoted throughout because the groups are small enough for one weak fold to move the mean; both statistics are given here so the two cannot be confused.

Novelty does not predict fold confidence in either direction (Figure 3B, n = 22). The most novel sequence in the folded set, gen_uncond_8 at 5.5% identity to any training sequence, scored 84.1, and gen_uncond_1 at 23.6% scored 90.6; conversely gen_2711_1, at 90.0% identity and therefore a near-copy, scored only 70.7, the lowest of any generated EC 2.7.11 fold except the two in the low 60s. No correlation is quoted because n = 22 across three heterogeneous groups would not support one. Each group has only 6 to 10 sequences, so these values are indicative only.

ESMFold confidence indicates a predictable fold, not a working kinase: only 2 of the 18 folded generated sequences contained all three catalytic motifs.

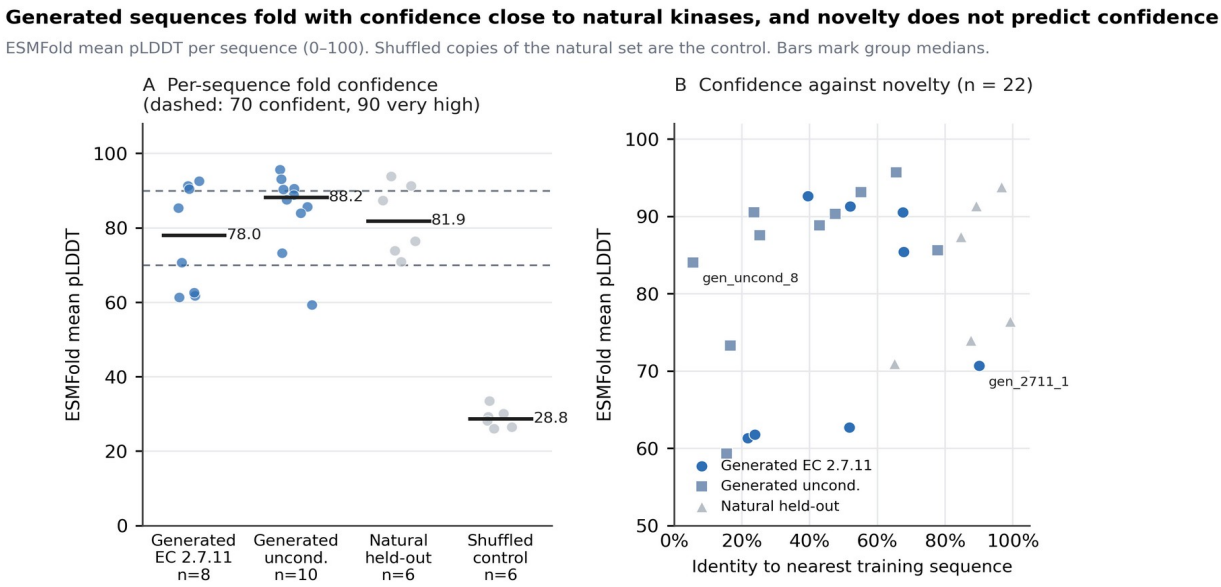


Figure 3. Generated sequences fold with confidence close to natural kinases, and novelty does not predict confidence. (A) Per-sequence fold confidence by group. (B) Mean pLDDT against identity to the nearest training sequence, for the 22 folded sequences that were also searched against the training set; the shuffled controls were not searched. Group medians (means in parentheses): generated EC 2.7.11 78.1 (77.1), n = 8; generated unconditioned 88.2 (84.9), n = 10; natural held-out EC 2.7.11 81.9 (82.3), n = 6; shuffled control 28.8 (29.0), n = 6. Per-sequence values are in fold.json. ESMFold public API, a separate Meta model; scripts/fold_eval.py; 30 folded sequences, 2 of 10 EC 2.7.11 folds failed on gateway timeouts and are excluded.

One generated structure is shown in Supplementary Figure S2. It was chosen by a rule, not by score: it is the only folded EC 2.7.11 sample that contains the P-loop, HRD and DFG motifs. It has a mean pLDDT of 85 and 68% identity to its nearest training sequence. Residues 1 to about 308 form a confident core (mostly pLDDT 85 to 99). The last 90 residues form an extended, low-confidence tail, which looks like a sequence truncated by the 400-residue generation limit rather than a finished domain.

3.5 Effect of the EC tag on generation

Conditioning on EC 2.7.11 (Ser/Thr kinases) lifted motif hit rates sharply, while small-molecule kinase conditioning (2.7.1) stayed near the untagged baseline, as expected because these motifs are specific to protein kinases (Figure 4). The share of samples containing all three motifs was 30% under EC 2.7.11 and 0% unconditioned in this seed-0 run.

With n = 40, each rate carries roughly 15 percentage points of sampling error. The 2.7.11 gap against the untagged baseline is well outside that; the 2.7.10 gap is smaller and less certain. A motif hit shows local motif presence, not an active enzyme: 70% of EC 2.7.11 samples still lacked at least one of the three motifs.

Two independent seeds support each rate. The seed-0 run was repeated with every sequence saved (scripts/seed0_motif_rerun.py, output in logs/navikinase_validation/seed0_motif_rerun.json) and reproduced the rates in Figure 4 exactly, with nothing tuned to match. A separate seed-1 pool (samples.json) gives 12 of 40 (30%) for EC 2.7.11, the same rate again, and 2 of 40 (5%) for unconditioned samples rather than 0%. Pooled across both seeds, EC 2.7.11 is 24 of 80 (30%) and unconditioned is 2 of 80 (2.5%). The conditioning effect is therefore robust across seeds, but the unconditioned baseline is not exactly zero and should not be read as such.

The same rerun recorded whether each sequence terminated on its own or hit the 400-residue cap, and conditioning changes this sharply. Under EC 2.7.11 only 3 of 40 samples emitted a stop and 37 reached the cap; the seed-1 pool agrees, with 39 of 40 at exactly 400 residues. Under EC 2.7.1 the position reverses, with 26 of 40 terminating on their own. Unconditioned samples fall between at 14 of 40, and EC 2.7.10 at 7 of 40. Sequence length under EC 2.7.11 conditioning is therefore censored by the generation limit rather than chosen by the model, and the three EC 2.7.11 samples that did terminate are too few to compare their motif content against the rest.

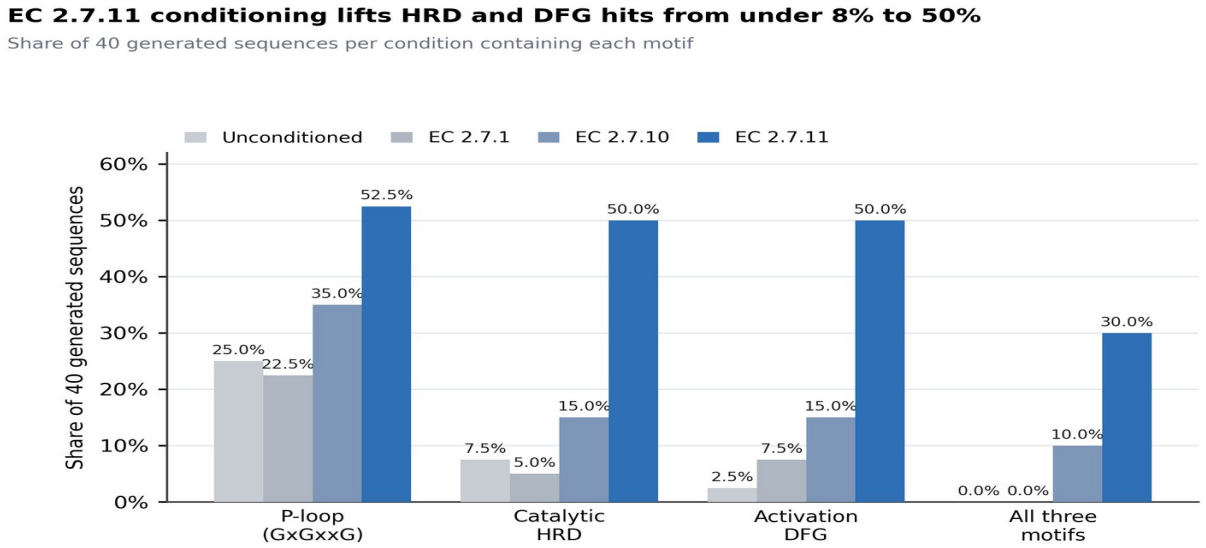


Figure 4. EC 2.7.11 conditioning lifts HRD and DFG hits from under 8% to 50%. Share of 40 generated sequences per condition containing each motif, for the P-loop (GxGxxG), catalytic HRD, activation DFG, and all three together. 40 samples per condition, temperature 0.8, seed 0; generate.py and motifs.py. The run was repeated with every sequence saved (scripts/seed0_motif_rerun.py, output logs/navikinase_validation/seed0_motif_rerun.json) and reproduced these rates exactly.

3.6 Scoring of known proteins

The three kinases scored far below the three non-kinases, which scored about as badly as shuffled sequence (Figure 5, Table 3). This is not generalisation evidence: ABL1, PKA and glucokinase are well-studied entries very likely present in the training corpus, and perplexities of 1.2 to 1.8 are far below the held-out figure of 4.32, which points to memorisation. The test shows only that the model separates in-scope from out-of-scope sequence.

Table 3. Motif content and EC ranking for six known proteins scored as a sanity check. Perplexities are shown in Figure 5. Sequences fetched from UniProt on 27 October 2026, truncated to 1,000 residues; score.py.

Protein (UniProt)	Class	Motifs found	True EC ranked first
ABL1 (P00519)	Tyr kinase, 2.7.10	3 of 3	Yes
PKA catalytic alpha (P17612)	Ser/Thr kinase, 2.7.11	2 of 3	Yes
Glucokinase (P35557)	Sugar kinase, 2.7.1	P-loop only	Yes
Lysozyme C (P00698)	Non-kinase	0 of 3	Not applicable
GFP (P42212)	Non-kinase	0 of 3	Not applicable
Hemoglobin alpha (P69905)	Non-kinase	0 of 3	Not applicable

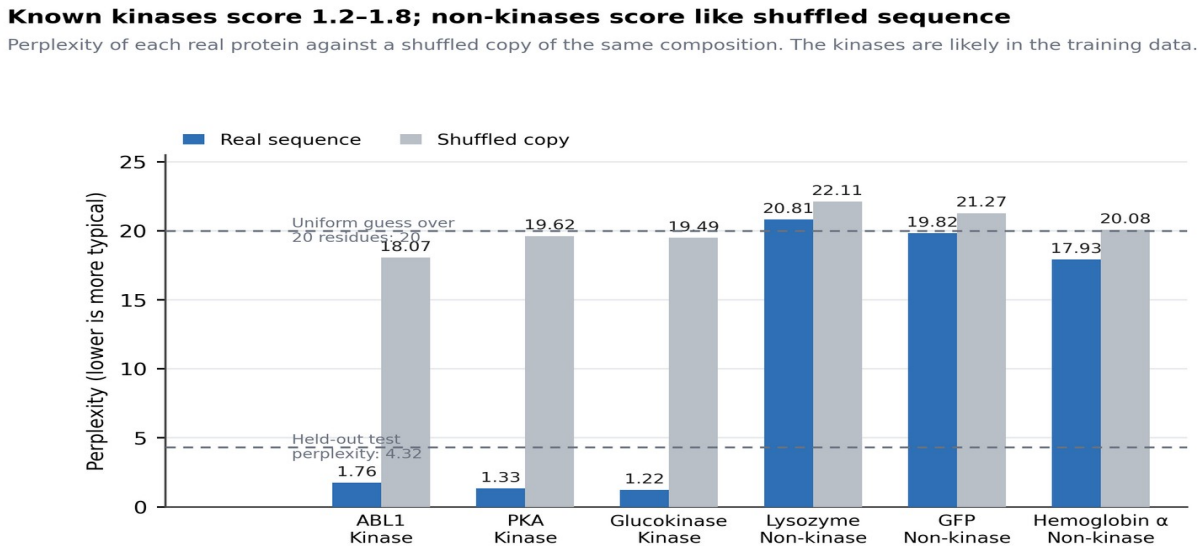


Figure 5. Known kinases score 1.2–1.8; non-kinases score like shuffled sequence (about 20). Perplexity of each real protein against a shuffled copy of the same amino-acid composition; lower is more typical. Reference lines at the held-out test perplexity of 4.32 and at the uniform guess over 20 residues. The kinases are likely present in the training data. Shuffled copies use seed 0; score.py.

3.7 Comparison with general protein language models

Two general models were compared with NaviKinase on the 800 held-out sequences used in Section 3.3. ESM-2 with 150M parameters (esm2_t30_150M_UR50D) [1] matches NaviKinase in size. ProtGPT2 [2] is an autoregressive model with 738M parameters trained on UniRef50, about five times larger.

EC identification. Both were used as frozen feature extractors. Mean-pooled final-layer embeddings of 6,052 training sequences (up to 500 per EC class, 100 to 600 residues, seed 4) were used to fit a regularised logistic-regression probe for each model, which then classified the 800 held-out sequences. NaviKinase used no labels at inference, only its tag-likelihood rule. The ESM-2 probe was also fitted on a 1,000-sequence subset to show the effect of label count.

On these 800 sequences, which are class-proportional rather than balanced (so the overall figure differs from the 95.6% balanced accuracy in Section 3.1), NaviKinase identified 97.0% and the ESM-2 and ProtGPT2 probes both 96.6% (773 of 800), a gap of 3 sequences (Figure 6). Below 50% identity to training, the ESM-2 probe scored 94.5%, NaviKinase 90.9% and the ProtGPT2 probe 89.1% (52, 50 and 49 of 55). At 90% identity or more, NaviKinase scored 98.4% against 94.5% for ESM-2 and 96.9% for ProtGPT2. These differences are within sampling error at these group sizes. With only 1,000 labels the ESM-2 probe fell to 89.4%.

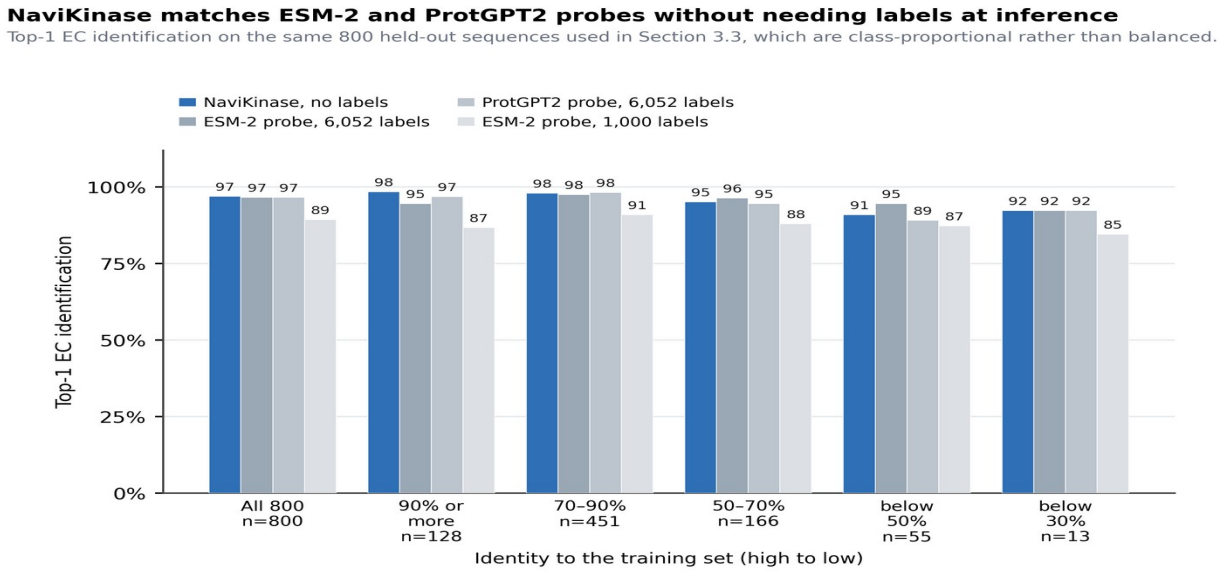


Figure 6. NaviKinase matches ESM-2 and ProtGPT2 probes without needing labels at inference. Top-1 EC identification on the same 800 held-out sequences used in Section 3.3, by identity to the training set. The probes were fitted on 6,052 labelled training sequences; the fourth series shows the ESM-2 probe fitted on 1,000. Values are percentages correct. Seed 4; scripts/baseline_esm2.py and scripts/baseline_protgpt2.py.

Likelihood. The two autoregressive models can also be compared directly on per-residue negative log-likelihood. NaviKinase was scored with its true EC tag and with no tag, and ProtGPT2 on the same sequences formatted as it was trained (60 residues per line). NaviKinase loss is summed over residue predictions only; ProtGPT2 loss is summed over all tokens, including line breaks. Both are divided by the residue count. A uniform guess over 20 residues scores 3.0 nats.

NaviKinase had the lower median loss overall (1.40 with its tag and 1.53 without, against 1.78 for ProtGPT2), and its advantage was largest when relatives exist in its training set: 0.79 against 1.51 at 90% identity or more (Figure 7). The order reverses for sequences below 50% identity to training (ProtGPT2 2.07, NaviKinase 2.17 without its tag and 2.10 with it) and below 30% (2.13 against 2.44 and 2.36, for 13 sequences). The larger, broader model therefore degrades more gracefully on unfamiliar sequences, while the specialised model is better when close relatives are in its training data.

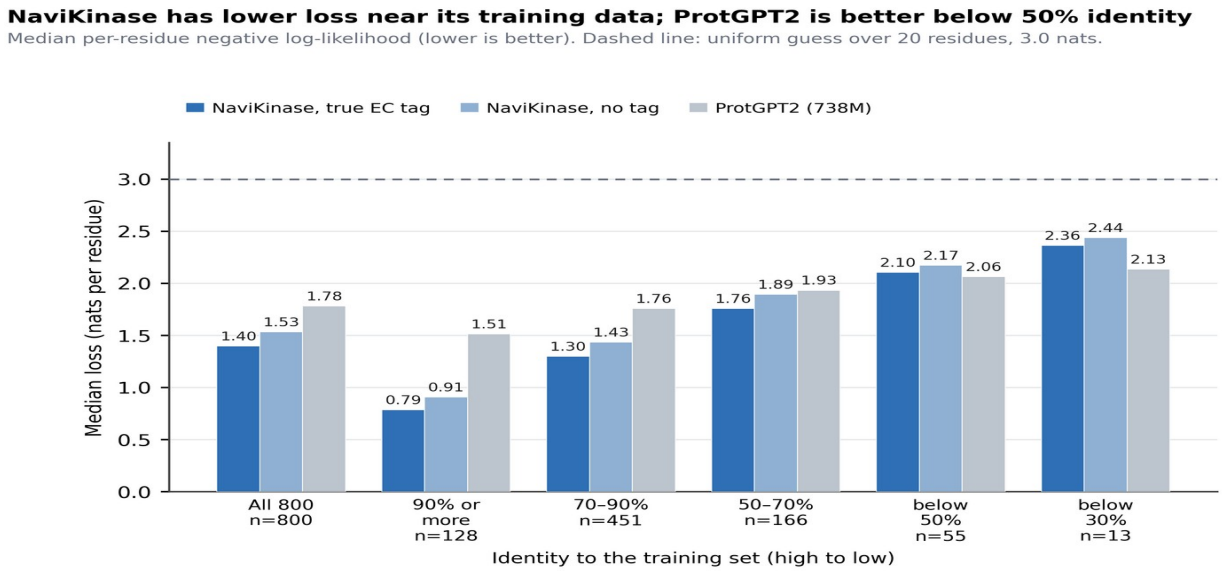


Figure 7. NaviKinase has lower loss near its training data; ProtGPT2 is better below 50% identity. Median per-residue negative log-likelihood in nats on the same 800 held-out sequences (lower is better), by identity to training. ProtGPT2 loss includes line-break tokens and is divided by the residue count. Dashed line: uniform guess over 20 residues, 3.0 nats. scripts/nk_nll.py and scripts/baseline_protgpt2.py.

Taken together, NaviKinase performs on par with same-size or larger general models at EC identification without needing labels at inference, and is better on likelihood near its training data, but it is not better on sequences dissimilar to its training set. Caveats: the broad training corpora of ESM-2 and ProtGPT2 (UniRef50) may include these test sequences or their relatives, which favours them; NaviKinase saw 2.45 million labelled sequences of this superfamily, which favours it; ProtGPT2's loss includes a small cost for line-break tokens; and no confidence intervals were computed for the likelihood differences. Generation quality and inference cost were not compared.

4. Discussion

4.1 What the evidence supports

- The model generalises to held-out clusters: perplexity of about 4.1 and 95.6% balanced EC identification (Section 3.1).
- Its EC signal survives low similarity to training data, at 91% for sequences below 50% identity, while likelihoods degrade (Section 3.3).
- The EC tag steers generation toward kinase motifs, and generated sequences are mostly novel and predicted to fold with confidence similar to natural kinases (Sections 3.2, 3.4 and 3.5).

4.2 What the evidence does not show

- **Function.** No generated sequence has been expressed or assayed, and ESMFold confidence is not evidence of activity.
- **Distant homology.** Identity estimates are lower bounds from a heuristic search, and only 55 sequences fall below 50% identity and 13 below 30%.
- **Advantage over other models.** Two baselines were run, ESM-2 150M [1] and ProtGPT2 [2], on EC identification and, for ProtGPT2, likelihood (Section 3.7). NaviKinase's 3-sequence lead in identification is within noise, it has higher loss than ProtGPT2 below 50% identity to training, and generation quality, inference cost and other tasks were not compared.
- **Robustness.** Samples are small (40 sequences per generation condition, 6 to 10 folded per group, 372 sequences for EC identification) and each analysis uses a single seed.
- **Novelty of every sample.** One of 40 EC 2.7.11 samples was a near-copy of a training sequence and a second matched at 90%, so some generated sequences are not new.

4.3 Further limitations

The EC-identification rule ranks tags by likelihood; the model was never trained as a classifier. The motif check is a regular-expression proxy limited to three protein-kinase motifs. Most EC 2.7.11 generations do not terminate on their own: 37 of 40 at seed 0 and 39 of 40 at seed 1 reach the 400-residue cap, including all 10 submitted for folding, so they carry truncated, low-confidence tails. Lengths under this conditioning are censored by the cap rather than chosen by the model (Section 3.5). Whether the model would emit a stop at a higher cap has not been tested. ESMFold was reached through a public API that timed out often, which excluded two folds. The model covers EC

2.7.- only and uses single sequences, with no alignment or structure input. The released chat.py is pattern matching, not language understanding.

5. Conclusion

NaviKinase 1.0 learned the statistics of the phosphotransferase family well enough to generalise to held-out clusters, to identify EC sub-subclasses of sequences that are only weakly similar to its training data, and to generate mostly novel sequences that ESMFold predicts will fold with confidence similar to natural kinases. The evidence supports a narrow, sequence-only generator and plausibility scorer. It does not show that generated sequences are functional, that the model outperforms general protein language models, or that it recognises remote homologs reliably.

The most valuable next steps are an exhaustive similarity search with MMseqs2 [12] or DIAMOND [13] in place of the 5-mer prefilter, baseline comparisons beyond these two models and tasks, including generation quality and inference cost, a larger ESMFold set compared with known kinase folds using Foldseek [14], and laboratory testing of selected sequences after biosecurity screening.

6. Declarations

Data and code availability

Model weights and inference code are at https://huggingface.co/NavitraTechnologies01/navikinase-1.0, released under CC BY-NC 4.0. The release contains model.safetensors (fp32, 598 MB), model.py, the frozen tokenizer.py for this checkpoint, config.json, and the generate.py, score.py, motifs.py, chat.py, nlu.py, research.py, lookup.py, structure.py and fold.py scripts.

Training data: UniProtKB entries with EC 2.7.- [5], deduplicated at UniRef90 [6]. The proteins in Section 3.6 were fetched from UniProt on 27 October 2026.

Evaluation scripts and raw outputs: the scripts listed in Table 4 and the contents of logs/navikinase_validation/ — the samples, identity searches, fold results and structure coordinates — are not part of the Hugging Face release. They are available from the corresponding author on request.

Software environment

Training: MLX 0.32.0 on Apple Silicon. Analysis: Python 3.12, NumPy 2.5.1, Biopython 1.88. Baselines: torch 2.14.1, transformers 5.21.0, scikit-learn 1.9.1.

Scripts and seeds

Table 4. Scripts and random seeds used for each result in Section 3.

Section	Scripts	Seed
3.1	scripts/eval_heldout.py	0
3.2	scripts/gen_samples.py, scripts/novelty.py	1
3.3	scripts/similarity_eval.py, scripts/novelty.py	3
3.4	scripts/fold_eval.py	1 (generation), 2 (shuffles)
3.5	generate.py, motifs.py, scripts/seed0_motif_rerun.py	0

Section	Scripts	Seed
3.6	score.py	0
3.7	scripts/baseline_esm2.py, scripts/baseline_protgpt2.py, scripts/nk_nll.py	4 (probe sample)

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The work was supported internally by Navitra Technologies Ltd.

Competing interests

The author is a director of Navitra Technologies Ltd, which developed NaviKinase 1.0 and licenses it commercially. NaviKinase 1.0 is released under CC BY-NC 4.0; commercial use requires a paid licence from Navitra Technologies Ltd. The author therefore has a financial interest in the model evaluated in this report.

Use of AI assistance

The evaluation scripts and this report were prepared with AI assistance (Claude, Anthropic). All reported numbers were produced by the scripts listed above. The author reviewed and takes responsibility for the content.

Biosecurity

Generated sequences carry no toxicity or hazard screening and must not be synthesised without independent expert review and a screening DNA provider.

Licences

NaviKinase 1.0 is released under CC BY-NC 4.0. ESMFold output is CC BY 4.0.

References

1. 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;379(6637):1123–1130.
2. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022;13:4348.
3. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. RoFormer: enhanced transformer with rotary position embedding. Neurocomputing. 2024;568:127063.
4. Zhang B, Sennrich R. Root mean square layer normalization. In: Advances in Neural Information Processing Systems 32. 2019. p. 12360–12371.
5. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–D531.
6. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–932.
7. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–197.
8. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915–10919.

9. 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;25(11):1422–1423.

10. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.

11. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 1995;9(8):576–596.

12. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–1028.

13. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.

14. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2):243–246.

Appendix A. Usage

NaviKinase speaks amino-acid sequences, not English: it has no chat template and no instruction following. Six commands cover most work (Table A1). Research and Fold need internet access; the rest run offline.

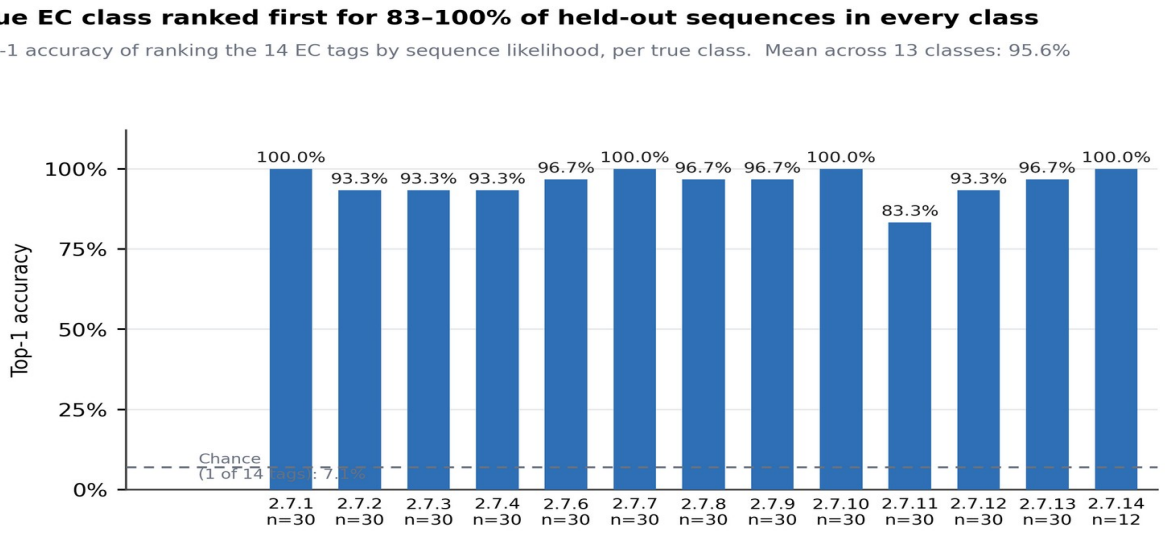
Two components are not part of the language model. chat.py is regular-expression and keyword matching that routes plain-English requests to the scripts below. fold.py calls the public ESMFold API; NaviKinase itself has no three-dimensional representation.

Table A1. The six NaviKinase 1.0 commands.

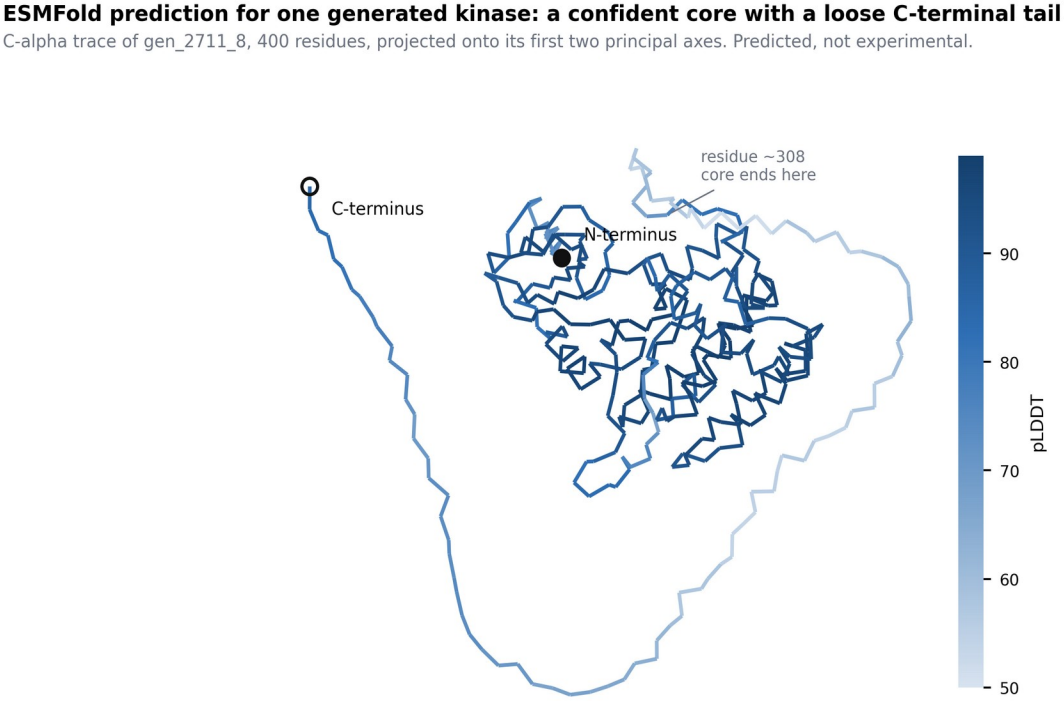
Task	Command	What it returns
Generate	python generate.py --n 5 --ecclass 2.7.11 --temperature 0.8	Novel sequences conditioned on an EC sub-subclass
Continue	python generate.py --n 3 --prefix MKTAYIAKQRQISFVKSHFSRQ	Completions of a real sequence prefix
Score	python score.py --sequence MKTAY... --per-position	Loss and perplexity, with per-residue loss to flag atypical stretches
Classify (experimental)	python score.py --sequence MKTAY... --classify	Loss under each of the 14 EC tags, ranked
Research	python chat.py --query "research protein P00519"	Fixed pipeline: UniProt fetch, score, classify, motif scan, PubMed search
Fold	python fold.py --sequence MKTAY... --out structure.pdb	3D structure from ESMFold, a separate Meta model; needs internet

Appendix B. Supplementary figures

These two figures support results stated in the main text.



Supplementary Figure S1. True EC class ranked first for 83–100% of held-out sequences in every class. Top-1 accuracy of ranking the 14 EC tags by sequence likelihood, per true class. Mean across 13 classes 95.6%; chance (1 of 14 tags) 7.1%. Class 2.7.99 had no eligible test sequences and 2.7.14 only 12. NaviKinase 1.0 on data/processed/test.bin; 372 held-out sequences, up to 30 per class, seed 0; scripts/eval_heldout.py. Summarised in Table 2 and Section 3.1.



Supplementary Figure S2. ESMFold prediction for one generated kinase: a confident core with a loose C-terminal tail. C-alpha trace of gen_2711_8, 400 residues, projected onto its first two principal axes; residue shading by pLDDT, which ranges from 50 to 99 (mean 85.4). The one folded EC 2.7.11 sample with all three catalytic motifs, chosen by that rule rather than by score. Coordinates (logs/navikinase_validation/gen_structure_ca.csv) are centred and re-oriented relative to fold.json, with identical geometry, so they should not be compared against raw ESMFold output. Predicted, not experimental. Discussed in Section 3.4.