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.

What enzyme-function scores hide: an audit of CLEAN with pre-registered tests

Prateek Paudel
Kennesaw State University
Prativa Khanal
prativakhanal1@gmail.com
Butwal Multiple Campus, Nepal

Research Article

Keywords: Enzyme function prediction, Enzyme Commission number, Evaluation metrics, Benchmark leakage, Label provenance, Calibration

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11732908/v1

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

Additional Declarations: The authors declare no competing interests.

What enzyme-function scores hide: an audit of CLEAN with pre-registered tests

Prativa Khanal[1*] and Prateek Paudel[2]

[1*]Butwal Multiple Campus, Butwal, Nepal.
[2]Kennesaw State University, Kennesaw, Georgia, United States.

*Corresponding author(s). E-mail(s): prativakhanal1@gmail.com;

Abstract

Context: Enzyme Commission (EC) number predictors such as CLEAN are evaluated with support-weighted precision and F1, which ignore the precision loss from predicted EC numbers absent from the test labels. This scoring, used in the public code of CLEAN, three subsequent methods and the EC-Bench benchmark, overlooks 67–96% of CLEAN's false positives on standard benchmarks, mostly incorrect fourth digits. Rescoring EC-Bench predictions for 16 methods showed score inflation rising with the number of EC numbers assigned and one significant pair reversal, while overall rankings remained similar (Kendall's $\tau_b = \mathbf{0.90}$). In a temporal benchmark of 2,099 Swiss-Prot enzymes added after CLEAN's training snapshot, 30.5% carried EC numbers unseen during training. For the remaining enzymes, inference from neighboring training members in CLEAN's embedding space outperformed its centroid rule (pre-registered: union-label macro-F1 **+0.043**), although the class-size correction added nothing. In descriptive analyses, BLASTp outperformed every embedding-based rule (micro-F1 0.735 vs 0.641 for the centroid rule and 0.703 for the best alternative, exemplar), with no significant difference from the centroid below 30% sequence identity. A subgroup analysis registered afterwards showed a BLASTp advantage on experimentally supported annotations (**+0.082**). CLEAN's confidence exceeded its accuracy by 0.16–0.30 for rare EC numbers, and model-based calibration budgets for 20% rare-tier error had lower bootstrap limits 4.5–6.7 times the available sample sizes. We propose an EC evaluation card for comparable, transparent reporting.

Methods: We audited the released CLEAN model without retraining, using ESM-1b embeddings, BLASTp, paired bootstrap tests and custom Python code on a temporal benchmark.

Keywords: Enzyme function prediction, Enzyme Commission number, Evaluation metrics, Benchmark leakage, Label provenance, Calibration

1 Introduction

Enzyme Commission (EC) numbers classify enzymatic reactions in a four-level hierarchical system [1]. An EC number is the standard unit by which the enzymatic activity of a protein is documented, compared and queried. Very few proteins are ever experimentally characterized. Thus, the majority of entries in the UniProt Knowledgebase are annotated through automatic procedures based on rules derived from experimentally characterized entries [2]. This means that computational assignment has become the default route to enzyme annotation, and any errors made during that process are propagated whenever its output is reused.

Learned predictors increasingly compete with homology transfer for this purpose, with CLEAN becoming a reference model among them [3]. CLEAN embeds each sequence using the ESM-1b protein language model [4], and the embedding is then mapped through a network trained with a contrastive objective. It then assigns EC numbers based on their training-set centroids, defined as the mean embedding of all training proteins associated with a given EC number, using a maximum-separation rule. In its original report, CLEAN was more accurate, reliable and sensitive than BLASTp. Since then, new models have followed a similar architecture and, in some cases, reused the same evaluation code. Thus, the numbers reported for CLEAN have become a yardstick, and the way in which they were produced shapes how a whole class of models is evaluated.

Subsequent independent evaluations have been less favorable. When experts examined the predictions of DeepECtransformer for more than 450 *Escherichia coli* proteins of unknown function, they found that machine learning largely failed to predict functions not observed during training and even made basic logical errors [5]. Across machine-learning-driven science, leakage between training and testing data is a major issue for reproducibility [6]. This suggests that a reported score may overestimate what a model can do on unseen proteins. However, it does not explain why.

A reported score is the result of more than the model alone. It also depends on the metric that turns predictions into a number, the test set on which the predictions are made, and the labels against which they are compared. For EC predictors, all of these factors have usually been taken for granted. Metrics are passed along with the evaluation code. Test sets are selected from the same curated database as the training data, sometimes years apart, and the fraction of EC numbers in a test set that a model has never seen is rarely disclosed. Labels used in benchmarks are considered ground truth, although many of them have been assigned by transferring function from a similar sequence. Each of these factors can affect a score, yet none of them is apparent from the score itself.

Deployment poses one more requirement. When an annotator is predicting EC numbers for a new genome, they should be aware not just of how accurate a model is on average, but also of when they can trust an individual prediction. Modern neural networks are often poorly calibrated, so their confidence scores cannot be interpreted as probabilities without verification [7]. This is particularly important for rare EC numbers. These are precisely the predictions that are most interesting, but they have the fewest labelled proteins against which to evaluate a model's confidence. Whether a

trustworthy prediction for a rare EC number is possible is therefore partly a question of how much labelled data is available.

A new architecture alone cannot satisfy this requirement. To determine whether a model has improved, the field first needs to understand how much of a reported score is due to the model and how much to the metric, the test set and the labels, as well as whether the confidence associated with a prediction can be trusted when data are scarce. We investigate these questions for CLEAN, the model on which much of the field builds. We reproduce the publicly available model without retraining it. We then study the support-weighted precision calculation in its released evaluation code and construct a temporal test set from enzymes added to Swiss-Prot after the snapshot used for training, excluding accession and identical-sequence overlap with CLEAN's training reference. On that test set, we compare CLEAN's centroid rule with nearest-member rules in the same frozen embedding space, as well as with a BLASTp baseline. We also investigate whether this comparison depends on label provenance through a subgroup analysis registered after the overall TEST comparison but before evidence-label retrieval. Finally, we quantify the amount of calibration data required to make a trustworthy prediction for a rare EC number. All confirmatory tests were specified in time-stamped analysis protocols before their corresponding analyses, with the timing of E7 reported separately (Section 3.1); comparisons with BLASTp outside that test are descriptive, and Appendix C lists every pre-registered hypothesis, its outcome and every deviation from the protocol. The methods are applicable to any EC predictor.

Each factor turns out to matter. The support-weighted precision in the publicly available code of CLEAN ignores the precision loss associated with 96% of CLEAN's false-positive predictions on its own Price-149 test set. Of the enzymes added to Swiss-Prot after CLEAN's training snapshot, 30.5% carry an EC number unseen by CLEAN. On the remaining enzymes, a standard BLASTp search against CLEAN's training sequences was more accurate than CLEAN in descriptive comparisons, and the subsequently registered subgroup analysis also found an advantage for BLASTp on labels with experimental evidence. Recalibration helped reduce the measured error in CLEAN's confidence estimates. For very rare predictions, a model-based calibration-budget analysis gave a lower bootstrap limit of 630 proteins, while only 102 were available. None of these findings concerns the architecture; all of them concern what a reported score measures and what it does not.

The rest of the paper is organized as follows. Section 2 discusses the previous work on which this audit is based and the gaps it leaves. Section 3 outlines the data, benchmark and analyses, Section 4 reports the results, Sections 5 and 6 discuss the implications and limitations, and Section 7 concludes.

2 Related work

2.1 Homology transfer and the reliability of EC labels

Homology transfer has been used to assign Enzyme Commission (EC) numbers for decades, and its reliability has been questioned for just as long. Rost [8] observed that enzyme function is less conserved between homologues than expected. Tian and Skolnick [9] estimated approximate thresholds. According to their estimates, about

40% sequence identity is needed for more than 90% of homologous pairs to share the first three EC digits, and about 60% identity is required for all four digits to match. Homology transfer is thus more reliable for the first digits of the EC hierarchy than for the last, and its errors tend to accumulate at the most specific level of the hierarchy.

Homology transfer also propagates errors. Across 37 enzyme families, misannotation was rare among manually reviewed Swiss-Prot annotations but common in other databases, including TrEMBL, GenBank NR and KEGG [10]. Annotation errors can also be widespread within a single class. In an experimental screen of 122 sequences, at least 78% of those annotated as EC 1.1.3.15 were inferred to be misannotated, and four alternative activities were confirmed [11].

Swiss-Prot provides the training labels for CLEAN and the reference labels for our benchmark, but not every label is backed by experimental evidence. The method used for each annotation is recorded using terms from the Evidence and Conclusion Ontology [12]: transfer from a characterised homologue ("by similarity", ECO:0000250), a family rule such as HAMAP (ECO:0000255), or experimental evidence (ECO:0000269). Therefore, the benchmark labels are partly the output of the same process used by a homology baseline. When a BLASTp top hit is compared with a learned model on such labels, part of the baseline's agreement may arise from the way those labels were assigned, whereas experimentally supported labels reduce direct dependence on homology transfer. In Gene Ontology prediction, the source of the annotations used for assessment significantly affected classifier accuracy, with computationally derived annotations increasing the measured accuracy [13]. To our knowledge, no comparison of EC predictors has separated test labels by their provenance in this manner.

2.2 Learned EC predictors

Machine-learning predictors were designed to overcome the limitations of pairwise transfer. ECPred used a separate classifier for each EC number [14]. It was also evaluated on a time-separated hold-out set, which proved very difficult, giving an F1 score of 0.14 at the substrate level. Deep sequence models followed. DeepEC used convolutional neural networks for accurate and high-throughput EC prediction [15], while ProteInfer used deep networks for protein function inference more generally [16]. DeepECtransformer used transformer layers and assigned EC numbers to 464 unannotated *E. coli* genes, three of which were validated [17]. The predictions of this model were later evaluated by experts, as noted in Section 1.

CLEAN moved the EC prediction task into a contrastively trained protein-language-model embedding space, and several subsequent models have built on it. CLEAN-Contact added predicted contact maps to the sequence embedding and reported improvements over CLEAN [18]. EnzymeHunter used hierarchy-aware contrastive learning for EC prediction [19]. Our analysis of their publicly available code shows that both reused CLEAN's support-weighted evaluation procedure; this is an observation about the code, not about their papers. Other work has tried to address the long-tailed distribution of labels directly. HIT-EC used a hierarchical, interpretable transformer to address the under-representation of EC numbers and incomplete annotations [20].

Two gaps arise from this work. Gains are reported using common evaluation conventions, so any bias in those conventions is carried over to the performance comparison of the entire family. In addition, CLEAN couples its learned embedding space with the centroid rule, so its reported accuracy combines the quality of the embedding space with the choice of inference rule. To our knowledge, no study has compared centroid and neighbour rules within one frozen EC embedding space.

2.3 Independent tests and the homology baseline

Several independent studies have attempted to test the claim made in CLEAN's paper: does a learned predictor perform better than homology transfer? Capela et al. [21] found that BLASTp was marginally better overall than protein-language-model classifiers, that the two approaches were complementary, and that language models were useful below about 25% sequence identity. A preprint by Sathyamoorthy and Puri [22] matched the BLASTp database to each model's training set. It reported comparable performance overall, but a clear advantage for language models on distant eukaryotic enzymes. Another preprint tested robustness rather than accuracy [23]. Both CLEAN and DIAMOND assigned the original EC number to more than 90% of decoy sequences with weakly perturbed active sites, suggesting that both homology transfer and CLEAN tend to over-assign function.

Together, these studies suggest that the gap between learned predictors and homology transfer is smaller than first reported and depends on the distance of a query from its nearest annotated neighbour. The peer-reviewed study reported that language models were useful below about 25% identity, while the training-set-matched preprint found a clear advantage for language models on distant enzymes. Neither study tested the released CLEAN model on enzymes annotated after its training snapshot, and none examined whether the test labels had themselves been assigned by similarity. If they were, the homology baseline would partly reproduce the curator's annotation process, and its advantage could partly be an artefact of the labels. Addressing this question therefore requires a temporal hold-out for the released model, a baseline searched against its own training sequences, and separation of labels by provenance.

2.4 Evaluation of function predictors: metrics, benchmarks and time

The Critical Assessment of Functional Annotation (CAFA) established time-delayed community evaluation of function prediction [24]. LAFA extended this approach to continuous, longitudinal evaluation under an evolving ground truth [25]. Time-delayed evaluation reduces supervised-label leakage when the training snapshot and annotation dates are verified. However, it does not prevent later labels from being assigned through homology transfer, which is the concern discussed in Section 2.1.

The choice of metric matters as much as the choice of test set. For Gene Ontology prediction, Plyusnin et al. [26] compared 37 evaluation metrics and found that some assign erroneously high scores to sets of false positives. This is the closest precedent for our metric analysis. There are two qualifications to consider in such an analysis. First, ground truth is incomplete, so a scored false positive could represent a true

function that has not yet been annotated [27]. Second, two commonly used macro-F1 averaging conventions can differ enough to rank classifiers differently, as an arXiv note demonstrated [28]. Thus, averaging conventions are not neutral, even for metrics that share the same name.

Recent EC benchmarks have become more systematic. EC-Bench unified the evaluation of EC predictors under one protocol, using support-weighted F1 to account for class imbalance [29]. CARE controls leakage using identity-based splits below 30% and between 30% and 50% sequence identity [30]. In support-weighted averaging, the precision of each EC number is weighted by its frequency among the test labels. An EC number that does not occur in the test labels therefore receives zero weight, and its false positives drop out of the score. This mathematics is straightforward, but to our knowledge, no study has analysed whether the averaging used to score EC predictors counts every false positive or how correcting it changes reported scores and method rankings. Identity-based splits, for their part, control sequence similarity but not time. They do not report the fraction of test enzymes carrying EC numbers unseen by the model, which bounds the recall of a closed-set model. They also do not treat training vintage as a property of each model, although a temporal test set that is clean for one model may overlap the training data of another.

2.5 Unseen and rare EC numbers, and trustworthy predictions

EnzPlacer used contrastive learning to predict EC numbers absent from its training data [31]. Its training data are more recent than CLEAN's, which illustrates the need to match training vintages for each model. Conformal prediction offers distribution-free guarantees for such calls. CPEC applies it to EC annotation with false discovery rate control [32], and Boger et al. [33] provided guarantees for protein mining at a user-specified risk, including searches using CLEAN's embeddings. These guarantees apply on average across test proteins, not separately within rarity tiers. Thus, the guarantees could be satisfied mainly through common EC numbers, while the error rate for predictions involving rare EC numbers remains unknown. To our knowledge, no study has estimated how many calibration samples a per-tier guarantee would require or whether curated databases can provide that amount of calibration data.

Thus, the literature leaves four questions open, each of which is important for the trustworthiness of reported EC scores. Does the metric used to evaluate EC predictors count every false positive, and how much does correcting it affect the reported rankings of methods? What fraction of a temporal test set falls outside a closed-set model's label space, and how should training vintage be matched? Within one frozen embedding space, how do centroid and neighbour rules perform compared with a homology baseline, and does the comparison hold when only experimentally supported labels are used? Finally, how much calibration data would a trustworthy prediction for a rare EC number require? The rest of the paper addresses these questions in turn.

3 Materials and methods

3.1 Design, model and data

We audited the released CLEAN model without any retraining. All analyses used the released split100 weights, training embeddings and predictions based on them. All confirmatory analyses were pre-registered in written, time-stamped analysis protocols that stated their hypotheses, endpoints and decision criteria before each corresponding analysis, with the timing of E7 disclosed separately below. For experiments E1–E6, this was the main protocol. All subsequent changes to the protocol were recorded (Table 9). The primary endpoint of H3b was changed from true-label to union-label macro-F1 before any TEST results were obtained (Section 3.3). Other changes, including the E5 secondary hypothesis and test direction, are listed in Table 9. The label-provenance analysis (E7) was pre-registered in a separate protocol. It was written after the descriptive BLASTp comparison on TEST (Section 3.3) was known, but before any label evidence was retrieved or the E7 analysis was performed. Each analysis output contains SHA-256 digests of the code and data that generated it. Each dataset was assigned one fixed role (Table 1). In particular, the calibration split (CAL) was the only data used for model tuning and calibration, while the test split (TEST) was never used for either purpose. TEST was later reused for the registered E7 analysis and post hoc analyses (Table 1). Appendix C lists all pre-registered hypotheses with their outcomes and all deviations from the protocols; analyses conducted after the confirmatory analyses are labelled post hoc throughout (Section 3.5).

CLEAN embeds a sequence using the mean ESM-1b representation from layer 33 (1,280 dimensions; sequences truncated to 1,022 residues), maps it to a 128-dimensional space, and predicts the EC numbers whose centroid distances are separated from the rest according to its maximum-separation rule. Its confidence is a two-component Gaussian-mixture score based on embedding distance. We computed the embeddings locally; an equivalence test against the released embeddings resulted in a maximum absolute difference of 1.76×10^{-7}. The reproduction was exact: 0 prediction mismatches out of 392 on New-392, with precision, recall and F1 of 0.5955, 0.4791 and 0.4967, respectively, bit-identical to CLEAN's own evaluation function. Price-149 confidences were also identical to four decimal places. Our code imports CLEAN's functions rather than copying them, because CLEAN is distributed for research use only.

The training reference for CLEAN and BLASTp is split100, containing 227,362 sequences, 5,242 EC numbers and 241,025 (EC, sequence) pairs from UniProt release 2022_01. CLEAN's embedding file stores rows in per-EC blocks, so the accession corresponding to each row cannot be recovered. A protein with several EC numbers appears in multiple blocks; therefore, co-annotated EC numbers were masked in every "other EC" computation. The split70 backbone (79,172 pairs, 5,093 EC numbers) served as a robustness check. New-392 and Price-149 are CLEAN's original test sets. Price-149 uses RefSeq identifiers and has low sequence identity to the training data, while New-392 contains 5 sequences identical to training sequences.

The SP-temporal benchmark was constructed from Swiss-Prot entries created on or after 1 March 2022, using UniProt release 2026_03 [34] and the corresponding

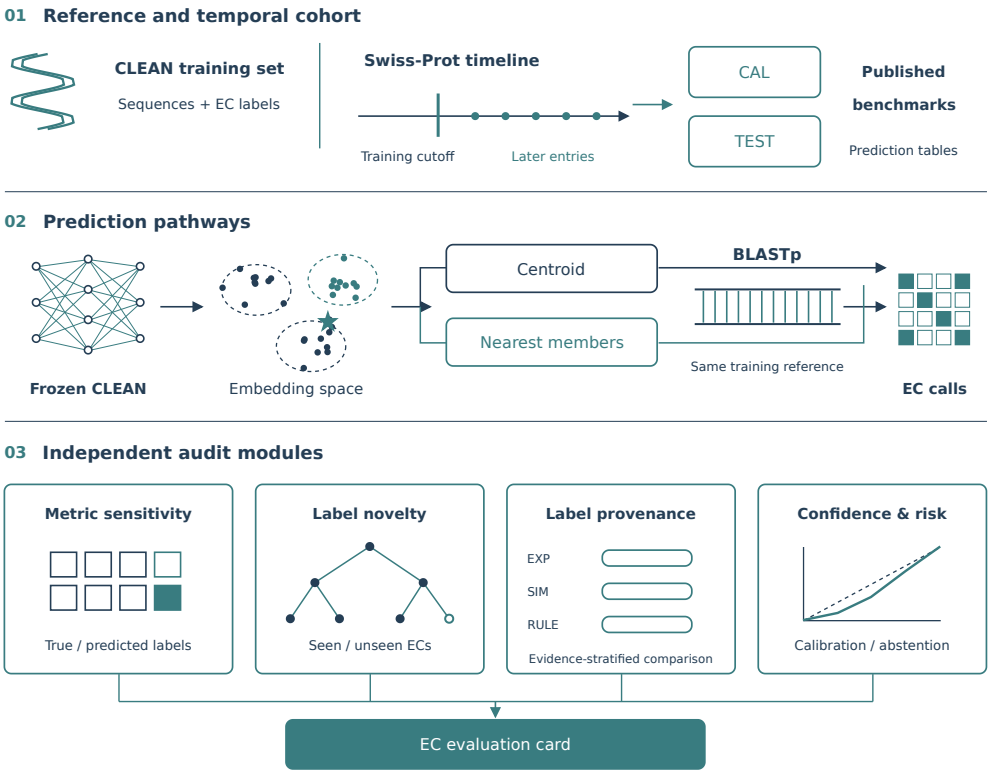


Fig. 1 Overview of the CLEAN audit: temporal benchmarking, inference comparisons, and evaluation of metric sensitivity, label novelty, annotation provenance, and calibration. Illustrations are schematic

ENZYME release [35]. Entries were selected using UniProt's `date_created` field, which does not establish when a sequence first became publicly available. Of the 1,260 TEST accessions, 906 have the older six-character format; accession format alone does not indicate prior sequence exposure, which we did not quantify. These annotated protein entries were absent from CLEAN's training reference, which is the requirement of the benchmark. An entry was selected if it was not a fragment, contained 30–1,022 residues, had an accession absent from all CLEAN splits and released test sets, had a sequence not identical to any CLEAN training sequence or another selected entry, and carried at least one complete EC number. Entries containing only partial EC numbers, or an enzyme keyword without an EC number, were excluded. Of 8,920 raw entries, 2,099 enzymes passed this filtering. They were randomly split 40/60 into CAL (839) and TEST (1,260), stratified by rarity tier, using seed 20260927. TEST includes 876 seen-EC and 384 unseen-EC proteins (331 pure, 53 mixed). A further 2,000 non-enzymes were sampled from 4,058 eligible entries. Labels are current complete EC numbers. Predictions from all methods, including BLASTp, which transfers 2022-era training labels, were mapped to current nomenclature using the ENZYME transfer

table before scoring; reclassifications without a recorded transfer were treated as open-set.

The rarity tier of a protein is determined by the split100 training count of its rarest true EC number after transfer mapping: unseen (0), singleton (1), very rare (2–3), rare (4–10) or common (more than 10). A protein is classified as seen-EC if all its true EC numbers are present in CLEAN's vocabulary; an unseen-EC protein is pure if none is present and mixed otherwise. The construction imposes two limitations. First, one family dominates the unseen tier: 78 fungal lytic polysaccharide monooxygenases with EC 1.14.99.56. Second, because CAL and TEST are random splits of the same pool, 27% of TEST proteins have at least 50% sequence identity to a CAL protein. This is addressed using the CAL-disjoint sensitivity subset.

Finally, we used the per-protein predictions of 16 methods in EC-Bench on its test set (468 proteins) and on Price-149, under both the 100% and 30% identity regimes. The predictions were obtained from a fixed version of the EC-Bench repository, verified by SHA-256 digests, and parsed exactly as in EC-Bench's scoring code. As a reproduction check, our scoring tool reproduced the support-weighted F1 values reported by EC-Bench for all 16 methods at EC levels 1 and 4, with a maximum absolute difference of 0.

Table 1 Datasets and their roles. CAL was the only set used for tuning and calibration. TEST outcomes motivated E7 and the explicitly post hoc analyses, but were never used to tune inference rules or perform calibration.

Dataset	Size	Role
split100	227,362 sequences; 5,242 ECs	Training reference for CLEAN and BLASTp database
split70	79,172 pairs; 5,093 ECs	Robustness backbone
New-392	392	Development set (disclosed); E7 pilot
Price-149	149	Confirmatory for rules fixed before confirmatory analysis; post hoc for class-balanced rules
SP-temporal CAL	839	Only set used for tuning and calibration
SP-temporal TEST	1,260 (876 seen-EC)	Evaluation; subsequently reused for registered E7 and post hoc analyses
Non-enzymes	2,000	Open-set evaluation
EC-Bench predictions	16 methods; test 468	Test set confirmatory for E5; Price-149 predictions used as E5 pilot

3.2 Endpoint metrics and the support-weighted blind spot

For EC depth $d \in \{1, 2, 3, 4\}$, obtained by truncating EC numbers to their first d fields, let T_i and P_i be the true and predicted EC sets of protein i, and let TP_c, FP_c and FN_c be the true-positive, false-positive and false-negative counts for class

c, respectively. Using the standard averaging schemes for multi-label metrics [36], we used three endpoint metrics:

$$\text{micro-F1} = \frac{2\sum_c \mathrm{TP}_c}{2\sum_c \mathrm{TP}_c + \sum_c \mathrm{FP}_c + \sum_c \mathrm{FN}_c}, \tag{1}$$

$$\text{union-label macro-F1} = \frac{1}{|\mathcal{C}_T \cup \mathcal{C}_P|} \sum_{c \in \mathcal{C}_T \cup \mathcal{C}_P} \mathrm{F1}_c, \tag{2}$$

$$\text{per-protein F1} = \frac{1}{N} \sum_{i=1}^{N} \frac{2|T_i \cap P_i|}{|T_i| + |P_i|}, \tag{3}$$

where $\mathcal{C}_T$ and $\mathcal{C}_P$ are the classes occurring among the true and predicted labels, respectively. Support-weighted F1 weights each class F1 by its true support $n_c = \mathrm{TP}_c + \mathrm{FN}_c$, as implemented by `average="weighted"` in scikit-learn:

$$\text{support-weighted F1} = \sum_{c \in \mathcal{C}_T} \frac{n_c}{\sum_{c'} n_{c'}} \mathrm{F1}_c. \tag{4}$$

True-label macro-F1, in contrast, averages $\mathrm{F1}_c$ over $\mathcal{C}_T$ without support weighting. These two scores are reported only for comparability and are labelled as blind to false positives outside the test label set and as gameable. Classification performance measures differ in their invariance properties [37], and the following invariance is relevant here.

Proposition 1 *In any label space, support-weighted precision and F1, as well as true-label macro scores, assign zero weight to any class with no true instance. They are therefore invariant to predictions of EC numbers absent from the test labels. Support-weighted recall equals micro-recall. Consequently, the omitted penalty for these false positives concerns precision; differences between weighted F1 and micro-F1 also depend on aggregation.*

The proof is provided in Appendix A. The proposition is elementary; our contribution is to estimate its effect and document its presence in public evaluation code. A false positive ignored by support-weighted F1 is still an incorrect prediction and increases the micro-precision denominator without increasing its numerator. If the true EC number is also missed, that miss is still counted against recall; however, the precision penalty associated with the wrong prediction is omitted. Abstentions contribute no predicted labels and count as false negatives when true labels are present. Scores computed after projecting predictions onto the test label set are invariant to such predictions by construction. We verified the proposition numerically on 82 real prediction tables and through unit tests. We define false positives associated with EC numbers outside the test label set as *invisible false positives*. To demonstrate how the blind spot can be exploited (E1), we emitted the top-k EC numbers for each protein, with $k = 1, \ldots, 10$, and measured every score. We also examined the public evaluation code of CLEAN, CLEAN-Contact, ProtDETR, EnzymeHunter and EC-Bench at

fixed versions; the functions examined are listed in Appendix Table 10. For any prediction file, we report an EC evaluation card containing, at each EC depth, micro-F1, union-label macro-F1 and per-protein F1, the support-weighted score with its count of invisible false positives, the number of abstentions and the sizes of the label spaces.

In E5, we rescored the EC-Bench predictions, which include PRIAM [38]. The preregistered hypothesis H5 states that at least one of the 120 method pairs (test set, 100% regime, EC level 4) is ranked one way by support-weighted F1 and significantly the opposite way by micro-F1. Significance required a Benjamini–Hochberg $q < 0.05$ across the 120 pairs and agreement between the bootstrap median and the observed difference in sign. The key secondary hypothesis, revised after the disclosed pilot and before confirmatory scoring, was the one-sided Spearman correlation between score inflation (support-weighted F1 minus micro-F1) and the average number of EC numbers predicted per protein. Other secondary analyses included Kendall's τ_b with a bootstrap interval, invisible false-positive shares, the 30% identity regime, EC levels 1–3, and a variant restricted to complete EC numbers.

3.3 Inference rules, the homology baseline and label provenance

In E2, we evaluated scoring rules in CLEAN's frozen embedding space, leaving its embeddings and maximum-separation rule unchanged. CLEAN's *centroid* rule scores an EC number by its distance to the mean embedding of the corresponding training members, making it a prototype classifier [39] built on a supervised contrastive embedding [40]. The alternative rules were the distance to the nearest member (*exemplar*); the mean distance to the three nearest members of each EC class (*kmean3*; all members for classes with fewer than three); and silhouette-gated k-means sub-prototypes (*multiproto*) for EC numbers with at least 20 members ($k \leq 4$, silhouette ≥ 0.25, each cluster containing at least 10% and at least 3 members). The class-balancing rules *balanced1* and *balanced3* scale the exemplar or kmean3 distance by n_c^{λ}, where n_c is the EC number's training size. This equal-prior density correction is motivated by the use of prototypes for long-tailed recognition [41]. The parameter λ was selected using CAL only, by maximising union-label macro-F1 at EC level 4 on seen-EC proteins over the grid $\{0, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.15, 0.2\}$. The primary endpoint (H3b; preregistered, with the metric switched to union-label macro-F1 before any TEST result was available) is the balanced3-minus-centroid difference in union-label macro-F1 at EC level 4 on the 876 seen-EC TEST proteins. Micro-F1 is the key secondary endpoint, per-tier tests are Holm-corrected, and split70 serves as the robustness backbone. The sensitivity analyses included a cluster bootstrap based on each protein's rarest true EC number (proteins sharing an EC number also share a centroid), a limit of 5 proteins per rarest EC number, the CAL-disjoint subset (< 50% identity to every CAL protein), and proteins with evidence at the protein level. In addition, E3 analysed the dispersion of EC numbers whose training members belonged to non-homologous Pfam-clan groups, representing candidate non-homologous isofunctional enzymes [42]. This analysis used a size-stratified permutation test with 10,000 permutations, testing distances to their own and to other EC numbers separately.

The homology baseline was BLASTp from BLAST+ 2.12.0 [43], an implementation of gapped BLAST [44] using default scoring parameters (BLOSUM62, gap opening penalty 11, gap extension penalty 1, and composition-based statistics), an E-value threshold of 10, and up to 500 subjects per query. The database was CLEAN's split100 training set, so both methods used the same training proteins. A query received the EC numbers of its best hit by bit score, with ties broken by lower E-value and then by subject accession; a query without a hit abstained. Sequence identity was calculated as the number of identical residues divided by the length of the shorter sequence, maximised over high-scoring segment pairs (HSPs). It was grouped into $< 30\%$, 30–50%, 50–70%, 70–90% and $\geq 90\%$ identity bins. The same search against CAL defined the CAL-disjoint subset. The hit tables were stored with SHA-256 digests of the searched sequences and checked against them before analysis. BLASTp replaced an approximate aligner (a k-mer prefilter followed by Smith–Waterman alignment of the top candidates) after the confirmatory analysis. The protocol had already recommended a standard external aligner for the final analysis, but the change was made after the results had been seen, and it affected the comparison. With the approximate aligner, the top-hit baseline achieved a seen-EC micro-F1 of 0.600 on TEST (CLEAN 0.641) and 0.505 on New-392 (CLEAN 0.498), compared with 0.735 and 0.584 with BLASTp. The approximate aligner missed some hits; it placed 36–38% of TEST and Price-149 proteins below 30% identity, compared with 22.6% of TEST proteins with BLASTp. This change affects the comparator, identity strata and CAL-disjoint subset, but not a pre-registered endpoint. All paired bootstrap comparisons between BLASTp and CLEAN outside E7 are descriptive.

For E7, we retrieved the current UniProtKB entry for every protein in New-392, CAL, TEST and the EC-Bench test set through the UniProt Representational State Transfer (REST) application programming interface (API), using a single release recorded with SHA-256 digests. Accessions were resolved to the primary accession, then to the unique active entry listing the query as a secondary accession, or finally to the entry into which the query was merged; all other cases were counted as not found. Price-149 was excluded because its proteins are not UniProtKB entries and its labels come from mutant-fitness experiments. The evidence for an EC label was defined as the set of Evidence and Conclusion Ontology codes (Section 2.1) attached to that EC number in the protein names, including domains, and in catalytic-activity annotations. Each label was classified, in order of precedence, as EXP (ECO:0000269, experimental), SIM (ECO:0000250, by similarity, citing a source protein), RULE (ECO:0000255, ECO:0000256 or ECO:0000259, a match to a sequence model such as a HAMAP rule), or OTHER. A protein was classified as LABEL_ABSENT if its entry was not found or any benchmark label was missing from it; otherwise, it was classified as SIM if any label was SIM, EXP if every label was EXP, RULE if any label was RULE, and OTHER otherwise. The PROPAGATED category contains SIM $\cup$ RULE. This design complements the time-delayed evaluation of newly accumulated annotations used in CAFA2 [45] and continued in CAFA3 [46].

E7 compares the BLASTp top hit with CLEAN at EC level 4 on the 876 seen-EC TEST proteins. The two primary endpoints form one Holm-corrected family. P7a is Δ_{EXP}, defined as the micro-F1 of BLASTp minus that of CLEAN on EXP proteins

(two-sided). P7b is the difference-in-differences:

$$\mathrm{DiD} = \Delta_{\mathrm{PROPAGATED}} - \Delta_{\mathrm{EXP}}, \tag{5}$$

with EXP and PROPAGATED proteins independently resampled (seeds 20260927 and 20260928). These primary endpoints are confirmatory only if gate G7 holds ($n_{\mathrm{EXP}} \geq 100$ and $n_{\mathrm{PROPAGATED}} \geq 50$); otherwise, all E7 results are descriptive. The secondary analyses included the SIM and RULE difference-in-differences separately (Holm-corrected over the two), identity-standardised versions of P7a and P7b, per-protein F1 instead of micro-F1, and the difference-in-differences for BLASTp minus EC-Bench's CLEAN retraining on the EC-Bench test set (descriptive unless both classes contain at least 30 proteins). Identity standardisation weights each bin's difference by the pooled EXP and PROPAGATED bin distribution:

$$\Delta^{\mathrm{std}} = \sum_b w_b \, \Delta_b, \qquad w_b = \frac{n_{\mathrm{EXP},b} + n_{\mathrm{PROP},b}}{n_{\mathrm{EXP}} + n_{\mathrm{PROP}}}, \tag{6}$$

where bins containing fewer than 10 proteins in either class were merged using a fixed rule (from the highest bin downward into the next lower bin, then from the lowest upward), and each (class, bin) cell was resampled independently. For SIM proteins, we additionally measured how often the BLASTp top hit was a source protein cited by the label, how often a cited source was among the retained hits or in CLEAN's training set, and the micro-F1 of both methods when the top hit was or was not a cited source, with exact binomial intervals [47]. The sensitivity analyses included an EC-cluster bootstrap and a limit of 5 proteins per rarest EC number. The interpretation of each combination of outcomes was fixed before the analysis (Appendix Table 7). New-392 served as the pilot, which could motivate changes to parsing and classification rules through dated amendments, but not to the endpoints.

3.4 Calibration and the certification budget

For the reliability analyses (E4 and E6), we used top-1 predictions under the centroid, exemplar and balanced3 rules and grouped proteins by the rarity tier of the *predicted* EC number. Calibration was measured using the expected calibration error (ECE) over 10 equal-width bins and the Brier score, with reliability diagrams showing exact per-bin intervals. We compared CLEAN's released confidence with isotonic regression [48] fitted on CAL, either globally or separately for each predicted tier. For certified abstention, we applied Selection with Guaranteed Risk (SGR) [49] to the raw distance, with $\delta = 0.1$ and $\alpha \in \{0.1, 0.2, 0.3\}$, using either one global threshold or one threshold per predicted tier, with δ Bonferroni-split over four tiers. Class-conditional guarantees of this kind are known to be difficult to obtain with many classes and few examples per class [50]. We report the realised selective risk on TEST with error counts and exact intervals, abstention on pure and mixed unseen-EC proteins and on non-enzymes, the area under the receiver operating characteristic curve (AUROC) of each score for distinguishing seen-EC enzymes from non-enzymes, and an EC-level-3 fallback fitted on the CAL proteins rejected by the level-4 rule. The standard SGR

guarantee assumes a fixed classifier and confidence function, as well as independent and identically distributed calibration data. Because balanced3 was tuned on CAL and the fallback reuses CAL after data-dependent rejection, we report the risk and coverage of these methods empirically. We do not claim additional finite-sample guarantees for these reuse steps or a joint guarantee across EC levels.

E6 estimates calibration-data requirements at a fixed operating point; it does not calculate the power of SGR's adaptive threshold search. For n accepted proteins, assumed selective risk r and per-test failure probability δ', define

$$\pi(n; r, \alpha, \delta') = \Pr_{K \sim \mathrm{Bin}(n,r)} \left[U_{\mathrm{CP}}(K, n; \delta') < \alpha \right], \tag{7}$$

where U_{CP} is the one-sided Clopper–Pearson upper bound. Binomial power has discrete downward steps. We therefore used n_{stable}, defined as the first accepted-sample size after the last value with power below 0.8 in the numerical search window, rather than the first crossing of 0.8. The window was expanded until its final fifth remained above 0.82, subject to a cap of 200,000 accepted proteins. For coverage c and tier t, the total calibration budget $m_t(c)$ was calculated by fixed-point iteration starting from $m = 2$:

$$\begin{aligned} m &\leftarrow \max\left\{ m, \left\lceil n_{\mathrm{stable}}\big(\hat{r}_t(c), \alpha, \delta'(m)\big)/c \right\rceil \right\}, \\ \widehat{m}_t &= \min_{c \in \mathcal{G}_t} m_t(c), \qquad \delta'(m) = \frac{\delta}{4 \max\{1, \lceil \log_2 m \rceil\}}. \end{aligned} \tag{8}$$

Here, $\mathcal{G}_t$ is the 0.1–0.9 coverage grid restricted to points with at least 10 accepted CAL proteins. Selective risks were estimated using the Jeffreys posterior mean (errors + 0.5)/(accepted + 1), a rule selected after the disclosed pilot. Values with estimated risk at least α, or without a solution within the search cap, were assigned an infinite budget. Thus, an infinite estimate does not establish that selection is impossible for every threshold or sample size.

The pre-registered primary endpoint H6 was the estimated deficit $\widehat{m}_t - m_{t,\mathrm{available}}$ at $\alpha = 0.2$ in the very-rare and singleton predicted tiers, with a bootstrap 95% interval (1,000 resamples; infinite estimates retained). Selecting favourable coverage makes this an optimistic planning calculation, not a proven lower bound on SGR's data requirements. We checked its relation to SGR through simulation. To illustrate the magnitude of the deficit, we divided it by the approximately 466 qualifying enzymes added annually between March 2022 and September 2026, multiplied by the tier's share of CAL predictions. This calculation assumes that future data resemble CAL and is not a forecast.

3.5 Statistical analysis and reproducibility

Comparisons used a paired nonparametric bootstrap over proteins, with the same resamples for every method compared (B = 2,000, seed 20260927). The reported effects are the observed full-sample differences with 95% percentile intervals. The two-sided bootstrap p-value was calculated as $p = 2\min\{\Pr^*(\hat{d}^* \le 0), \Pr^*(\hat{d}^* \ge 0)\}$, where $\hat{d}^*$ is the difference in a resample, with the value capped at 1. A p-value of 0 was reported as $p < 1/B$. The bootstrap mean of a macro-averaged metric is biased, so

it was never reported as the effect. As a post hoc precision summary, we report an approximate minimum detectable effect at 80% power for a two-sided normal-theory test at nominal $\alpha = 0.05$:

$$\mathrm{MDE}_{80} = 2.80 \times \widehat{\mathrm{SE}}_{\mathrm{boot}}. \tag{9}$$

This approximation uses $z_{0.975} + z_{0.8} \simeq 2.80$; it is not a hard detection threshold, a confidence bound or a multiplicity-adjusted power calculation. Multiplicity was controlled using the Holm step-down procedure [51] across tiers and within each pre-registered family, and the Benjamini–Hochberg procedure [52] across the 120 EC-Bench pairs. Risks and shares were reported with exact binomial intervals. Proteins sharing an EC number were handled using a cluster bootstrap. After the confirmatory analysis, an error in the cluster bootstrap was corrected so that per-protein F1 was divided by each resample's total weight rather than the number of proteins. The error affected only the per-protein F1 rows of that sensitivity analysis, none of which is reported here. We report the data, optimisation, model and evaluation following the DOME recommendations for supervised machine learning in biology [53]. Table 2 lists the primary endpoints, and Appendix C provides the full register.

After the confirmatory analyses, we added post hoc analyses using only the stored outputs of the original analyses (Table 6). None was pre-registered, and each is labelled post hoc wherever reported. These analyses included micro-precision, micro-recall and per-protein F1 for each rule and BLASTp; Holm-adjusted tests across the five identity bins; the EC depth at which invisible false positives agree with a true label of the same protein; bootstrap intervals for per-tier ECE and the confidence–accuracy gap (2,000 protein resamples within each tier); the unseen-EC share excluding EC 1.14.99.56 and across sequence clusters; yearly accrual; the evidence codes of the OTHER class excluded from E7; and a sequence-cluster bootstrap. For the latter, all TEST proteins were aligned against each other using the approximate aligner in the analysis code, and proteins were grouped by single linkage at $\geq 50\%$ identity over the shorter sequence, yielding 966 clusters (699 among the 876 seen-EC proteins and 419 among the 477 EXP proteins in E7). The paired bootstrap was then repeated by resampling clusters instead of individual proteins. Because the approximate aligner misses some hits, these clusters may be finer than those obtained with BLASTp, making this check less conservative than BLASTp-based clustering.

Table 2 Pre-registered primary endpoints. Paired performance comparisons use 2,000 bootstrap resamples at EC level 4; H6 uses 1,000 resamples for its budget-deficit interval. The full register, including secondary hypotheses, outcomes and all deviations, is provided in Appendix C.

ID	Hypothesis	Metric	Population	Test and correction	Registered
H3b (E2)	balanced3 beats centroid	union-label macro-F1	TEST seen-EC ($n = 876$)	two-sided, $\alpha = 0.05$; tiers Holm	Before the analysis; metric amended before any TEST result
H5 (E5)	≥ 1 pair reversed between support-weighted F1 and micro-F1	both	EC-Bench test, 100% regime, 120 pairs	BH $q < 0.05$ with sign agreement	Before the analysis
H6 (E6)	budget exceeds available CAL proteins at $\alpha = 0.2$	deficit, Eq. (8)	very-rare and singleton predicted tiers	bootstrap 95% CI (1,000)	Before the analysis
P7a (E7)	$\Delta_{\mathrm{EXP}} \neq 0$	micro-F1	TEST seen-EC, EXP	two-sided; Holm over P7a, P7b; gate G7	Before the E7 analysis; after the BLASTp result
P7b (E7)	DiD $\neq 0$	micro-F1, Eq. (5)	TEST seen-EC, EXP and PROPAGATED	two-sided; Holm over P7a, P7b; gate G7	Before the E7 analysis; after the BLASTp result

4 Results

4.1 The reported metric misses the precision cost of most false positives

The released model was reproduced exactly. On New-392, our predictions differed from CLEAN's for 0 of 392 proteins, and CLEAN's own evaluation function returned precision, recall and F1 of 0.5955, 0.4791 and 0.4967, the values reported in the original paper [3]. Our ESM-1b embeddings matched the released ones with a maximum absolute difference of 1.76×10^{-7}, our scorer reproduced all 32 weighted F1 values reported by EC-Bench (16 methods at EC levels 1 and 4) exactly, and Proposition 1 held on all 82 prediction tables we scored.

Most of CLEAN's false positives do not enter its reported precision (Fig. 2, Table 3). These are predictions of EC numbers that do not occur among the test labels: 150 of 224 on New-392 (67%), 106 of 110 on Price-149 (96%) and 616 of 830 on the SP-temporal TEST set (74%). On Price-149, precision falls from 0.584 to 0.392 and F1 from 0.495 to 0.426 once every prediction is counted. On New-392 and TEST, the two

F1 scores are almost equal (0.497 against 0.498, and 0.443 against 0.446), because the precision gap is offset by differences in averaging. This agreement is not a property of the metric, since on Price-149 the two F1 scores differ by 0.07. We had predicted that the invisible share on TEST would be below New-392's 0.67, since TEST has a larger label space. It was 0.74, so this pre-registered secondary prediction (S7) was not met: the size of the label set alone does not determine how many false positives the metric misses.

Most invisible false positives are near misses rather than unrelated functions (post hoc, Table 6). For CLEAN's centroid rule, 60%, 67% and 63% of them on New-392, Price-149 and TEST share the first three EC digits with a true EC number of the same protein, while 15%, 9% and 20% fall into a different EC class altogether. The metric thus mostly misses incorrect fourth digits, which are the errors that matter most for substrate-level annotation.

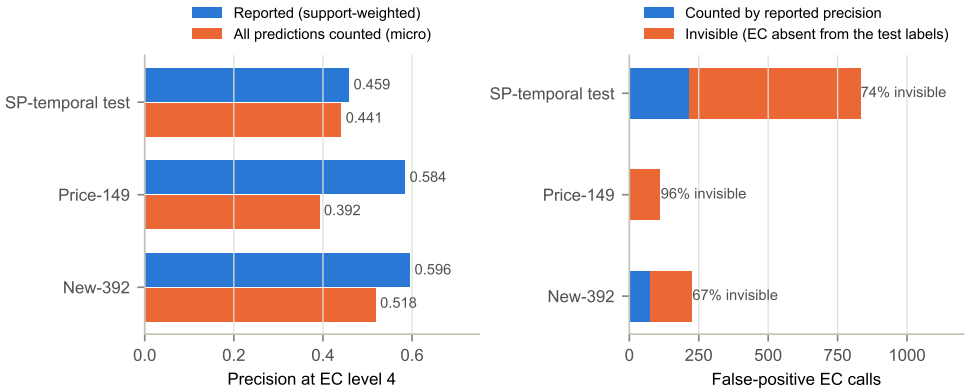


Fig. 2 The blind spot of support-weighted precision for the released CLEAN model (maximum-separation rule, EC level 4). Left: precision as reported (support-weighted over the test labels) and with every prediction counted (micro). Right: false-positive EC calls split into those the reported precision counts and those it cannot see because the predicted EC number is absent from the test labels

Table 3 Reported and micro-averaged scores of the released CLEAN model at EC level 4. "Invisible" false positives are predictions of an EC number that no protein in the test set carries; their precision cost is ignored by the reported score.

Test set	Reported P / R / F1	Micro P / R / F1	False positives	Invisible
New-392	0.596 / 0.479 / 0.497	0.518 / 0.479 / 0.498	224	150 (67%)
Price-149	0.584 / 0.467 / 0.495	0.392 / 0.467 / 0.426	110	106 (96%)
SP-temporal TEST	0.459 / 0.451 / 0.443	0.441 / 0.451 / 0.446	830	616 (74%)

In the evaluated datasets, score increases under over-prediction occurred only on Price-149. When we emitted the top 10 EC numbers for every Price-149 protein, the reported precision rose from 0.519 (top 1) to 0.636 and the reported F1 reached 0.618, above CLEAN's own 0.495, while micro-precision fell from 0.403 to 0.070 and micro-F1 was 0.127. On New-392 and TEST, the reported precision fell under the same manipulation (0.560 to 0.482, and 0.453 to 0.279), though far less than micro-precision (0.541 to 0.093, and 0.479 to 0.067). The first half of hypothesis H1, that the reported score ignores most false positives, held on every test set; the second half, that it rises under over-prediction, held only on Price-149. CLEAN's p-value rule shows the same blind spot (247 of 398 false positives invisible on New-392, 62%). New-392 also contains some label noise: 5 of its proteins are identical to training sequences, and 7 of its 392 label sets have changed since 2022 (3 changed, 2 gained an EC number, 1 lost one, and 1 is no longer an enzyme entry).

The scoring code of three successors, CLEAN-Contact, ProtDETR and EnzymeHunter, is a verbatim copy of CLEAN's evaluation function. EC-Bench weights by support over the union of true and predicted labels, which means that an EC number appearing only among the predictions receives zero weight. All four are therefore invariant to predictions outside the test label set (Appendix Table 10). To examine whether this changes conclusions, we re-scored the per-protein predictions released by EC-Bench for 16 methods [29] (Appendix Table 15). The pre-registered hypothesis H5 was met by a single pair of low-scoring methods. DeepECtransformer-learnt ranks above PRIAM by the reported F1 (0.115 against 0.099) but significantly below it by micro-F1 (-0.041 [-0.075, -0.006], BH $q = 0.026$); the same pair also reversed significantly under union-label macro-F1. The reversal reflects abstention as much as invisible false positives: DeepECtransformer predicts about 1.1 EC numbers per protein, and 393 of its 452 false positives are invisible, whereas PRIAM abstains on 402 of 468 proteins and is correct on 64% of the calls it makes. Restricting labels to complete EC numbers gave one significant reversal of a closely related pair (DeepECtransformer-regular against PRIAM), and per-protein F1 reversed two other pairs (ECRECer against the two ProteinBERT variants). Six pairs changed order under micro-F1, five of them without reaching significance, and the two rankings otherwise agreed closely (Kendall's $\tau_b = 0.90$ [0.83, 0.93]). Across methods, inflation of the reported score was associated with the number of EC numbers predicted per protein (Spearman's $\rho = 0.75$, one-sided $p = 0.0004$; key secondary met), although the 16 methods include regular and learnt variants of the same architectures and are not independent. The median share of invisible false positives was 0.65 (range 0.25–0.87) on EC-Bench's test set and 0.96 on its Price-149 split. In the 30% clustering regime, agreement was lower ($\tau_b = 0.82$), and three pairs reversed significantly. In the 100% regime, union-label macro-F1 showed even lower agreement with the reported ranking ($\tau_b = 0.57$, four significant reversals).

4.2 What a temporal test set contains

Of the 2,099 enzymes curated into Swiss-Prot after CLEAN's training snapshot that passed our filters, 640 (30.5%) carry at least one EC number absent from CLEAN's vocabulary. In TEST, these include 384 proteins, 331 with no EC number known to

CLEAN and 53 that also carry a known one. Of the unseen-EC proteins, 209 carry an EC number higher than any number CLEAN knows in its sub-subclass, which most likely indicates a number created after 2022, and 173 carry an existing EC number with no training member. Seven carry an unseen EC number in a sub-subclass absent from the training vocabulary. These categories are non-exclusive because a protein can carry multiple unseen EC numbers. Part of this novelty is therefore due to nomenclature rather than newly discovered chemistry. These closed-set methods cannot predict EC numbers absent from their reference vocabulary; on unseen-EC proteins, exact-EC micro-F1 was 0.02–0.03 for every rule. One family is particularly prominent: the 45 TEST proteins with EC 1.14.99.56, lytic polysaccharide monooxygenases, make up 3.6% of TEST. Excluding them, the unseen share of TEST is 27.9%, and when counted over sequence clusters rather than proteins, it is 28.4%, since 274 of TEST's 966 clusters contain an unseen-EC protein (both post hoc).

The remaining 876 seen-EC proteins are typical of new Swiss-Prot entries rather than remote ones. Against CLEAN's training set, 22.6% of TEST proteins have less than 30% identity, 40.9% have at least 50%, and 7.4% have at least 90%. Within the benchmark, 345 TEST proteins (27%) are at least 50% identical to a CAL protein, and 59 have at least 90% identity. Within TEST, the 876 seen-EC proteins form 699 sequence clusters at 50% identity (post hoc). Training vintage matters as much as identity for any newer model. The released training set of EnzPlacer [31] contains 631 of the 1,260 TEST accessions (50.1%), 430 of the 839 CAL accessions and 316 of the 392 New-392 sequences, with entries created up to April 2025. TEST is a valid hold-out for CLEAN, but half of it would be training data for EnzPlacer.

We also checked whether CLEAN's false positives correspond to functions that curators have since added. Of the 224 false-positive calls CLEAN makes on New-392, one (0.45%) is now annotated on the protein in UniProt release 2026_03; for BLASTp, the corresponding number is 0 of 203 (post hoc). Later curation has therefore confirmed almost none of them. However, because an absent annotation is not proof of absent activity, this provides weak evidence that they are wrong, not proof.

Test sets differ greatly in what they can detect. For the comparison of BLASTp with CLEAN's centroid rule, the minimum detectable effect at 80% power (MDE_{80}) is 0.032 in micro-F1 on TEST, 0.052 on New-392 and 0.096 on Price-149. For balanced3 against the centroid, it is 0.023, 0.038 and 0.039. The median MDE_{80} across EC-Bench's 120 method pairs is 0.063 (all post hoc). Price-149 therefore has limited power to detect modest differences between methods.

4.3 Inference rules and the homology baseline

Within CLEAN's frozen embedding space, replacing the centroid rule with a member-level rule improved exact-EC prediction (Table 4, Fig. 3). The pre-registered primary endpoint (H3b) was met. On the 876 seen-EC TEST proteins, balanced3 increased union-label macro-F1 from 0.432 to 0.475 (+0.043 [+0.030, +0.060], $p < 0.001$) and micro-F1 from 0.641 to 0.689 (+0.047 [+0.032, +0.063]). The gain held under the EC-cluster bootstrap (+0.043), with at most 5 proteins per EC number (+0.041), on the CAL-disjoint subset (+0.045 [+0.030, +0.065], $n = 649$), on proteins with protein-level evidence (+0.035) and on the split70 backbone (+0.036 [+0.022, +0.052]). In

post hoc analyses, the gain also held for per-protein F1 (+0.046 [+0.032, +0.059]) and when sequence clusters rather than proteins were resampled (macro-F1 +0.043 [+0.030, +0.061], micro-F1 +0.047 [+0.032, +0.064]). However, it did not replicate on CLEAN's original test sets: +0.025 [−0.000, +0.046] on New-392, which had been used during development, and +0.008 [−0.016, +0.027] on Price-149.

The class-size correction is not what drives the gain. The tuned exponent was close to zero ($\lambda = 0.02$), and the uncorrected rules performed at least as well as balanced3 in the descriptive comparisons: exemplar increased union-label macro-F1 by +0.079, kmean3 by +0.063 and balanced1 by +0.055, while multiproto did not differ significantly from the centroid (+0.007 [−0.005, +0.016]). With kmean3 at 0.494, the correction reduced union-label macro-F1 to 0.475. The pre-registered test therefore confirms that the size-balanced 3-nearest-neighbour (3-NN) rule outperforms the centroid rule; the mechanism supported by the data is inference from nearby training members rather than from a class mean. The gain is not an artefact of predicting more labels: the rules predicted 1.17–1.23 EC numbers per seen-EC protein (post hoc, Table 4).

Table 4 Exact-EC (level 4) performance on the 876 seen-EC TEST proteins. Differences are against CLEAN's centroid rule v paired bootstrap intervals ($B = 2{,}000$). Only the balanced3 row is the pre-registered primary comparison; all other rows are de Micro-precision (P), micro-recall (R) and EC numbers predicted per protein were added post hoc.

Rule	P	R	Micro-F1	Union-label macro-F1	ECs/protein	Δ macro-F1	Δ micr
centroid (CLEAN)	0.632	0.652	0.641	0.432	1.17	–	–
multiproto	0.628	0.659	0.643	0.439	1.19	+0.007 [−0.005, +0.016]	+0.002 [−0.0
balanced3	0.671	0.707	0.689	0.475	1.19	+0.043 [+0.030, +0.060]	+0.047 [+0.0
balanced1	0.665	0.723	0.693	0.486	1.23	+0.055 [+0.038, +0.071]	+0.052 [+0.0
kmean3	0.687	0.717	0.702	0.494	1.18	+0.063 [+0.048, +0.079]	+0.060 [+0.0
exemplar	0.676	0.732	0.703	0.510	1.22	+0.079 [+0.059, +0.093]	+0.062 [+0.0
BLASTp top hit	0.745	0.724	0.735	0.542	1.10	+0.111 [+0.087, +0.134]	+0.093 [+0.0

The gain is concentrated in common EC numbers. Exemplar minus centroid in micro-F1 was +0.082 [+0.060, +0.106] for common EC numbers (Holm-adjusted $p < 0.005$), and +0.025 [−0.011, +0.063], +0.048 [−0.020, +0.109] and −0.020 [−0.068, +0.026] for rare, very rare and singleton EC numbers (Holm-adjusted $p = 0.48$ each). Hypothesis H3a was therefore partly met: member-level inference helps common EC numbers but does not significantly harm rare ones on TEST. Balanced3 behaved similarly (+0.070 [+0.051, +0.089] for common EC numbers; +0.002, +0.016 and −0.022, not significant, in the rarer tiers). On Price-149, the full trade-off appeared, with significant losses in the three rarer tiers, but that pattern had already been observed during development and is not confirmatory. For CLEAN itself, micro-F1 declined with rarity, from 0.70 for common to 0.38 for singleton EC numbers, and with sequence identity, from 0.89 at ≥ 90% identity to 0.41 below 30% (Appendix Tables 11 and 12, which also report these results for New-392 and Price-149).

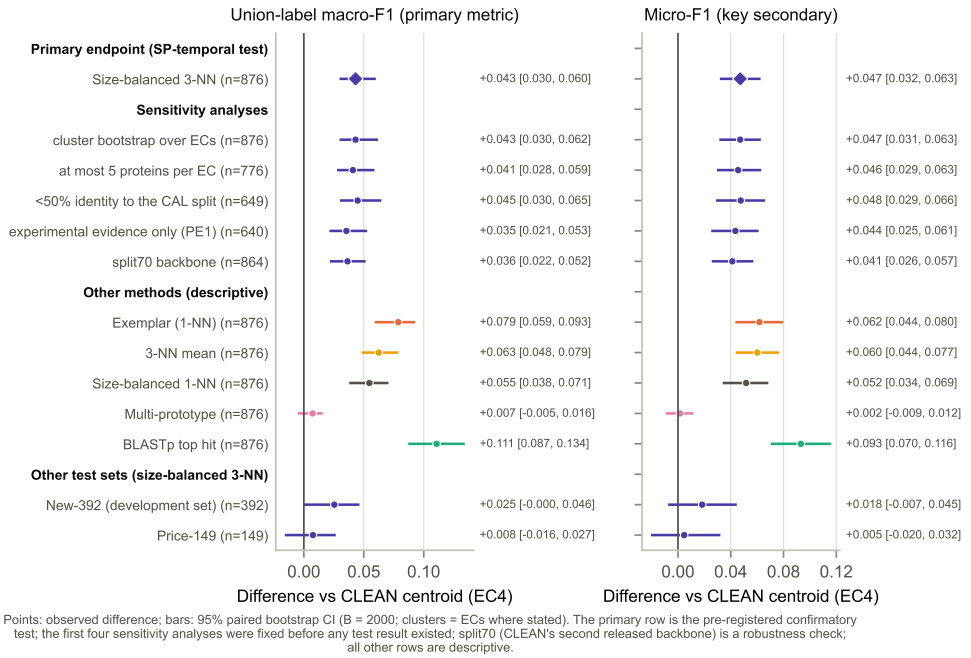


Fig. 3 The pre-registered primary endpoint (balanced3 minus CLEAN's centroid rule on the seen-EC TEST proteins), its pre-registered sensitivity analyses, the descriptive comparisons with the other rules and with BLASTp, and the same comparison on New-392 and Price-149. Points are observed differences; bars are 95% paired bootstrap intervals

In the descriptive comparisons, a standard homology search outperformed every rule in CLEAN's embedding space (Table 4). Transferring the EC numbers of the best BLASTp hit in CLEAN's own training set gave a seen-EC micro-F1 of 0.735 on TEST, ahead of CLEAN's centroid rule by +0.093 [+0.070, +0.116], balanced3 by +0.046 [+0.027, +0.065] and exemplar by +0.031 [+0.013, +0.051]. The advantage came from both precision (0.745 against 0.632) and recall (0.724 against 0.652), with fewer labels predicted per protein (1.10 against 1.17; post hoc). It also held when sequence clusters were resampled (+0.093 [+0.067, +0.117]; post hoc). On New-392, BLASTp was also ahead (+0.086 [+0.051, +0.123]; micro-F1 0.584 against 0.498). On Price-149, it performed worse, but the difference was not significant (−0.051 [−0.116, +0.015]); the interval is compatible with a substantial BLASTp deficit as well as a small advantage. Price-149's approximate MDE_{80} of 0.096 indicates limited power to detect modest differences. By sequence identity, BLASTp's advantage over the centroid rule on TEST was +0.113, +0.109, +0.079 and +0.057 in the four bins from 30% upward, each significant after Holm adjustment across the five bins (post hoc, $p_{\mathrm{Holm}} \leq 0.002$). Below 30% identity, no significant difference was detected (+0.048 [−0.016, +0.113], $n = 166$, $p_{\mathrm{Holm}} = 0.13$); however, the interval is mostly positive, and equivalence was not tested. Compared with exemplar, BLASTp was significantly ahead only at 30–50%

identity (+0.056 [+0.019, +0.091], uncorrected), and the two made identical predictions above 90% identity. On TEST, we found no identity range in which CLEAN's embedding-based rules outperformed homology transfer. The advantage held in every pre-registered sensitivity subset (+0.09 to +0.11). With the approximate aligner used in the original analysis, the same top-hit baseline had scored 0.600 on TEST and 0.505 on New-392 (Section 3.3); the change in aligner, rather than the transfer rule, accounts for this difference.

An additional analysis (H3c) examined whether member-level rules are more helpful when the same EC number is associated with unrelated proteins. In CLEAN's embedding space, 183 of 2,493 EC numbers (7.3%) have training members belonging to at least two non-homologous Pfam-clan groups. Their members lie farther from their own centroid than expected for their class size (AUROC 0.82, $p < 10^{-4}$), but are not closer to the centroids of other EC numbers ($p = 0.36$). The H3c comparison included 874 of the 876 seen-EC TEST proteins; two proteins were excluded because none of their true EC numbers matched an entry in the E3 EC-geometry table. The gain in per-protein F1 from exemplar on TEST proteins carrying these EC numbers (0.073, $n = 87$) was comparable to that observed for the remaining eligible proteins (0.062, $n = 787$; $p = 0.36$). Thus, H3c was not met (Appendix Table 13; New-392 gave the same result).

4.4 BLASTp's advantage holds on experimentally supported labels

Swiss-Prot curators often assign EC numbers based on similarity to characterised proteins, which could favour a method that transfers labels using sequence similarity. The pre-registered E7 analysis examined this possibility (Fig. 4). The analysis was registered after the descriptive comparison above was known, so it tested whether BLASTp's advantage still held on experimentally supported labels rather than predicting the existence of that advantage. Gate G7 was satisfied: of the 876 seen-EC TEST proteins, 477 had only experimentally supported labels (EXP), 218 had a label assigned by similarity (SIM), 61 had a rule-based label (RULE) and 120 belonged to OTHER (excluded from both arms; all 120 carried a label supported only by curator inference, ECO:0000305; post hoc). All 2,959 queried accessions were found in UniProt release 2026_03. On experimentally supported labels, BLASTp outperformed CLEAN by +0.082 [+0.052, +0.114] in micro-F1 (P7a, Holm-adjusted $p < 0.0005$, $\mathrm{MDE}_{80} = 0.044$). The interval remained above zero under the pre-registered EC-cluster bootstrap (+0.082 [+0.050, +0.113]) and when the 419 sequence clusters of EXP proteins were resampled (+0.082 [+0.048, +0.115]; post hoc). BLASTp's advantage on propagated labels was not detectably larger: the difference-in-differences (DiD) was +0.012 [−0.037, +0.059] (P7b, $p = 0.61$, $\mathrm{MDE}_{80} = 0.068$). According to the interpretation defined before the analysis (Appendix Table 7), BLASTp's advantage holds on experimentally supported labels, and there is no detectable provenance effect. P7b's confidence interval remains compatible with both a modest reduction and a modest increase in BLASTp's advantage on propagated labels. The non-significant DiD does not establish equivalence between the evidence classes; the positive result for experimentally supported labels is based on P7a.

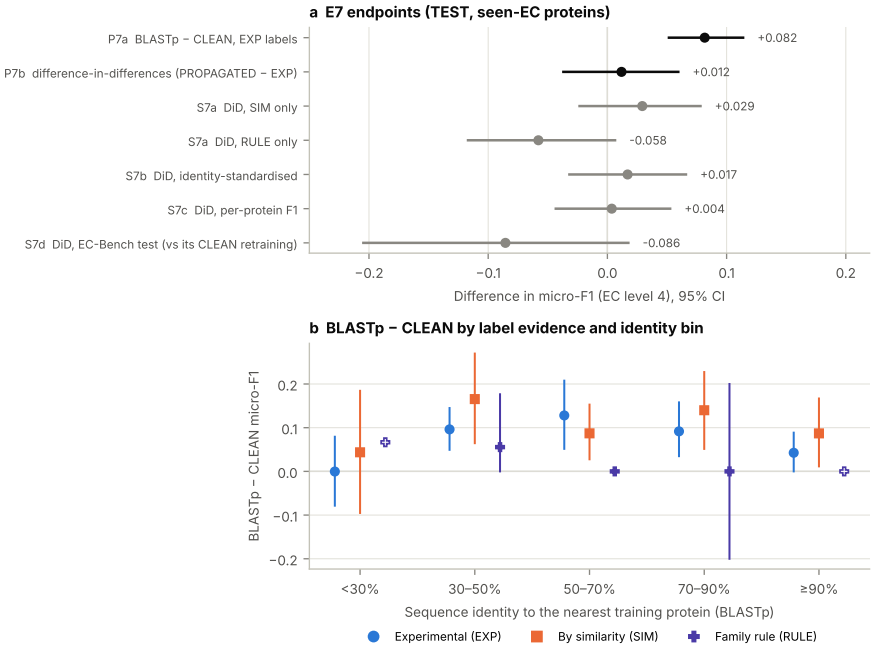


Fig. 4 Label provenance (E7). (a) The two pre-registered primary endpoints (black) and the secondary endpoints (grey) for BLASTp minus CLEAN, with 95% paired bootstrap intervals. P7a is the micro-F1 difference on EXP labels; P7b compares PROPAGATED (SIM and RULE) with EXP. Secondary contrasts use the evidence groups, metrics and adjustments indicated in the plot. (b) BLASTp minus CLEAN by evidence class and identity bin on the seen-EC TEST proteins; open markers represent cells with fewer than 10 proteins and are shown without an interval

None of the secondary endpoints changed this interpretation. The DiD was +0.029 [−0.023, +0.078] for SIM labels and −0.058 [−0.117, +0.006] for RULE labels ($q = 0.134$); after identity standardisation, it was +0.017 [−0.032, +0.066], with $\Delta_{\mathrm{EXP}} = +0.078$ [+0.050, +0.108]; and using per-protein F1, it was +0.004 [−0.043, +0.053]. Both sensitivity analyses agreed (P7b +0.012 with the EC-cluster bootstrap and +0.019 with at most 5 proteins per EC number). On EC-Bench's test set, compared with EC-Bench's retrained CLEAN model, BLASTp outperformed CLEAN on experimentally supported labels (+0.065 [+0.001, +0.136], $n = 98$), while the DiD was −0.086 [−0.205, +0.018] ($n = 46$ propagated), again not significant. In the New-392 pilot, the DiD was +0.060 [−0.016, +0.131]. Below 30% identity, no significant BLASTp advantage was detected on experimentally supported labels (−0.0003 [−0.078, +0.079], $n = 103$), consistent with the result across all labels.

The source-link analysis helps explain the provenance comparison. Of the 218 SIM proteins, only 45 (0.21 [0.16, 0.27]) had a BLASTp top hit that was a source protein cited by the label, although a cited source protein was among the retained BLASTp hits for 128 (0.59) and was present in CLEAN's training set for 164 (0.75). BLASTp outperformed CLEAN both when its top hit was a cited source protein (micro-F1 0.942 against 0.796, $n = 45$) and when it was not (0.853 against 0.751, $n = 173$).

4.5 Confidence, abstention and the certification budget

CLEAN's released confidence scores overestimated accuracy for rare predictions (Table 5). By tier of the predicted EC number on TEST, mean confidence was close to overall accuracy for common predictions (−0.01 [−0.05, +0.02] difference between mean confidence and accuracy), whereas for rare, very rare and singleton predictions confidence exceeded accuracy by +0.16 [+0.10, +0.22], +0.30 [+0.22, +0.38] and +0.28 [+0.19, +0.35] (intervals post hoc). The expected calibration error rose from 0.19 [0.16, 0.22] for common EC numbers to 0.22 [0.17, 0.28], 0.32 [0.25, 0.40] and 0.32 [0.26, 0.41] (hypothesis H4a satisfied). Expected calibration error is biased upwards in small samples, so we read the confidence-accuracy difference as the more reliable summary; both indicate poorer reliability in the rarer tiers. Isotonic recalibration fitted on CAL independently for each tier reduced the ECE on TEST to 0.05, 0.05, 0.08 and 0.10.

Recalibration reduced calibration error but could not overcome the lack of calibration data. The global SGR threshold chosen at $\alpha = 0.2$ had a risk of 0.12–0.14 on TEST for all three rules. While a global bound on the risk does not imply a similar bound for individual tiers, within the singleton tier the point estimate of the realised risk exceeded 0.2 for every rule: 4 of 14 errors for centroid (0.29 [0.08, 0.58]), 5 of 15 for exemplar (0.33 [0.12, 0.62]) and 11 of 24 for balanced3 (0.46 [0.26, 0.67]). Only the interval of balanced3 nominally excludes 0.2; these within-tier intervals were not adjusted for individual rules (Appendix Table 14 shows every tier and rule). With separate SGR thresholds for each predicted tier, all three rules rejected every prediction in the rarer tiers at $\alpha = 0.2$; the centroid rule also rejected all common-tier predictions. Their rare-tier realised risk is thus not estimable: hypothesis H4b did not establish within-tier accuracy. Instead we report the minimal risk at which the SGR accepted predictions on CAL. For the centroid rule this was 0.25 for common predictions and 0.65, 0.60 and 0.60 for rare, very rare and singleton ones; exemplar reached 0.15 for common predictions. An empirical level-3 fallback increased coverage. At $\alpha = 0.2$, the per-tier centroid rule that falls back to the first three digits of the EC number if the full number is not certified accepted 1,059 of 1,260 TEST proteins (84%) with the realised level-3 risk of 0.145.

Abstention additionally provided protection against inputs the model cannot process. With a global threshold set at $\alpha = 0.2$, the centroid rule of CLEAN abstained from 98% of unseen-EC proteins (98% of the pure and 98% of the mixed groups) and on all 2,000 non-enzymes but accepted only 8% of TEST. Exemplar accepted 37% of TEST and still abstained from 91% of unseen-EC proteins and 99.7% of non-enzymes. The raw scores distinguished seen-EC enzymes from non-enzymes with the AUROC of 0.90 (0.89–0.91 across rules) but distinguished seen-EC from unseen-EC enzymes with a lesser AUROC of 0.72–0.77.

Model-based budget estimates refer to the average risk of accepted calls within each tier, not to the risk of individual EC numbers (Fig. 5, Table 5). With $\alpha = 0.2$, the common tier estimate was 635 calibration proteins compared with 525 available; the estimate of the deficit was 110 [−250, 2,055]. For the very rare and singleton tiers, the estimate of selective risk was at least 0.2 at every available coverage value, yielding infinite point estimates; only 46% and 21% of the bootstrap resamples yielded finite estimates. The lower endpoint of the bootstrap intervals of the required total was 630

and 605 proteins, compared with 102 and 90 available. The corresponding intervals of the deficit started at 528 and 515, satisfying the registered H6 decision rule in this budget model. The rare tier point estimate was unstable (49,340), and the lower endpoint of the bootstrap interval was 550 compared with 122 available. These lower endpoints are 4.5–6.7 times the available sample sizes and are not universal minimum requirements.

Estimates depend on the target risk. With $\alpha = 0.3$, the common tier budget model was within the available sample size, while the lower endpoint of the deficit intervals was 178 and 198 proteins for the rare and very rare tiers and 135 for the singleton tier, whose point estimate remained infinity. With $\alpha = 0.1$, the lower endpoint of the deficit was 455 even for the common tier (post-hoc summary of secondary decisions). SGR simulation on the common tier produced certification probability 0.56 at the estimated budget and 0.92 at twice that budget, illustrating the difference between the fixed-operating-point calculation and adaptive certification.

Conversion of the lower endpoints of the deficit to the accrual rate of 2022–2026 gave 6.3, 9.3 and 10.3 years for the rare, very rare and singleton tiers. Annual counts varied substantially: 357 qualifying enzymes in 2023, 614 in 2024 and 221 in 2025; 2022 and 2026 were partial years. Conversion based on the highest and lowest full-year rates provides approximately 5–22 years. These conditional scenarios provide an indication of the scale of years rather than months, not a forecast and not a guaranteed waiting time.

Table 5 Reliability and model-based calibration budgets for CLEAN's centroid rule by tier of predicted EC number. Panel A reports TEST top-1 predictions and the number of errors among calls accepted by SGR with global $\alpha = 0.2$. Panel B reports CAL quantities. Budget intervals pertain to the fixed-operating-point model, not proven lower bounds for adaptive SGR. Years are conditional scenario-based conversions of the lower endpoint of the deficit interval at the average accrual rate.

A. Calibration and observed risk on TEST

Tier	n	Accuracy	Confidence	ECE	ECE (isotonic)	Errors/accepted
Common	785	0.61	0.59	0.19	0.05	7/63
Rare	205	0.36	0.51	0.22	0.05	1/17
Very rare	129	0.24	0.54	0.32	0.08	1/11
Singleton	141	0.16	0.44	0.32	0.10	4/14

B. CAL thresholds and model-based budgets at target risk 0.2

Tier	Min. target	Available	Budget [95% interval]	Deficit interval	Years
Common	0.25	525	635 [275, 2,580]	[−250, 2,055]	–
Rare	0.65	122	49,340 [550, ∞)	[428, ∞)	6.3
Very rare	0.60	102	∞ [630, ∞)	[528, ∞)	9.3
Singleton	0.60	90	∞ [605, ∞)	[515, ∞)	10.3

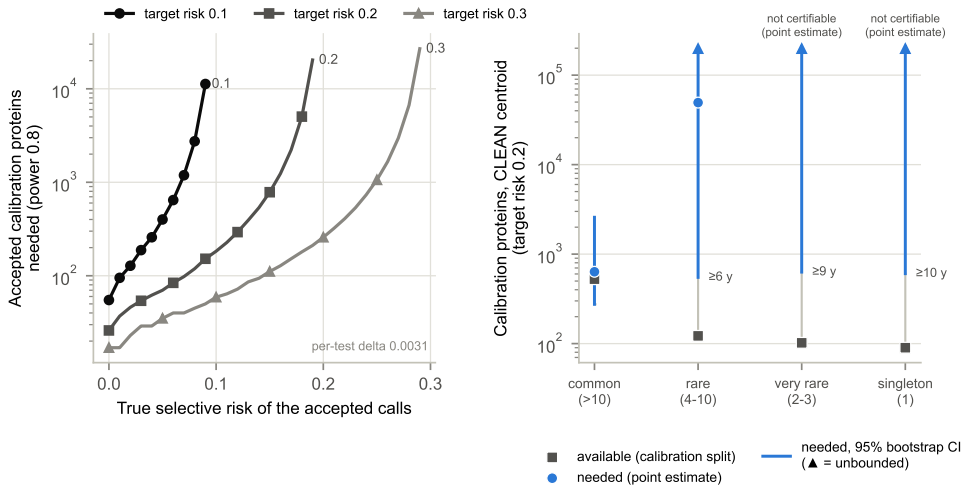


Fig. 5 Model-based calibration budgets for tier-level average risk. Left: accepted calibration counts under the fixed-operating-point binomial-power model as a function of assumed selective risk. Right: estimated total calibration budgets and available counts per predicted tier for CLEAN's centroid rule at target risk 0.2. Arrows denote unbounded model intervals. Year labels are conditional conversions of the lower bootstrap deficit endpoint at the 2022–2026 average rate; they are not guaranteed minimum waiting times

Table 6 Post-hoc analyses performed after the confirmatory analyses. None is pre-registered; all are based on the stored outputs of the original analyses, $B = 2{,}000$ and seed 20260927. "Cluster" resampling draws TEST sequence clusters (single linkage at $\geq 50\%$ identity over the shorter sequence) instead of proteins.

Question	Result	Claim it qualifies
Do intervals survive family-level resampling?	Yes: H3b +0.043 [+0.030, +0.061] and BLASTp +0.093 [+0.067, +0.117] over 699 clusters; P7a +0.082 [+0.048, +0.115] over 419	H3b, BLASTp, P7a
Are BLASTp's identity-bin differences significant after correction?	30–50% to $\geq 90\%$: $p_{\text{Holm}} \leq 0.002$; $< 30\%$: $p_{\text{Holm}} = 0.13$	BLASTp by identity
Is BLASTp's advantage precision or recall?	Both: P 0.745 vs 0.632, R 0.724 vs 0.652; 1.10 vs 1.17 ECs per protein	BLASTp
Does H3b hold in per-protein F1?	+0.046 [+0.032, +0.059]	H3b
Are invisible false positives near misses?	Correct at EC level 3: 60%, 67%, 63% (New-392, Price-149, TEST)	Blind spot
Is the ECE rise real?	Confidence - accuracy: −0.01, +0.16, +0.30, +0.28 with intervals excluding 0 in the rarer tiers	H4a
How robust is the unseen share?	TEST: 27.9% without EC 1.14.99.56; 28.4% of sequence clusters (274 of 966)	Unseen-EC share
Is accrual stable?	357 (2023), 614 (2024), 221 (2025) qualifying enzymes per year	Years of accrual
Have false positives since been annotated?	New-392: 1 of 224 (CLEAN), 0 of 203 (BLASTp)	Blind spot
What is the excluded OTHER class?	120 proteins, all with a curator-inference label (ECO:0000305)	E7
How precise are the test-set comparisons?	MDE_{80}: 0.032 (TEST), 0.052 (New-392), 0.096 (Price-149)	Price-149 nulls

5 Discussion

The audit demonstrates the impact of evaluation choices on the interpretation of enzyme function prediction. Support-weighted scores hide the precision cost of EC numbers unseen in the test set labels. The mathematics is simple; the practice is that this convention hides 67–96% of CLEAN's false positives across the evaluated datasets and is shared by several later evaluation pipelines. Its effect on ranking was modest: the overall EC-Bench ranking was mostly preserved, and one pair reversed significantly, although six pairs changed order. Evaluation cards should thus report micro-F1, union-label macro-F1, per-protein F1, false positive count and omitted share, in addition to the results by rarity and identity.

Temporal benchmarking also requires coordination with the training snapshot of the model. In the cohort, 30.5% of the proteins have an EC number outside the vocabulary of CLEAN; the TEST share is 27.9% without EC 1.14.99.56. The proportion of such proteins highlights the limitation hidden in the evaluation of known labels. It reflects the curation of Swiss-Prot and should not be considered as the frequency of novel functions in all sequenced genomes.

Against the same reference, BLASTp outperformed the released version of CLEAN on seen-EC proteins, which is consistent with independent benchmarks [21, 29]. The advantage also held on experimentally supported labels in E7 although the subgroup analysis was registered after the overall result became known. The nearest-member inference recovered part of the performance gap within CLEAN's frozen embedding space, implicating both inference and representation. Below 30% identity the interval was too wide to establish either an advantage or equivalence. These findings apply to the released CLEAN, not to protein language models in general.

Deployment requires separate improvement of calibration and risk control. Tier-specific recalibration reduced measured calibration error, but rare-tier abstention often eliminated the coverage. The empirical EC3 fallback suggests that coarser annotations can preserve usefulness when exact-EC calls are uncertain; the adaptive calibration should be validated before a joint guarantee can be claimed. Model-based budget estimates indicate significant data requirements under the evaluated assumptions. Future research should validate hierarchical selection with independent calibration and audit another model with a matching temporal benchmark.

6 Limitations

We evaluated one released model without retraining. Curation of Swiss-Prot, related proteins across CAL and TEST, a prominent unseen-EC family and unmeasured sequence exposure during pre-training restrict generalization. Cluster and CAL-disjoint checks supported the key comparisons, but the approximate aligner may overlook relationships. The evidence codes reflect the current release, not the annotation-time provenance, and experimental labels do not eliminate family selection bias. E7 reused the evaluated cohort; BLASTp replaced the approximate aligner after the results became known, and the comparisons outside E7 are descriptive. We did not reproduce the BLASTp setup of the original CLEAN paper. The precision of

estimates from Price-149 and the low-identity subsets is limited; the EC-Bench reversal concerns one pair. Finally, budget estimates depend on the risk estimation, finite coverage grid, numerical search limits and future data similar to CAL. They are planning estimates, not proven lower bounds. CAL reuse for balanced3 and the fallback is assessed empirically, without an additional finite-sample guarantee.

7 Conclusions

Reliable evaluation of EC predictors requires counting every false positive, matching benchmarks to training history and comparing against a suitable homology baseline. In CLEAN, support-weighted precision ignored most false positives, newly curated Swiss-Prot enzymes frequently carried unseen EC numbers, and exemplar inference improved on the centroid rule. BLASTp was stronger on the evaluated seen-EC cohort, including its experimental-label subset. Rare-tier confidence required recalibration, while abstention and model-based budget analyses exposed limited calibration support. Reporting these distinctions alongside aggregate scores would make claims about enzyme annotation more transparent and reproducible.

Acknowledgements. The CLEAN Software, used by Prativa Khanal and Prateek Paudel, was developed by the Department of Chemical and Biomolecular Engineering at the University of Illinois Urbana-Champaign.

A Proof of Proposition 1

Let $\mathcal{C}_T$ be the set of classes with at least one true instance, and let $c^* \notin \mathcal{C}_T$ be a predicted class with no true instance, so $n_{c^*} = \mathrm{TP}_{c^*} + \mathrm{FN}_{c^*} = 0$. In support-weighted averaging, the weight of class c is $n_c / \sum_{c'} n_{c'}$, so c^* has weight 0 and its precision, recall and F1 do not enter the average. Every prediction of c^* is a false positive of c^* alone: it changes FP_{c^*} and leaves TP_c, FP_c and FN_c of every $c \in \mathcal{C}_T$ unchanged. Hence the per-class scores of all classes with non-zero weight, and the weighted average, are unchanged. True-label macro scores average over $\mathcal{C}_T$ only, so the same argument applies. For recall, with the sum taken over classes with positive support, $\sum_c \frac{n_c}{\sum_{c'} n_{c'}} \cdot \frac{\mathrm{TP}_c}{n_c} = \frac{\sum_c \mathrm{TP}_c}{\sum_c n_c}$, which is micro-recall. Adding out-of-label predictions does not change micro-recall. It adds false positives that micro-precision and micro-F1 count, but support-weighted precision and F1 ignore. This invariance does not imply that the difference between weighted F1 and micro-F1 is determined by their aggregated precisions alone. Projecting predictions onto $\mathcal{C}_T$ before scoring deletes every prediction of c^*, so scores computed after projection are invariant by construction. □

B E7 interpretation table

Table 7 E7 interpretation, fixed before the analysis. "Significant" refers to Holm-adjusted primaries.

Outcome	Interpretation
G7 fails	Descriptive only
P7a significant positive, P7b not	BLASTp's advantage holds on experimental labels; no detectable provenance effect
Both significant positive	The advantage holds on experimental labels and is larger on propagated labels; the provenance-matched estimate is Δ_{EXP}
P7a not significant, P7b significant positive	The advantage is larger on propagated labels; no significant advantage is detected on experimental labels
Neither significant	No significant difference on experimental labels and no detectable provenance effect
Any significant negative effect	Reported as such

C Pre-registration register and deviations

Table 8 lists every pre-registered hypothesis and endpoint with its outcome, and Table 9 lists every change made after a pre-registration was frozen. Hypotheses H1–H6 and their secondary endpoints were specified in the main analysis protocol before each corresponding analysis; P7a, P7b and S7a–S7d were specified in the E7 protocol after the descriptive BLASTp comparison on TEST was known and before the E7 analysis. The H3b metric was switched to union-label macro-F1 before any TEST result existed. This and other amendments, including the E5 secondary hypothesis and test direction, are documented in Table 9. Both protocols and all amendments are available with the analysis materials (see Data and code availability).

Table 8: Register of pre-registered hypotheses and endpoints. "Met" means the pre-registered decision rule was satisfied. "Main": main protocol, specified before the corresponding analysis; "E7": E7 protocol, specified before the E7 analysis. E7 was registered after the descriptive BLASTp comparison on TEST was known (Table 9).

ID	Hypothesis or endpoint	Family and correction	Registered	Outcome
H1	Support-weighted precision and F1 ignore most false positives and rise under over-prediction	none (descriptive, per test set)	Main	First half met on all three test sets; second half met only on Price-149
S7	Invisible share of CLEAN's false positives on TEST below New-392's 0.67	none	Main	Not met (0.74)

Continued on next page

ID	Hypothesis or endpoint	Family and correction	Registered	Outcome
H2	Exact-EC performance falls with rarity tier and identity; a large share of new enzymes carry unseen EC numbers	none (descriptive)	Main	Met (micro-F1 0.70 to 0.38 by tier, 0.89 to 0.41 by identity; 30.5% unseen)
H3a	Exemplar helps common and hurts rare tiers	Holm over 4 tiers	Main	Half met: common +0.082 ($p_{\text{Holm}} < 0.005$); no significant harm in rarer tiers on TEST
H3b (primary)	balanced3 beats centroid in union-label macro-F1, TEST seen-EC (metric amended before any TEST result)	single primary, two-sided $\alpha = 0.05$	Main	Met: +0.043 [+0.030, +0.060]
H3b key secondary	Same contrast in micro-F1	none	Main	Met: +0.047 [+0.032, +0.063]
H3c	Exemplar gain larger for EC numbers with non-homologous members	none	Main	Not met ($p = 0.36$); dispersion part met
H4a	Confidence miscalibration (ECE) rises from common to singleton predicted tiers	none	Main	Met: 0.19, 0.22, 0.32, 0.32
H4b	Per-tier SGR keeps per-tier risk $\leq \alpha$ on TEST; a global threshold does not	none	Main	Rare-tier risk not estimable at zero coverage; global singleton interval excludes target for balanced3 (nominal)
H4c	Smallest certifiable risk differs by tier	none	Main	Met: 0.25 (common) vs 0.60–0.65
H5 (primary)	≥ 1 of 120 EC-Bench pairs reversed between support-weighted F1 and micro-F1	BH $q < 0.05$ over 120 pairs, sign agreement	Main	Met by 1 pair (DeepECtransformer-learnt vs PRIAM, $q = 0.026$)
S1	H5 reversal count using union-label macro-F1 and per-protein F1	BH over 120 pairs for each metric	Main	4 significant reversals for union-label macro-F1; 2 for per-protein F1
S2	Kendall's τ_b between reported and alternative rankings, with bootstrap intervals	none	Main	EC level 4: $\tau_b = 0.90$ (micro-F1), 0.57 (union-label macro-F1), 0.90 (per-protein F1)
S3	Share of false positives outside the test label set, by method	none (descriptive)	Main	Test set: median 0.65, range 0.25–0.87; Price-149: median 0.96
S4 (key secondary)	Spearman ρ(inflation, ECs per protein) > 0	one-sided	Main (amended after pilot)	Met: $\rho = 0.75$, $p = 0.0004$
S5	S1–S4 in the 30% regime, EC levels 1–3, Price-149	none	Main	Reported in Results

Continued on next page

ID	Hypothesis or endpoint	Family and correction	Registered	Outcome
S6	Complete-EC-only variant of H5	BH over 120 pairs	Main	1 significant reversal of a different pair (DeepECtransformer-regular vs PRIAM)
H6 (primary)	At $\alpha = 0.2$ the budget exceeds available CAL proteins in the very-rare and singleton predicted tiers	bootstrap 95% CI of the deficit	Main	Met within the budget model: deficit endpoint 528 for very-rare, 515 for singleton
H6 secondaries	$\alpha \in \{0.1, 0.3\}$, other rules, years of accrual, SGR simulation	none	Main	Reported in Results
P7a (primary)	$\Delta_{\text{EXP}} \neq 0$ (BLASTp minus CLEAN, micro-F1, EXP labels)	Holm over P7a, P7b; gate G7	E7	Met: +0.082 [+0.052, +0.114]
P7b (primary)	DiD $\neq 0$ (PROPAGATED minus EXP)	Holm over P7a, P7b; gate G7	E7	Not significant: +0.012 [−0.037, +0.059], $p = 0.61$
S7a	DiD for SIM and for RULE separately	Holm over 2	E7	Neither significant
S7b	Identity-standardised P7a and P7b	none	E7	Δ_{EXP} +0.078; DiD +0.017, not significant
S7c	P7a and P7b in per-protein F1	none	E7	DiD +0.004, not significant
S7d	DiD on EC-Bench test vs EC-Bench's CLEAN	none	E7	−0.086 [−0.205, +0.018], not significant

Table 9: Deviations from the analysis protocols, in order of occurrence. Amendments made before the confirmatory analysis are distinguished from later changes. The H3b primary-metric amendment preceded any TEST result; the E5 secondary-hypothesis and test-direction amendment followed the pilot and preceded confirmatory scoring.

Stage	Change	Reason	Numbers affected
Before any TEST result	H3b metric and λ criterion switched from true-label to union-label macro-F1	E1 showed the true-label score rises under over-prediction	Primary endpoint definition; no TEST result existed
After the E5 pilot, before confirmatory scoring	E5 key secondary set to S4 with a one-sided direction	Price-149 pilot: no significant reversal, $\rho = 0.94$	None; primary unchanged
After the E6 pilot, before the confirmatory analysis	E6 Jeffreys shrinkage and the $\geq$ 10-accepted-protein rule	Pilot on half of New-392	Budget estimates

Continued on next page

Stage	Change	Reason	Numbers affected
After the confirmatory analysis	Approximate aligner replaced by BLASTp for the homology baseline, identity bins and CAL-disjoint subset	The protocol recommended a standard aligner for the final analysis; the approximate aligner missed hits	Top-hit baseline 0.600 to 0.735 (TEST) and 0.505 to 0.584 (New-392); identity strata; CAL-disjoint primary +0.045, $n = 649$
After the confirmatory analysis	EC-cluster bootstrap: per-protein F1 divided by resample weight instead of N	Implementation error	Per-protein F1 rows of that sensitivity analysis only; none is reported
After the confirmatory analysis	Reporting: exact CIs on realised risks; infinite E6 bounds; years at the deficit's lower bound	Unstable point estimates near the certifiability boundary	Presentation only
After the BLASTp comparison, before the E7 analysis	E7 protocol specified	Test whether the BLASTp advantage, already known, survives on experimental labels	E7 is confirmatory for P7a and P7b only, conditional on the known parent result
After all confirmatory analyses	Post hoc analyses (Table 6)	Robustness and interpretation	All labelled post hoc

D Evaluation code inspected

Table 10 Public scoring code inspected for the support-weighted blind spot (Proposition 1). "Invariant" means the reported score does not change when a method predicts an EC number that no test protein carries. Commits are the upstream versions inspected; the CLEAN release used for all analyses has the same `evaluate.py` as commit `f2bf2a4`.

Code base, version and repository	Location	Scoring	Consequence
CLEAN, commit `f2bf2a4` (`evaluate.py` identical to that in `71f425b`) https://github.com/tttianhao/CLEAN	`app/src/CLEAN/evaluate.py`, `get_eval_metrics`, lines 301–323	Binarizer fitted on the test labels; scikit-learn scores with `average='weighted'`	Invariant
CLEAN-Contact, commit `0e423db` https://github.com/pnnl-predictive-phenomics/clean-contact	`src/CLEAN/evaluate.py`, lines 320–342	Copy of CLEAN's function	Invariant
ProtDETR, commit `aa49809` https://github.com/yangzhao1230/ProtDETR	`util/clean.py`, lines 311–331	Copy of CLEAN's function	Invariant
EnzymeHunter, commit `fd2aa18` https://github.com/cgxbio/EnzymeHunter	`src/evaluate.py`, lines 241–262	Copy of CLEAN's function	Invariant
EC-Bench, commit `5e98a2c` https://github.com/dsaeedeh/EC-Bench	`code/results.ipynb`, cell 60	Support-weighted F1 over the union of true and predicted labels	Invariant (zero support)

E Detailed results

This appendix gives the numerical results behind the stratified analyses summarised in Section 4: micro-F1 by rarity tier (Table 11) and by sequence identity (Table 12), the H3c comparison (Table 13), risk control by predicted tier (Table 14) and the re-scoring of EC-Bench (Table 15). The TEST rows and the EC-Bench re-scoring belong to the pre-registered analyses H2, H3a, H3c, H4b and H5, with identity bins computed with BLASTp (Table 9); rows for New-392 and Price-149 are given for comparison.

Table 11 Exact-EC micro-F1 by rarity tier of the true EC number, with 95% bootstrap intervals ($B = 2{,}000$). Tiers are those of Section 3.1 (split100 training count of the rarest true EC number). An asterisk marks a significant difference from CLEAN's centroid rule after Holm adjustment over the four seen-EC tiers ($p_{\mathrm{Holm}} < 0.05$); BLASTp was not tested against the centroid rule by tier. The unseen-EC row covers TEST proteins with an EC number absent from CLEAN's training set.

Test set	Tier	n	centroid (CLEAN)	balanced3	exemplar	BLASTp top hit
TEST	Common	634	0.701 [0.666, 0.734]	0.770 [0.740, 0.801]*	0.782 [0.752, 0.812]*	0.811 [0.781, 0.839]
	Rare	122	0.585 [0.503, 0.667]	0.587 [0.502, 0.671]	0.610 [0.525, 0.691]	0.596 [0.505, 0.678]
	Very rare	52	0.473 [0.356, 0.597]	0.489 [0.366, 0.620]	0.522 [0.389, 0.657]	0.571 [0.443, 0.705]
	Singleton	68	0.382 [0.280, 0.486]	0.360 [0.267, 0.461]	0.362 [0.257, 0.474]	0.476 [0.368, 0.588]
	Unseen EC	384	0.022 [0.009, 0.036]	0.026 [0.013, 0.041]	0.025 [0.013, 0.041]	0.025 [0.011, 0.041]
New-392	Common	217	0.621 [0.557, 0.681]	0.662 [0.596, 0.721]*	0.680 [0.619, 0.739]*	0.745 [0.690, 0.796]
	Rare	79	0.414 [0.316, 0.521]	0.420 [0.313, 0.535]	0.462 [0.363, 0.562]	0.523 [0.416, 0.633]
	Very rare	58	0.453 [0.341, 0.573]	0.424 [0.307, 0.546]	0.465 [0.355, 0.575]	0.489 [0.372, 0.611]
	Singleton	38	0.152 [0.060, 0.275]	0.151 [0.061, 0.270]	0.157 [0.068, 0.269]	0.177 [0.080, 0.298]
Price-149	Common	74	0.338 [0.244, 0.431]	0.380 [0.284, 0.473]	0.400 [0.309, 0.486]*	0.318 [0.215, 0.421]
	Rare	29	0.531 [0.364, 0.700]	0.507 [0.347, 0.677]	0.448 [0.295, 0.606]*	0.254 [0.119, 0.413]
	Very rare	18	0.696 [0.560, 0.837]	0.619 [0.450, 0.780]	0.565 [0.409, 0.720]*	0.737 [0.550, 0.900]
	Singleton	28	0.349 [0.194, 0.526]	0.355 [0.197, 0.533]	0.308 [0.159, 0.464]*	0.421 [0.246, 0.607]

Table 12 Exact-EC micro-F1 of seen-EC proteins by BLASTp identity to the nearest protein in CLEAN's training set, with 95% bootstrap intervals ($B = 2{,}000$). Intervals are omitted for bins with fewer than 10 proteins. Holm-adjusted tests of BLASTp against CLEAN's centroid rule on TEST are reported in Section 4.3 (post hoc).

Test set	Identity	n	centroid (CLEAN)		balanced3		exemplar		BLASTp top hit	
TEST	$< 30\%$	166	0.414	[0.343, 0.484]	0.457	[0.385, 0.530]	0.464	[0.392, 0.538]	0.462	[0.391, 0.536]
	30–50%	285	0.580	[0.529, 0.636]	0.609	[0.556, 0.664]	0.638	[0.587, 0.689]	0.693	[0.641, 0.743]
	50–70%	187	0.710	[0.647, 0.772]	0.761	[0.701, 0.819]	0.793	[0.737, 0.848]	0.819	[0.764, 0.869]
	70–90%	150	0.794	[0.729, 0.859]	0.866	[0.807, 0.918]	0.847	[0.782, 0.906]	0.873	[0.815, 0.925]
	$\geq 90\%$	88	0.885	[0.811, 0.956]	0.938	[0.886, 0.983]	0.942	[0.886, 0.989]	0.942	[0.886, 0.989]
New-392	$< 30\%$	59	0.277	[0.175, 0.388]	0.292	[0.187, 0.403]	0.288	[0.176, 0.405]	0.400	[0.277, 0.511]
	30–50%	118	0.365	[0.283, 0.447]	0.389	[0.307, 0.473]	0.394	[0.309, 0.478]	0.461	[0.370, 0.553]
	50–70%	101	0.672	[0.572, 0.775]	0.691	[0.589, 0.789]	0.696	[0.595, 0.802]	0.745	[0.649, 0.843]
	70–90%	71	0.521	[0.408, 0.632]	0.538	[0.428, 0.650]	0.573	[0.483, 0.663]	0.582	[0.480, 0.686]
	$\geq 90\%$	43	0.684	[0.557, 0.807]	0.700	[0.581, 0.810]	0.736	[0.617, 0.852]	0.736	[0.617, 0.852]
Price-149	$< 30\%$	26	0.211	[0.069, 0.370]	0.203	[0.067, 0.357]	0.214	[0.071, 0.370]	0.111	[0.000, 0.255]
	30–50%	75	0.471	[0.378, 0.570]	0.479	[0.380, 0.580]	0.444	[0.358, 0.531]	0.377	[0.273, 0.487]
	50–70%	21	0.468	[0.267, 0.652]	0.478	[0.286, 0.652]	0.458	[0.255, 0.640]	0.489	[0.280, 0.682]
	70–90%	25	0.473	[0.302, 0.630]	0.491	[0.327, 0.644]	0.467	[0.321, 0.621]	0.538	[0.360, 0.720]
	$\geq 90\%$	2	0.500		0.400		0.500		0.500	

Table 13 Hypothesis H3c: mean gain of exemplar over CLEAN's centroid rule in per-protein F1, for proteins whose EC number has training members in at least two non-homologous Pfam-clan groups and for other eligible seen-EC proteins. The TEST comparison includes 874 proteins; two of the 876 seen-EC TEST proteins had no true EC number matching an entry in the E3 EC-geometry table and were excluded. p: two-sided Mann–Whitney test.

	Non-homologous members		Other EC numbers		
Test set	Gain	n	Gain	n	p
TEST	0.073	87	0.062	787	0.36
New-392	0.035	25	0.041	367	0.58

Table 14 Risk control on TEST at target risk $\alpha = 0.2$ by predicted tier (hypothesis H4b). Each cell gives accepted/total proteins; errors among accepted calls (realised risk [exact 95% binomial interval]). "Global": one SGR threshold selected on CAL for all tiers; "Per tier": a separate SGR threshold selected on CAL within each predicted tier. "–": nothing accepted, so the realised risk is not estimable.

Rule	Threshold	All	Common	Rare	Very rare	Singleton
centroid	Global	105/1260; 13 (0.12 [0.07, 0.20])	63/785; 7 (0.11 [0.05, 0.22])	17/205; 1 (0.06 [0.00, 0.29])	11/129; 1 (0.09 [0.00, 0.41])	14/141; 4 (0.29 [0.08, 0.58])
	Per tier	0/1260; –	0/785; –	0/205; –	0/129; –	0/141; –
exemplar	Global	471/1260; 64 (0.14 [0.11, 0.17])	392/883; 49 (0.12 [0.09, 0.16])	41/176; 6 (0.15 [0.06, 0.29])	23/92; 4 (0.17 [0.05, 0.39])	15/109; 5 (0.33 [0.12, 0.62])
	Per tier	403/1260; 52 (0.13 [0.10, 0.17])	403/883; 52 (0.13 [0.10, 0.17])	0/176; –	0/92; –	0/109; –
balanced3	Global	438/1260; 63 (0.14 [0.11, 0.18])	362/829; 45 (0.12 [0.09, 0.16])	36/167; 5 (0.14 [0.05, 0.29])	16/96; 2 (0.12 [0.02, 0.38])	24/168; 11 (0.46 [0.26, 0.67])
	Per tier	374/1260; 50 (0.13 [0.10, 0.17])	374/829; 50 (0.13 [0.10, 0.17])	0/167; –	0/96; –	0/168; –

Table 15 Re-scoring of the per-protein predictions released by EC-Bench for 16 methods on its test set (100% clustering regime, EC level 4, $n = 468$ proteins), ordered by EC-Bench's reported F1. FP: false-positive EC calls; invisible: false positives on EC numbers absent from the test labels. †: the pair with a significant reversal between reported F1 and micro-F1 (H5, BH $q = 0.026$).

Method	Reported F1	Micro-F1	Union-label macro-F1	ECs/protein	Abstained	FP	Invisible FP
Stacked	0.639	0.653	0.316	1.07	0	170	73
Majority voting-learnt	0.515	0.516	0.201	1.09	0	247	133
Majority voting-regular	0.508	0.505	0.198	1.10	0	258	133
BLASTp	0.502	0.497	0.203	0.92	86	196	133
EnzBert-learnt	0.475	0.468	0.182	1.05	0	257	131
EnzBert-regular	0.468	0.477	0.205	1.00	0	235	109
ProteinBERT-learnt	0.414	0.366	0.104	1.13	0	338	258
ProteinBERT-regular	0.394	0.362	0.106	1.03	0	302	221
ECRECer	0.359	0.399	0.109	1.00	0	273	133
ECPred	0.273	0.292	0.093	0.83	81	256	64
CLEAN (SupCon)	0.251	0.224	0.110	1.22	0	450	393
DeepEC	0.127	0.184	0.122	0.26	353	64	49
DeepECtransformer-learnt†	0.115	0.118	0.051	1.09	0	452	393
PRIAM†	0.099	0.158	0.103	0.15	402	26	16
DeepECtransformer-regular	0.095	0.100	0.052	1.01	0	426	365
CatFam	0.074	0.124	0.106	0.12	419	20	16

References

[1] McDonald AG, Tipton KF (2023) Enzyme nomenclature and classification: the state of the art. FEBS J 290:2214–2231. https://doi.org/10.1111/febs.16274

[2] MacDougall A, Volynkin V, Saidi R, Poggioli D, Zellner H et al (2020) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics 36:4643–4648. https://doi.org/10.1093/bioinformatics/btaa485

[3] Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H (2023) Enzyme function prediction using contrastive learning. Science 379:1358–1363. https://doi.org/10.1126/science.adf2465

[4] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A 118:e2016239118. https://doi.org/10.1073/pnas.2016239118

[5] de Crécy-Lagard V, Dias R, Sexson N, Friedberg I, Yuan Y, Swairjo MA (2025) Limitations of current machine learning models in predicting enzymatic functions for uncharacterized proteins. G3 (Bethesda) 15:jkaf169. https://doi.org/10.1093/g3journal/jkaf169

[6] Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4:100804. https://doi.org/10.1016/j.patter.2023.100804

[7] Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, pp 1321–1330.

[8] Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608. https://doi.org/10.1016/S0022-2836(02)00016-5

[9] Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity?. J Mol Biol 333:863–882. https://doi.org/10.1016/j.jmb.2003.08.057

[10] Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605

[11] Rembeza E, Engqvist MKM (2021) Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLoS Comput Biol 17:e1009446. https://doi.org/10.1371/journal.pcbi.1009446

[12] Nadendla S, Jackson R, Munro J, Quaglia F, Mészáros B, Olley D, Hobbs ET, Goralski SM, Chibucos M, Mungall CJ, Tosatto SCE, Erill I, Giglio MG (2022) ECO: the Evidence and Conclusion Ontology, an update for 2022. Nucleic Acids Res 50:D1515–D1521. https://doi.org/10.1093/nar/gkab1025

[13] Rogers MF, Ben-Hur A (2009) The use of gene ontology evidence codes in preventing classifier assessment bias. Bioinformatics 25:1173–1177. https://doi.org/10.1093/bioinformatics/btp122

[14] Dalkiran A, Rifaioglu AS, Martin MJ et al (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics 19:334. https://doi.org/10.1186/s12859-018-2368-y

[15] Ryu JY, Kim HU, Lee SY (2019) Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A 116:13996–14001. https://doi.org/10.1073/pnas.1821905116

[16] Sanderson T, Bileschi ML, Belanger D, Colwell LJ (2023) ProteInfer, deep neural networks for protein functional inference. eLife 12:e80942. https://doi.org/10.7554/eLife.80942

[17] Kim GB, Kim JY, Lee JA, Norsigian CJ, Palsson BO, Lee SY (2023) Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nat Commun 14:7370. https://doi.org/10.1038/s41467-023-43216-z

[18] Yang Y, Jerger A, Feng S, Wang Z, Brasfield C, Cheung MS, Zucker J, Guan Q (2024) Improved enzyme functional annotation prediction using contrastive learning with structural inference. Commun Biol 7:1690. https://doi.org/10.1038/s42003-024-07359-z

[19] Cao G, Ouyang J, Xiong X, Liu C, Zhang Y, Yang S, Shi T, Wu J (2026) EnzymeHunter: achieving fine-grained enzyme function prediction with a hierarchically aware contrastive learning framework. Patterns :101567. https://doi.org/10.1016/j.patter.2026.101567

[20] Dumontet L, Han SR, Oh TJ, Kang M (2026) Trustworthy prediction of enzyme commission numbers using a hierarchical interpretable transformer. Nat Commun 17:1146. https://doi.org/10.1038/s41467-026-68727-3

[21] Capela J, Zimmermann-Kogadeeva M, van Dijk ADJ et al (2025) Comparative assessment of protein large language models for enzyme commission number prediction. BMC Bioinformatics 26:68. https://doi.org/10.1186/s12859-025-06081-9

[22] Sathyamoorthy R, Puri M (2026) Protein language models outperform BLAST for evolutionarily distant enzymes: a systematic benchmark of EC number prediction. bioRxiv preprint. https://doi.org/10.64898/2026.03.31.715487

[23] Sartori J et al (2026) A novel benchmark dataset for enzyme function prediction reveals the limitations of state-of-the-art models. bioRxiv preprint. https://doi.org/10.64898/2026.08.21.746242

[24] Radivojac P et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. https://doi.org/10.1038/nmeth.2340

[25] Phan A, Wang Y, Boadu F, Cheng J, Radivojac P, Friedberg I (2026) LAFA: a framework for reproducible longitudinal assessment of protein function annotation models. Bioinform Adv :vbag221. https://doi.org/10.1093/bioadv/vbag221

[26] Plyusnin I, Holm L, Törönen P (2019) Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences. PLoS Comput Biol 15:e1007419. https://doi.org/10.1371/journal.pcbi.1007419

[27] Jiang Y, Clark WT, Friedberg I, Radivojac P (2014) The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. Bioinformatics 30:i609–i616. https://doi.org/10.1093/bioinformatics/btu472

[28] Opitz J, Burst S (2019) Macro F1 and macro F1. arXiv preprint arXiv:1911.03347.

[29] Davoudi S, Henry CS, Miller CS, Banaei-Kashani F (2026) EC-Bench: a benchmark for enzyme commission number prediction. Bioinform Adv 6:vbag004. https://doi.org/10.1093/bioadv/vbag004

[30] Yang J, Mora A, Liu S, Wittmann BJ, Anandkumar A, Arnold FH, Yue Y (2024) CARE: a benchmark suite for the classification and retrieval of enzymes. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.

[31] Ma X, Joshi P, Friedberg I, Li Q (2026) How not to be seen: predicting unseen enzyme functions using contrastive learning. Bioinformatics 42:btag215. https://doi.org/10.1093/bioinformatics/btag215

[32] Ding K, Luo J, Luo Y (2024) Leveraging conformal prediction to annotate enzyme function space with limited false positives. PLoS Comput Biol 20:e1012135. https://doi.org/10.1371/journal.pcbi.1012135

[33] Boger RS, Chithrananda S, Angelopoulos AN, Yoon PH, Jordan MI, Doudna JA (2025) Functional protein mining with conformal guarantees. Nat Commun 16:85. https://doi.org/10.1038/s41467-024-55676-y

[34] The UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 53:D609–D617. https://doi.org/10.1093/nar/gkae1010

[35] Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305. https://doi.org/10.1093/nar/28.1.304

[36] Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26:1819–1837. https://doi.org/10.1109/TKDE.2013.39

[37] Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[38] Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31:6633–6639. https://doi.org/10.1093/nar/gkg847

[39] Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems 30 (NIPS 2017).

[40] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2020) Supervised contrastive learning. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[41] Sharma S, Xian Y, Yu N, Singh A (2023) Learning prototype classifiers for long-tailed recognition. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23). https://doi.org/10.24963/ijcai.2023/151

[42] Omelchenko MV, Galperin MY, Wolf YI, Koonin EV (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct 5:31. https://doi.org/10.1186/1745-6150-5-31

[43] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

[44] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

[45] Jiang Y, Oron TR, Clark WT et al (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 17:184. https://doi.org/10.1186/s13059-016-1037-6

[46] Zhou N et al (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 20:244. https://doi.org/10.1186/s13059-019-1835-8

[47] Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404–413. https://doi.org/10.1093/biomet/26.4.404

[48] Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 694–699. https://doi.org/10.1145/775047.775151

[49] Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp 4878–4887.

[50] Ding T, Angelopoulos AN, Bates S, Jordan MI, Tibshirani RJ (2023) Class-conditional conformal prediction with many classes. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[51] Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70. https://www.jstor.org/stable/4615733

[52] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[53] Walsh I, Fishman D, Garcia-Gasulla D et al (2021) DOME: recommendations for supervised machine learning validation in biology. Nat Methods 18:1122–1127. https://doi.org/10.1038/s41592-021-01205-4

Statements and Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests. The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate. Not applicable. This study used only publicly available protein sequence and annotation data and involved no human participants or animals.

Consent for publication. Not applicable.

Data availability. The derived benchmark accession lists and labels, per-protein predictions, BLASTp hit tables, evidence-class assignments, result tables and figure source data are deposited at Zenodo (https://doi.org/10.5281/zenodo.23934633). Public source data are available from UniProtKB, ENZYME and Pfam, with release identifiers given in the Materials and methods. The analysis code is available from

the corresponding author upon reasonable request. CLEAN weights, embeddings and split files must be obtained from their original source under the applicable licence, and ESM-1b weights from their official source; EC-Bench predictions were used from version 5e98a2c of the EC-Bench repository.

Materials availability. Not applicable.

Code availability. The analysis code is available from the corresponding author upon reasonable request.

Author contributions. Prativa Khanal conceived and designed the study, performed the analyses, prepared the figures and tables, and wrote the main manuscript text. Prateek Paudel reviewed the work and the manuscript. Both authors approved the final manuscript.