LOCUS: LABEL-GATED REGION PROPOSALS FOR PROTEIN FUNCTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Where function resides in a protein is tightly coupled to what that function is. Recent predictors pair protein language model representations with structural encoders and reduce them to a single protein-level vector. The design is effective but leaves no way to express how one region bears on one label. Catalytic residues and binding pockets are sharply localized, while many biological process and cellular component terms have no compact structural correlate at all. Routing every label through one pooled vector and routing every label through a site bottleneck therefore fail for opposite reasons. We introduce LOCUS, a framework that gives localized evidence its own pathway and keeps the global one. A class-agnostic region proposal network inspired by object detection emits multi-scale soft 3D regions over the residue graph, supervised only by residue-level annotations of active sites, binding sites and motifs. Graph non-maximum suppression selects among the proposals, and each survivor is pooled into a region-of-interest token. A classifier adds this regional evidence to global sequence and structure features as label-specific gated residuals. Every label then decides how much local evidence to admit, and no functional label enters the model as an input. Since the fusion is additive and decomposes exactly, the model that uses the proposal pathway can also measure its own contribution. On the standard Gene Ontology and enzyme classification benchmarks LOCUS exceeds published sequence- and structure-based predictors under the official evaluator. The regional pathway contributes in a strongly label-dependent way: it is worth a margin clear of zero on all three Gene Ontology branches and limited contributions on enzyme classification.

1 INTRODUCTION

Assigning functions to proteins from sequence and structure is central to biological annotation, yet experimental characterization covers only a fraction of known protein sequences (The UniProt Consortium, 2023). Gene Ontology (GO) represents function through the molecular function (MF), biological process (BP), and cellular component (CC) ontologies (Aleksander et al., 2023), while Enzyme Commission (EC) numbers organize catalytic activities. Predicting these labels is difficult in the remote-homology regime, where test proteins share little sequence identity with the training set and similarity-based transfer becomes unreliable.

Modern predictors combine protein language-model representations with structural encoders (Lin et al., 2023; Gligorijevic et al., 2021; Gu et al., 2023; Lee et al., 2024; Hu et al., 2024; Ouyang-Zhang et al., 2025; Zhu et al., 2026) and produce increasingly informative protein-level representations. A single pooled vector cannot distinguish residues that carry direct local evidence from residues that supply global fold or cellular context. The opposite extreme is inadequate too: many BP and CC terms cannot be reduced to one compact site. Local and global evidence therefore need separate computational roles. Functional residues are often sparse along the sequence yet coherent in three dimensions, which makes region-level learning natural for catalytic and binding evidence. A global route must stay available for diffuse functions so that each label can decide whether proposed regions add information. Unlike post-hoc saliency, such regions take part in the prediction itself and can be evaluated as an explicit model component.

We introduce LOCUS, a local–global architecture for direct multi-label protein function prediction. LOCUS projects frozen ESM-2 residue embeddings into a multi-scale sequence encoder, propagates

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information on a residue graph, and summarizes both streams with latent tokens. A functional region proposal network (RPN) inspired by object detection (Ren et al., 2015) learns multi-scale soft regions from the active-site, binding-site and motif annotations in VenusX (Tan et al., 2025). Each proposal becomes a region-of-interest (ROI) token. The final predictor adds label-specific gated structure and ROI residuals to a global sequence prediction. Compact local evidence can then contribute without forcing every function through a site bottleneck.

Our contributions can be summarized as follows:

- We introduce LOCUS, a local–global architecture for protein function prediction. A class-agnostic region proposal network supervised only by residue annotations emits multi-scale soft 3D regions, and their pooled ROI tokens enter the classifier as label-specific gated residuals on an always-available global pathway. Each label decides whether a compact site is relevant, rather than every function being forced through a site bottleneck.
- We show that this fusion makes the local pathway measurable rather than assumed. The fusion is additive and decomposes exactly, so the ROI term can be isolated with the rest of the network held fixed. That separates what the term is worth from what the classifier has learned to rely on. The two quantities are routinely conflated, and we report them apart.
- Experiments show that LOCUS is competitive on the Gene Ontology and enzyme-classification benchmarks, and that the proposal pathway contributes on every Gene Ontology branch while adding nothing measurable on enzyme classification. We treat that asymmetry as a result in its own right: a learned local pathway is not universally useful, and regional evidence earns its place only where it carries information the global routes do not. Gate and proposal-geometry diagnostics then separate localization quality from protein-level benefit and show that the two need not move together.

2 RELATED WORK

Protein function prediction. DeepFRI uses graph convolutions over residue contacts (Gligorijevic et al., 2021), HEAL combines hierarchical graph attention with contrastive learning (Gu et al., 2023), and DeepGO-SE treats prediction as approximate semantic entailment (Kulmanov et al., 2024). These methods optimize protein-level labels directly. LOCUS adds a trained local evidence pathway while retaining the global information diffuse functions require.

Representation learning and localized evidence. Protein language models learn residue features from large sequence corpora (Lin et al., 2023). Structure-aware learning has incorporated residue graphs through geometric pretraining (Zhang et al., 2023), equivariant interaction models (Kong et al., 2024), and structural modules inserted into protein language models (Hartout et al., 2025). Structure can also be coupled dynamically with sequence features (Hu et al., 2024), distilled into a sequence model (Ouyang-Zhang et al., 2025), or combined with surface reconstruction (Lee et al., 2024). FlexRibbon extends joint sequence–structure pretraining across prediction and design tasks (Zhu et al., 2026). These approaches strengthen global encoders but do not explicitly route supervised spatial regions into individual function predictions. LOCUS adds this local route alongside global sequence and structure summaries.

Functional residue learning and region proposals. VenusX provides residue-, fragment-, and domain-level tasks with identity-controlled splits (Tan et al., 2025), which makes supervised functional-region learning feasible at scale. Region proposal networks were introduced for object detection, where class-agnostic candidate regions are scored before any category decision is made (Ren et al., 2015). LOCUS adapts this mechanism to proteins through four design choices: proposal supervision is derived from four curated VenusX residue tasks; proposals are adaptive multi-radius three-dimensional regions rather than fixed graph neighborhoods; the global sequence pathway remains ungated; and local evidence enters only as a learned residual whose contribution is observable through label-specific gates. This design avoids the assumption that every GO label corresponds to one compact site.

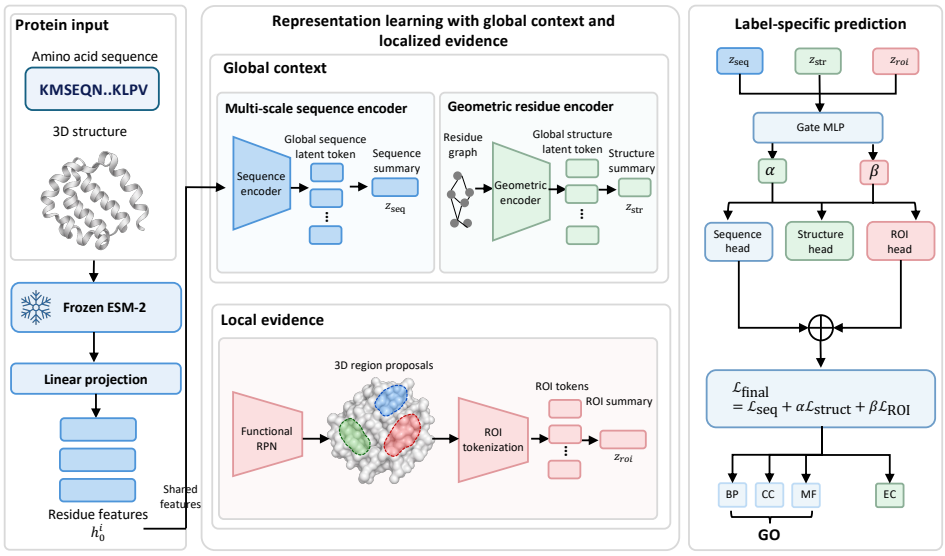


Figure 1: Overview of LOCUS. Frozen protein language model embeddings feed a sequence stream and a geometric stream over the residue graph. The functional RPN proposes soft 3D regions that are pooled into ROI tokens, and label-specific gates fuse the three protein summaries at the logit level.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Let a protein contain L residues with amino-acid sequence $x = (x_1, \ldots, x_L)$, coordinates $\boldsymbol{r}_i \in \mathbb{R}^3$ when observed, confidence values q_i, and a residue-validity mask. For a label space o containing G_o GO terms or EC classes, the target is $\boldsymbol{y}_o \in \{0,1\}^{G_o}$. LOCUS predicts all labels directly; label identifiers, descriptions, and ontology embeddings are never model inputs.

Figure 1 summarizes the architecture. A frozen protein language model supplies residue embeddings. A sequence stream encodes motifs at several receptive fields, and a geometric stream propagates those features on a residue graph. In parallel, the functional RPN proposes up to $K = 16$ soft regions. The model pools proposals into ROI tokens and forms three protein summaries: $\boldsymbol{z}_{\mathrm{seq}}$, $\boldsymbol{z}_{\mathrm{str}}$, and $\boldsymbol{z}_{\mathrm{roi}}$. Label-specific gates fuse the summaries at the logit level.

3.2 GLOBAL SEQUENCE AND STRUCTURE REPRESENTATIONS

We project frozen ESM-2 650M layer-33 embeddings $\boldsymbol{e}_i \in \mathbb{R}^{1280}$ to $\boldsymbol{h}_i^0 \in \mathbb{R}^{256}$. The sequence encoder applies three depthwise-separable convolutional branches with kernel–dilation pairs $(3,1)$, $(9,1)$, and $(9,3)$. A residue-wise softmax gate combines their outputs with a residual connection:

$$\boldsymbol{h}_i^{\mathrm{seq}} = \mathrm{LN}\left(\boldsymbol{h}_i^0 + \sum_{s=1}^{3} \gamma_{is}\, f_s(\boldsymbol{H}^0)_i\right), \qquad \boldsymbol{\gamma}_i = \mathrm{softmax}(g(\boldsymbol{h}_i^0)). \tag{1}$$

Eight learned latent tokens attend to valid residues in two cross-attention blocks, and their mean gives $\boldsymbol{z}_{\mathrm{seq}} \in \mathbb{R}^{256}$.

The geometric encoder operates on the union of short-range sequence edges, $|i-j| \le 2$ within a chain, and 30-nearest-neighbor spatial edges among residues with coordinates. Edge attributes encode radial basis distances, clipped sequence separation, chain identity, coordinate confidence, and edge type. Four fixed-coordinate message-passing blocks produce $\boldsymbol{H}^{\mathrm{str}}$. A second eight-token pooler yields $\boldsymbol{z}_{\mathrm{str}}$. Keeping the two summaries separate lets the output head measure whether geometry adds evidence beyond the sequence representation.

3.3 SUPERVISED FUNCTIONAL-REGION PROPOSALS

The RPN predicts spatial regions that may contain functional evidence without conditioning on the requested GO or EC label. For residue i, it forms

$$\boldsymbol{u}_i = [\boldsymbol{h}_i^{\mathrm{seq}}; \boldsymbol{h}_i^{\mathrm{str}}; \boldsymbol{h}_i^{\mathrm{seq}} \odot \boldsymbol{h}_i^{\mathrm{str}}; \boldsymbol{z}_{\mathrm{str}}] \tag{2}$$

and predicts a seed score, scale scores, and bounded center and per-scale radius offsets. For every selected seed, the RPN instantiates candidates with all three base radii $(6, 12, 20)$ Å; scale logits contribute to their ranking. Training uses straight-through top-k seed selection, whereas evaluation uses deterministic top-k. For a proposal a with center $\boldsymbol{c}_a$ and radius r_a, its soft membership for residue i is

$$m_{ia} = \sigma\left(\frac{r_a - \|\boldsymbol{r}_i - \boldsymbol{c}_a\|_2}{\tau_d}\right) \sigma\big(h_{\mathrm{mask}}([\boldsymbol{h}_i^{\mathrm{str}}; \boldsymbol{u}_a])\big)\,. \tag{3}$$

Soft-Dice non-maximum suppression removes redundant regions and retains at most 16 proposals.

We derive proposal targets from the Act, BindI, BindB, and Motif residue tasks in VenusX. The executed preprocessing snapshot preserves every distinct source-defined site instance as one exact component. An exact mask supervises residue refinement, while a 6 Å-expanded mask supervises region matching. Anchors with intersection-over-union at least 0.5 are positive, those at most 0.2 are negative, and the best anchor is forced positive when a site would otherwise be unmatched.

The site objective combines focal objectness loss, region Dice loss, focal-plus-Dice exact-mask loss, smooth-ℓ_1 radius regression, and pairwise proposal diversity:

$$\mathcal{L}_{\mathrm{RPN}} = \mathcal{L}_{\mathrm{obj}} + \mathcal{L}_{\mathrm{region}} + \mathcal{L}_{\mathrm{mask}} + 0.1\mathcal{L}_{\mathrm{radius}} + 0.1\mathcal{L}_{\mathrm{div}}. \tag{4}$$

The RPN is selected by macro-average validation residue AUPR across the four tasks. Functional-site annotations supervise a class-agnostic proposal mechanism; they are not converted into GO labels.

3.4 ROI TOKENS AND GATED RESIDUAL FUSION

For proposal a, weighted means of sequence and structure features are concatenated with objectness, radius, relative size, mean confidence, and coordinate coverage. An MLP maps this vector to a 256-dimensional ROI token. A two-layer, eight-head Transformer exchanges information among the retained tokens, and masked attention pooling produces $\boldsymbol{z}_{\mathrm{roi}}$.

For label space o, separate heads compute sequence, structure, and ROI logits. The final logits are

$$\boldsymbol{\alpha}_o = \sigma(g_o^{\mathrm{str}}([\boldsymbol{z}_{\mathrm{seq}}; \boldsymbol{z}_{\mathrm{str}}; \boldsymbol{z}_{\mathrm{roi}}])), \tag{5}$$

$$\boldsymbol{\beta}_o = \sigma(g_o^{\mathrm{roi}}([\boldsymbol{z}_{\mathrm{seq}}; \boldsymbol{z}_{\mathrm{str}}; \boldsymbol{z}_{\mathrm{roi}}])), \tag{6}$$

$$\boldsymbol{\ell}_o = \boldsymbol{\ell}_o^{\mathrm{seq}} + \boldsymbol{\alpha}_o \odot \boldsymbol{\ell}_o^{\mathrm{str}} + \boldsymbol{\beta}_o \odot \Delta\boldsymbol{\ell}_o^{\mathrm{roi}}. \tag{7}$$

The ROI gate is initialized conservatively so that training begins near the global predictor. Equation 7 also exposes the route by which local evidence changes each prediction.

3.5 FUNCTION TRAINING

We transfer only shared representation and proposal parameters from site pretraining; label heads are freshly initialized. All reported results use the same validation-selected site checkpoint. The site objective is optimized in that stage alone and never alongside function training; Appendix D reports what happens when it is. The source snapshot that produced the site checkpoint is identified by digest but is no longer on disk, so it must be reconstructed before artifact release. Appendix E describes how the evaluation pipeline was consolidated to a single revision. Function training uses asymmetric multi-label loss (Ridnik et al., 2021) on the fused logits and auxiliary sequence, structure, and ROI logits:

$$\mathcal{L}_o = \mathcal{L}_{\mathrm{ASL}}(\boldsymbol{\ell}_o, \boldsymbol{y}_o) + 0.2\mathcal{L}_{\mathrm{ASL}}(\boldsymbol{\ell}_o^{\mathrm{seq}}, \boldsymbol{y}_o) + 0.1\mathcal{L}_{\mathrm{ASL}}(\boldsymbol{\ell}_o^{\mathrm{str}}, \boldsymbol{y}_o) + 0.1\mathcal{L}_{\mathrm{ASL}}(\Delta\boldsymbol{\ell}_o^{\mathrm{roi}}, \boldsymbol{y}_o). \tag{8}$$

An auxiliary ROI masked-reconstruction loss predicts stop-gradient projected PLM features for a fraction of the proposal-weighted residues. Only the non-PLM parameters are trained, with AdamW (Loshchilov & Hutter, 2019).

4 EXPERIMENTS

4.1 DATASETS AND LEAKAGE CONTROL

For protein-level prediction, we use the HEAL adaptation of the DeepFRI benchmark (Gligorijevic et al., 2021; Gu et al., 2023). GO prediction uses 1,943 BP, 320 CC, and 489 MF labels; EC prediction uses 538 classes. All evaluations use the test subsets capped at 95% sequence identity to training proteins, following the protocol of the published rows we compare against so that the comparison is made on the membership those rows were scored on.

Training pool. For protein-level labels, GO models train on the benchmark's PDBch+AFch training split and EC models on its EC split, and no protein outside those splits contributes a function label. The proposal network is pretrained separately on residue-level VenusX annotations, so LOCUS draws on supervision that the rows of Table 1 do not use; that table matches them in metric, threshold rule and protein-level training data, and the residue-level supervision is an addition our ablations quantify. Every baseline we train ourselves uses the same pool as the LOCUS model it is compared against, including the sequence-only probe of Section 4.3. We score the 3,410 GO test chains and the 1,919 EC test proteins that survive the frozen embedding and structure intersection. The benchmark also distributes stricter identity subsets of the test set; we report the 95% cut only.

Site supervision uses the VenusX MP50 partitions (Tan et al., 2025). A record is kept only after residue mapping, coordinate validation, duplicate removal and homology filtering against the GO validation and test proteins. Ambiguous and coordinate-missing records are excluded, leaving 9,352 training, 1,578 validation and 1,643 untouched test records.

4.2 EVALUATION PROTOCOL

GO metrics follow the frozen HEAL ontology and evaluation implementation, with predictions propagated over the GO DAG before scoring. We report function-centric AUPR and semantic distance (Jiang et al., 2016), where the reported S is the HEAL-normalized value rather than the unnormalized CAFA sum.

HEAL `fmax` returns the maximum protein F-score over the threshold grid $t/100$ for $t = 1, \ldots, 99$ evaluated on the test set, and $S_{\min}$ is the corresponding minimum. Every published row of Table 1 is that quantity, so we adopt it unchanged and denote it $F_{\max}$. This concerns reporting only. Checkpoint selection, early stopping and every other training decision are made on validation and never consult the test split, exactly as in HEAL. The secondary EC evaluation applies the same sweep to the raw 538-label matrix, which has no ontology DAG.

For site localization, we report residue AUPR by task because unannotated residues dominate accuracy. The matched ESM-2 probe uses the same VenusX test partitions and validation-selected thresholds.

4.3 GENE ONTOLOGY PREDICTION

On the frozen GO test evaluation LOCUS reaches $F_{\max} = 0.616$, 0.684 and 0.774 on BP, CC and MF, with function-centric AUPR of 0.394, 0.465 and 0.741 and $S_{\min}$ of 0.499, 0.463 and 0.308 (Table 5, *full* column). Table 1 places these values beside those published for this benchmark lineage, where LOCUS leads every column but the cellular-component semantic distance. Metric definition, threshold rule and training set all agree with the cited rows, so the comparison turns on membership alone: those rows score 3,416 chains and LOCUS scores the 3,410 that survive the frozen embedding and structure intersection.

The language-model rows of Table 1 are the relevant control for a model built on frozen ESM-2 features, but they are published numbers on a different test set. To control the split as well, we train a sequence-only frozen ESM-2 650M layer-33 mean-pool MLP on our exact training proteins, embeddings, label maps, evaluator and reporting protocol, with validation-selected checkpoints. It reaches $F_{\max}$ of 0.557, 0.682 and 0.671 on BP, CC and MF at the 95% cut, with function AUPR of 0.292, 0.449 and 0.604. The probe shares everything with LOCUS but the read-out, so the difference between them isolates what the structural encoder and the proposal pathway add to a language model that already has the same data. That difference is large on BP ($+0.060$ $F_{\max}$, $+0.102$ AUPR) and

on MF (+0.103, +0.137). It is narrow on CC, where LOCUS leads by 0.002 $F_{\max}$ and 0.016 AUPR. The pattern tracks the biology. A protein's compartment is set by targeting signals that are sequence-local, so on cellular component neither three-dimensional geometry nor a compact proposed region carries information that a mean over residues does not already hold.

Benchmark position establishes that the architecture is competitive. It does not isolate what the proposal pathway accounts for, which is the question the rest of this section addresses. Section 4.5 removes that pathway while keeping the encoder, and Section 4.7 varies how functional-site supervision reaches it.

Table 1: GO prediction on the DeepFRI/HEAL benchmark. Bold marks the best value in each column.

	$F\uparrow$			Function AUPR ↑			Semantic distance ↓		
Method	BP	CC	MF	BP	CC	MF	BP	CC	MF
BLAST (Altschul et al., 1990)	0.336	0.448	0.328	0.067	0.097	0.136	0.651	0.628	0.632
FunFams (Das et al., 2015)	0.500	0.627	0.572	0.260	0.288	0.367	0.579	0.503	0.531
DeepGO (Kulmanov et al., 2018)	0.493	0.594	0.577	0.182	0.263	0.391	0.577	0.550	0.472
DeepFRI (Gligorijevic et al., 2021)	0.540	0.613	0.625	0.261	0.274	0.495	0.543	0.527	0.437
HEAL (Gu et al., 2023)	0.581	0.673	0.708	0.298	0.415	0.630	0.504	**0.462**	0.369
ProtT5 (Elnaggar et al., 2021)	0.327	0.623	0.511	0.056	0.284	0.294	0.637	0.526	0.552
SAProt (Su et al., 2023)	0.374	0.506	0.287	0.034	0.037	0.018	0.643	0.612	0.662
ESM-2 (Lin et al., 2023)	0.351	0.633	0.531	0.062	0.282	0.298	0.622	0.507	0.563
DeepGO-SE (Kulmanov et al., 2024)	0.566	0.636	0.654	0.233	0.423	0.495	0.530	0.481	0.435
TAWFN (Meng & Wang, 2024)	0.548	0.609	0.711	0.279	0.346	0.674	0.561	0.539	0.393
PFresGO (Pan et al., 2025)	0.568	0.674	0.692	0.293	0.361	0.602	0.535	0.498	0.417
LOCUS	**0.616**	**0.684**	**0.774**	**0.394**	**0.465**	**0.741**	**0.499**	0.463	**0.308**

4.4 ENZYME CLASSIFICATION

On the secondary 538-label EC task LOCUS reaches $F_{\max} = 0.880$ and function-centric AUPR of 0.881 on 1,919 proteins, with $S_{\min} = 0.129$ on the raw 538-label matrix. EC has no ontology DAG, so this last value is not comparable to the GO column. Table 2 lists published EC F-scores under the same sweep; they differ in split membership and report no comparable AUPR. LOCUS is the highest of them, ahead of GearNet-Edge with multiview pretraining by 0.006. That margin is smaller than the differences in split membership between the sources. We therefore read the table as placing LOCUS among the strongest EC predictors rather than as establishing an ordering at the top.

EC earns its place in this paper for a second reason, and it is the more interesting one. A central finding of this study is that a learned local pathway is not universally useful. On EC LOCUS predicts strongly, and yet the ROI residual itself is statistically inert: Section 4.5 shows both its $F_{\max}$ and its AUPR effects spanning zero. The two facts are not in tension, because they concern different things — how well the model predicts, and whether one particular route inside it carries information the others do not. Explicit regional evidence should be admitted only where it contributes beyond the global sequence and structure routes, and EC is the label space where it does not.

Table 2: EC prediction, ordered by F-score. Bold marks the best value.

Method	F-score ↑
DeepFRI (Gligorijevic et al., 2021)	0.631
GearNet (Zhang et al., 2023)	0.730
GearNet-Edge (Zhang et al., 2023)	0.810
GET-IEConv (Kong et al., 2024)	0.835
ESM-1b (Rives et al., 2021)	0.864
GearNet-Edge + multiview pretraining (Zhang et al., 2023)	0.874
LOCUS	**0.880**

4.5 CONTRIBUTION OF THE PROPOSAL PATHWAY

Before asking whether site supervision helps, we ask whether the ROI pathway contributes anything to protein-level accuracy. We disable it at inference in the trained models of Section 4.3: the residual

Table 3: LOCUS with and without the ROI residual. Bold marks intervals excluding zero.

		$F_{\max}$ ↑			Function AUPR ↑		
Task	n	full	no ROI	Δ [95% CI]	full	no ROI	Δ [95% CI]
GO-BP	3,410	0.616	0.606	**+0.010** [+0.009, +0.012]	0.394	0.376	**+0.017** [+0.016, +0.019]
GO-CC	3,410	0.684	0.679	**+0.005** [+0.004, +0.007]	0.465	0.454	**+0.011** [+0.009, +0.014]
GO-MF	3,410	0.774	0.768	**+0.006** [+0.004, +0.007]	0.741	0.730	**+0.011** [+0.009, +0.013]
EC	1,919	0.880	0.880	+0.001 [−0.001, +0.001]	0.881	0.881	+0.000 [−0.001, +0.001]

$\boldsymbol{\beta}_o \odot \Delta\boldsymbol{\ell}_o^{\mathrm{roi}}$ is dropped and $\boldsymbol{z}^{\mathrm{roi}}$ is zeroed inside the gate input, leaving $\boldsymbol{\ell}_o = \boldsymbol{\ell}_o^{\mathrm{seq}} + \boldsymbol{\alpha}_o \odot \boldsymbol{\ell}_o^{\mathrm{str}}$. The RPN still runs; only its route into the classifier is cut. Reusing the full model's threshold would bias the comparison, since that threshold was chosen with the residual active. We therefore repeat the validation selection for the ablated model and score it at its own threshold. Uncertainty is a paired bootstrap over test proteins, resampling both arms with the same indices.

The pathway's value varies in magnitude across label spaces but never in sign. On BP it is worth $+0.010$ $F_{\max}$ [$+0.009$, $+0.012$] and $+0.017$ AUPR [$+0.016$, $+0.019$]. On CC the two figures are $+0.005$ [$+0.004$, $+0.007$] and $+0.011$ [$+0.009$, $+0.014$], and on MF $+0.006$ [$+0.004$, $+0.007$] and $+0.011$ [$+0.009$, $+0.013$], with every interval clear of zero. On EC the pathway is inert: $+0.001$ $F_{\max}$ and $+0.000$ AUPR, both with intervals spanning zero. The resulting ordering BP $>$ CC $\approx$ MF $\gg$ EC tracks how far each label space depends on localized evidence in the first place, which is what a pathway carrying such evidence should produce.

CC repays a closer look, because there the two instruments disagree. A matched sequence-only probe already reaches LOCUS's protein-level accuracy on that branch (Section 4.3), yet removing the ROI residual still costs as much as it does on MF. The pathway is supplying evidence that the global read-out can also reach by another route, which is not the same as supplying nothing. Appendix A shows the same branch from the inside.

The removal answers a narrower question than it first appears to, and a second measurement separates the two. Dropping the residual also zeroes $\boldsymbol{z}^{\mathrm{roi}}$ inside the gate input, so $\boldsymbol{\alpha}_o$ is recomputed on an input the gate MLP never saw in training. The ablation thus perturbs the structure gate as well as removing the ROI term. Appendix A isolates the ROI term alone by holding both gates at the values the full forward pass produced. On EC the two measurements agree to the fourth decimal, as inertness requires. On the GO branches they differ: on BP the gate-fixed ROI term is worth $+0.020$ function AUPR against the removal's $+0.017$, and the gap is the structure gate adapting to expect a route that is then cut. We quote the gate-fixed figure when asking what the ROI term is worth, and the removal when asking how much the trained classifier relies on the route being present.

Two quantities therefore exist and should not be interchanged: what the ROI term is worth with both gates held fixed, and what removing the route costs a classifier trained with it available. The first measures the term and the second measures the reliance the classifier has learned on it. Both should be read against the margin over the matched sequence-only probe of Section 4.3, which shares the training set and the embeddings and differs from LOCUS only in the read-out.

4.6 WHAT THE PROPOSAL BRANCH LOCALIZES

Localization is strongly task-dependent. On the VenusX test partitions the site checkpoint reaches residue AUPR 0.680 on BindI and 0.715 on Motif against a matched ESM-2 probe's 0.374 and 0.494, and leads it on the four-task macro average by 0.458 to 0.310. It leads on Act (0.223 versus 0.198) and on BindB (0.214 versus 0.174) as well, though by margins far narrower than on the other two tasks. Appendix B gives the full table and traces that asymmetry to two matched causes. The first is the smallest radius the RPN may propose: adding finer scales while keeping the coarse ones doubles Act and raises BindB by half without costing BindI. The second is a genuine conflict among the four tasks over training time, which the macro selection criterion silently resolves against Act and BindB, so the reported checkpoint is not the one that would serve those two tasks best. These numbers describe the site checkpoint, whose proposal geometry differs markedly from that of the function checkpoints (Appendix C). The case that follows concerns the regions an MF model actually uses.

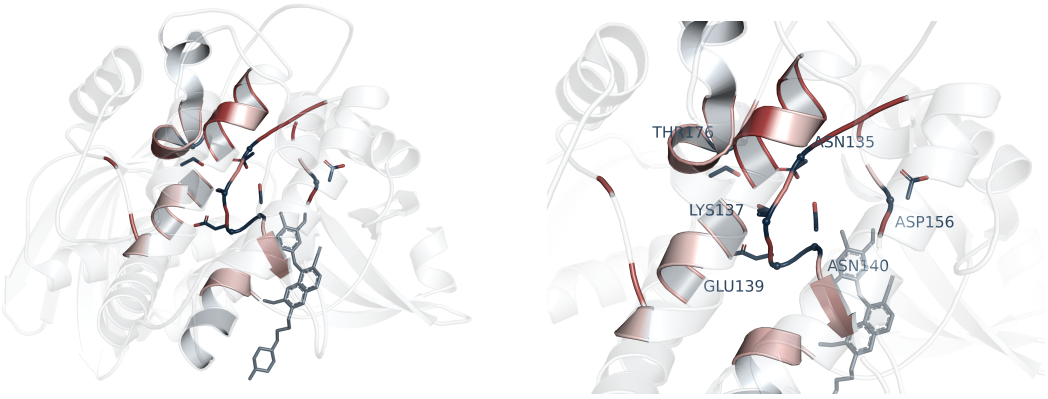


Figure 2: Proposal mass on 3SOA chain A. Left: whole chain. Right: the active site.

Table 4: MF ablation on the site warm start.

	95% identity ($n = 3{,}410$)		
Site supervision enters via	$F_{\max} \uparrow$	AUPR $\uparrow$	$S_{\min} \downarrow$
Pretrained proposals (LOCUS)	**0.774**	**0.741**	**0.308**
No site supervision	0.757	0.725	0.327

Figure 2 shows one such region. The chain is 3SOA A, a 436-residue kinase with no close sequence neighbor among the training proteins. Its backbone is colored by per-residue ROI mass on the contribution ramp of Figure 1, and the structure outside the region is faded so that the region, which sits in a buried cleft, remains visible through it. Sticks mark the six residues BioLiP annotates as catalytic, and the co-crystallized inhibitor DB8 is drawn in slate for orientation only. All six catalytic residues (N135, K137, E139, N140, D156 and T176) fall inside the ten highest-weighted ROI residues, and 42% of the chain's proposal mass sits on those six of 436 positions. The four remaining residues in that top ten are immediate sequence or spatial neighbors of the annotated ones, so the mass forms one contiguous pocket rather than a scatter of high-scoring positions. Neither the ligand nor any site annotation is a model input, and BioLiP's catalytic-residue field is not the supervision the proposal network saw. The case shows what a proposal region looks like on a chain the model has no close training neighbor for, and what structure the pathway converges on when left to choose freely.

4.7 ROUTES BY WHICH SITE SUPERVISION ENTERS

Appendix B shows that site-pretrained proposals localize InterPro binding sites and motifs well. It does not follow that this localization ability improves protein-level prediction. To test the link we trained two MF models that share configuration, data, splits, optimizer, seed, rank count and global batch. They differ in one thing: whether the shared encoder and proposal network are warm-started from the VenusX site checkpoint, which is LOCUS's design, or left randomly initialized. The proposal pathway is architecturally present in both arms, and neither co-trains the site loss during function training. Each arm selects its own threshold on validation, so the comparison isolates the warm start.

The experiment separates two claims that are easily conflated. One is whether the pathway contributes to protein-level accuracy, which Section 4.5 answers affirmatively. The other is whether the proposal mechanism learns to localize particular annotation classes, which Appendix B answers affirmatively on VenusX. What remains open is whether the second property can be converted into the first. The warm start is the most direct route by which residue annotations could do the converting, and it is worth $+0.017$ $F_{\max}$, $+0.017$ function AUPR and -0.019 in semantic distance. Annotations therefore do transfer, but the margin they buy is close to what cutting the ROI route costs on the same branch, so the pathway matters more than the supervision that initializes it.

Neither arm co-trains the site loss, so Appendix D tests that remaining option on EC, where the runs are complete. Optimizing the site objective alongside function prediction is worse than leaving it off on both metrics, and the proposal telemetry in Section 4.8 indicates why.

4.8 Gate and proposal diagnostics

The ablations establish where the pathway helps; the gates and the proposal geometry indicate why. Mean ROI gate values are 0.24 on BP, 0.25 on CC, 0.24 on MF and 0.12 on EC, while the structure gate runs 0.25, 0.32 and 0.39 on the first three. Neither residual is uniformly enabled or disabled, and the model discriminates between label spaces where the ablations say it should. Recorded mean soft proposal masses are of order 10^{-5}, far below the multi-residue regions the model is meant to emit. That figure is uninformative on its own, because `GraphROIPool` divides the pooled summaries by the same mass and only the relative weighting reaches the ROI tokens. The cancellation is not complete, and Appendix C records the two ROI metadata channels it leaves inert. Measuring the normalized geometry directly there, on a fixed set of 256 proteins, contradicts the natural reading: function training does not shrink the regions but widens them, from the site checkpoint's median 6.0 effective residues to 23.6 on the EC function checkpoint. The 10^{-8} denominator clamp is never reached, so pooling remains a well-defined weighted average. EC sharpens the picture. The regions there are broad and the pathway is nonetheless inert, and the cosine between the ROI summary and a masked mean of the same features reaches 0.85, consistent with a pathway that has collapsed onto global pooling. Breadth alone therefore does not make a region useful. What matters is whether the region isolates evidence the global route does not already carry.

5 Discussion and Conclusion

We introduced Locus, a local–global model that learns functional-region proposals from residue annotations and admits them through label-specific residual gates. We also used it as an instrument for measuring when such a pathway is worth having. The gates make both uses possible: each label decides how much local evidence to take, and the fusion stays additive enough to be decomposed exactly. Measured that way, the pathway pays for itself on all three Gene Ontology branches while contributing nothing measurable on enzyme classification, and the ROI term itself is worth somewhat more than an inference-time ablation alone would suggest. That asymmetry is a central finding rather than a shortfall of the enzyme model, which predicts as strongly as any row we compare against: a learned local pathway is not universally useful, and explicit regional evidence should be admitted only where it contributes beyond the global sequence and structure routes.

A second question follows: whether the residue annotations the pathway is trained on are what make it work. Warm-starting the proposal network from those annotations is worth $+0.017$ function AUPR on molecular function, a margin comparable to what cutting the ROI route costs there, while co-training the same objective on enzyme classification is worse on both metrics. The proposal telemetry indicates why localization quality and protein-level benefit come apart. Site supervision sharpens the regions toward the sparse annotation rate, far enough under co-training to clamp the pooling denominator on most proposals, so faithfulness to annotations and usability by the classifier pull against each other. The component to change is therefore the pooling operator rather than the amount of supervision. A pooling operator that stays well posed as masks sharpen is the target this analysis leaves for future work.

AI use statement

In this work, we used generative AI tools (Claude, Anthropic) to assist with running experiments and editing the manuscript text. We did not use generative AI tools to generate the core scientific ideas or experimental design. All AI-assisted code, text, and figures were reviewed by the authors before inclusion, and all reported numerical results were verified against raw experiment output logs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work analyzes publicly available protein sequence, structure, ontology, and functional-site resources and does not involve human participants or identifiable personal data. Protein-function predictions can nevertheless be unreliable outside the evaluated data regime. The model outputs should not be treated as validated biomedical conclusions without appropriate experimental confirmation.

Reproducibility statement

Section 3 specifies the model and objectives, and Section 4 defines the split and threshold protocol. The appendix records the evaluation protocol and the provenance digests each run is pinned to. Frozen run and evaluation artifacts record label-map, split, ontology, metric-code, checkpoint and evaluation-code digests. A complete anonymous release still requires reconstruction and testing in a clean environment. Three gaps remain at submission time. Two model modules, `functional_rpn.py` and `geometric_encoder.py`, are present in our working tree only as compiled bytecode and in frozen execution snapshots. The working tree is not under version control. The source snapshot that produced the site warm start, which initializes every model reported here, is identified by digest but is no longer on disk. Reconstructing all three is a prerequisite for the artifact release.

References

Suzi A. Aleksander, James Balhoff, Seth Carbon, J. Michael Cherry, Harold J. Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L. Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2.

Sayoni Das, David Lee, Ian Sillitoe, Natalie L. Dawson, Jonathan G. Lees, and Christine A. Orengo. Functional classification of CATH superfamilies: A domain-based approach for protein function annotation. *Bioinformatics*, 31(21):3460–3467, 2015.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021.

Vladimir Gligorijevic, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, 2021.

Zhonghui Gu, Xiao Luo, Jiaxiao Chen, Minghua Deng, and Luhua Lai. Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics*, 39(7):btad410, 2023.

Philip Hartout, Dexiong Chen, Paolo Pellizzoni, Carlos Oliver, and Karsten Borgwardt. Endowing protein language models with structural knowledge. *Bioinformatics*, 41(11):btaf582, 2025. doi: 10.1093/bioinformatics/btaf582.

Bozhen Hu, Cheng Tan, Jun Xia, Yue Liu, Lirong Wu, Jiangbin Zheng, Yongjie Xu, Yufei Huang, and Stan Z. Li. Learning complete protein representation by dynamically coupling of sequence and structure. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yuxiang Jiang, Tal Ronnen Oron, Wyatt T. Clark, Asma R. Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S. Funk, Indika Kahanda, Karin M. Verspoor, Asa Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1):184, 2016. doi: 10.1186/s13059-016-1037-6.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Generalist equivariant transformer towards 3D molecular interaction learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25149–25175, 2024.

Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2018. doi: 10.1093/bioinformatics/btx624.

Maxat Kulmanov, Francisco J. Guzman-Vega, Paula Duek Roggli, Lydie Lane, Stefan T. Arold, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, 6(2):220–228, 2024.

Youhan Lee, Hasun Yu, Jaemyung Lee, and Jaehoon Kim. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In *International Conference on Learning Representations*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu Meng and Xiaoran Wang. TAWFN: A deep learning framework for protein function prediction. *Bioinformatics*, 40(10):btae571, 2024. doi: 10.1093/bioinformatics/btae571.

Jeffrey Ouyang-Zhang, Chengyue Gong, Yue Zhao, Philipp Krähenbühl, Adam R. Klivans, and Daniel J. Diaz. Distilling structural representations into protein sequence models. In *International Conference on Learning Representations*, 2025.

T. Pan, G. I. Webb, S. Imoto, and J. Song. Integrating gene ontology relationships for protein function prediction using PFresGO. In *Methods in Molecular Biology*, volume 2947, pp. 161–169. Springer, 2025. doi: 10.1007/978-1-0716-4662-5_9.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SAProt: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023. doi: 10.1101/2023.10.01.560349.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. VenusX: Unlocking fine-grained functional understanding of proteins. *arXiv preprint arXiv:2505.11812*, 2025. doi: 10.48550/arXiv.2505.11812.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations*, 2023.

Jianwei Zhu, Yu Shi, Ran Bi, Peiran Jin, Chang Liu, Zhe Zhang, et al. FlexRibbon: Joint sequence and structure pretraining for protein modeling. In *International Conference on Learning Representations*, 2026.

A PATHWAY DECOMPOSITION

Section 4.5 removes the ROI residual and Section 4.3 compares against a matched sequence-only model. Neither isolates what LOCUS's own global sequence pathway achieves alone. Because the fusion is exact,

$$\boldsymbol{\ell}_o = \boldsymbol{\ell}_o^{\mathrm{seq}} + \boldsymbol{\alpha}_o \odot \boldsymbol{\ell}_o^{\mathrm{str}} + \boldsymbol{\beta}_o \odot \boldsymbol{\ell}_o^{\mathrm{roi}},$$

every partial sum can be scored from a single forward pass. We record the three logit tensors and both gates for each test protein and read out four arms: the sequence pathway alone, either residual added to it, and the full model. No arm has a training or checkpoint-selection advantage over another; only the summation differs. Recomputing the full arm reproduces the frozen score matrices to within 2×10^{-7}, which is what licenses reading the other three as exact decompositions rather than approximations.

Table 5: Pathway decomposition of LOCUS.

	LOCUS read-outs				ESM-2
Metric	seq	+ ROI	+ str	full	ESM-2
GO-BP					
$F_{\max} \uparrow$	0.597	0.612	0.607	**0.616**	0.351
AUPR $\uparrow$	0.366	0.386	0.378	**0.394**	0.062
$S_{\min} \downarrow$	0.515	0.503	0.508	**0.499**	0.622
GO-CC					
$F_{\max} \uparrow$	0.666	0.679	0.674	**0.684**	0.633
AUPR $\uparrow$	0.438	0.457	0.452	**0.465**	0.282
$S_{\min} \downarrow$	0.479	0.467	0.471	**0.463**	0.507
GO-MF					
$F_{\max} \uparrow$	0.751	0.762	0.768	**0.774**	0.531
AUPR $\uparrow$	0.706	0.725	0.731	**0.741**	0.298
$S_{\min} \downarrow$	0.334	0.320	0.318	**0.308**	0.563
EC					
$F_{\max} \uparrow$	0.878	0.878	0.880	0.880	0.677
AUPR $\uparrow$	0.874	0.877	0.880	**0.881**	0.667
$S_{\min} \downarrow$	0.131	0.130	0.129	**0.129**	0.347

Two conclusions follow, and the second bears on which component to change.

The fusion behaves as designed on every label space. Function AUPR increases monotonically from *seq* through either residual to *full* on BP, CC, MF and EC, and $S_{\min}$ decreases in step with no exception. The two residuals together are worth $+0.028$ AUPR on BP, $+0.027$ on CC, $+0.035$ on MF and $+0.007$ on EC at the 95% cut, and the 2×10^{-7} agreement recorded above makes these exact partial sums rather than approximations. The one place a residual costs anything is EC $F_{\max}$, where adding the ROI term moves the fourth decimal the wrong way while raising AUPR. That is the same inertness Section 4.5 reports for EC.

Which residual is the larger one is not fixed across label spaces, and the split is informative. On MF and EC the structure residual dominates, adding $+0.025$ and $+0.006$ AUPR against the ROI residual's $+0.019$ and $+0.003$. On BP and CC the order reverses: the ROI residual is worth $+0.020$ and $+0.019$ AUPR on its own against $+0.013$ for the structure residual on both. Those are the two branches whose labels are least reducible to a single fold. They are also where removing the ROI route costs the most relative to what the structural encoder supplies, so the gate-fixed decomposition and the removal ablation agree on where the proposal pathway matters.

The rightmost column places the decomposition against a language-model baseline. LOCUS's global sequence pathway alone, before either residual is added, already exceeds it by 0.304 function AUPR on BP, 0.156 on CC and 0.408 on MF, so the residuals extend a lead the global pathway holds rather than repairing a deficit.

B RESIDUE LOCALIZATION ON VENUSX

Table 6 reports residue AUPR on the four VenusX tasks for the site checkpoint and for a matched ESM-2 probe. The site checkpoint leads on BindI, on Motif and on the macro average, and trails on Act and on BindB.

Table 6: Residue AUPR on the VenusX test partitions. Both rows use the same partitions and validation-selected thresholds. Bold marks the best value in each column.

Model	Act	BindI	BindB	Motif	Macro
ESM-2 probe	0.198	0.374	0.174	0.494	0.310
LOCUS site	**0.223**	**0.680**	**0.214**	**0.715**	**0.458**

The error pattern is consistent with a geometric mismatch between spherical proposals and some annotation types. BindI and Motif often provide spatially coherent or contiguous supervision, whereas BioLiP contacts can be dispersed, ligand-dependent, or sensitive to structure and chain mapping. Act annotations can contain small catalytic constellations that expanded spherical regions match poorly.

Two measurements replace that reading with something more specific. Both come from a sweep of sixteen site-training arms sharing driver, epoch budget, global batch and seed. The arms differ only in `proposal_radii` and in whether the region Dice term is active, so comparisons within the sweep are matched.

Table 7: Proposal-radius sweep on the VenusX test partitions. Arms within each block differ only in `proposal_radii`.

	Proposal radii (Å)	Act	BindI	BindB	Motif	Macro
A. macro-selected						
control	$\{6, 12, 20\}$	0.083	0.420	0.093	**0.664**	0.315
spanning	$\{2, 3, 4, 6, 8, 12, 16, 20\}$	**0.168**	**0.507**	**0.140**	0.634	**0.362**
B. macro-selected, region Dice active						
control	$\{6, 12, 20\}$	0.086	**0.648**	0.089	**0.708**	0.383
spanning	$\{2, 4, 6, 8, 12, 20\}$	**0.133**	0.612	**0.148**	0.645	**0.384**
C. Act+BindB-selected, same runs as A						
control	$\{6, 12, 20\}$	0.169	0.266	0.127	**0.512**	0.268
spanning	$\{2, 3, 4, 6, 8, 12, 16, 20\}$	**0.198**	**0.289**	**0.140**	0.509	**0.284**

First, the smallest radius the RPN may propose is a real constraint, and relieving it is close to free. Adding radii below 6 Å while keeping the existing ones doubles Act residue AUPR ($0.083 \rightarrow 0.168$) and raises BindB by half ($0.093 \rightarrow 0.140$). It also raises Act proposal recall from 0.038 to 0.161. The gain does not trade against the other two tasks: BindI improves as well ($0.420 \rightarrow 0.507$) and only Motif gives up 0.030, so the macro average rises from 0.315 to 0.362. An earlier version of this sweep replaced the large radii rather than adding to them, and it did trade Act and BindB against BindI and Motif. That trade was an artifact of removing the coarse scales, not a property of adding fine ones. Block B repeats the comparison with the region Dice term the published row used, and the same pattern holds at unchanged macro (Act $+55\%$, BindB $+66\%$). That term is what BindI and Motif want: it lifts the control's macro from 0.315 to 0.383, close enough to the site checkpoint's 0.458 to show that the sweep's lower absolute level is mostly the missing Dice term and a shorter budget rather than a different model.

Second, the four tasks are in genuine conflict over training time, and the selection rule silently resolves it against Act and BindB. In a matched run at the published radii, validation Act peaks at 0.188 at the very start of training and then decays monotonically, through 0.089 to 0.046, while BindB follows it down from 0.135 through 0.102 to 0.053. BindI and Motif move the other way over the same span, rising from 0.243 to 0.434 and from 0.455 to 0.631. Act and BindB are therefore at their best before the proposal network has learned anything task-specific, and the macro criterion peaks where BindI and Motif peak, so the checkpoint it selects has already given up more than half

of the initial Act. Block C makes the cost explicit: selecting the same runs on validation Act+BindB instead roughly doubles Act again, to 0.169 for the control and 0.198 for the spanning ladder, and halves BindI and Motif.

Taken together, the two measurements make BindB a resolution problem and Act a selection one. Against the ESM-2 probe of Table 6, the spanning ladder closes most of the BindB gap, from 0.174 against 0.093 to 0.174 against 0.148, and finer radii alone carry Act from 0.083 to 0.168 without touching the other tasks. Closing the Act gap entirely takes the second change as well: selecting on validation Act+BindB rather than on the macro brings Act to 0.198, level with the probe, at the cost of half of BindI and Motif. Act is therefore reachable, but not while a macro criterion is choosing the checkpoint.

C Proposal geometry across checkpoints

We measure the normalized geometry of the pooled regions directly, on a fixed protein set common to every checkpoint, reporting an effective residue count per proposal and how concentrated the mass is.

The site warm start pools a median 6.0 effective residues, 2.7% of the chain, with 95% of the mass on ten residues. The EC function checkpoint pools 23.6 effective residues, or 11% of the chain, with 42% of the mass on ten residues and a cosine of 0.851 against a plain masked mean; the clamp is reached on none of its 4,096 proposals. The two EC continuation arms of Appendix D differ sharply. With the site loss off they give 43.2 effective residues, 37% top-ten mass, cosine 0.818 and no clamp hits. With it on they give 2.4 effective residues, 99% top-ten mass, cosine 0.488 and 2,900 clamp hits out of 4,096.

Three conclusions follow. First, the near-zero raw mass is not evidence of degenerate regions: after normalization the function checkpoint pools over tens of residues and the clamp is never reached. Second, the site checkpoint and the EC function checkpoint have qualitatively different geometry, 6.0 against 23.6 effective residues, so the site results and Table 3 do not describe the same regions. Localization quality measured on the former should not be attributed to the latter. Third, the only genuinely degenerate arm is the one that keeps the site loss active during function training. It concentrates on 2.4 effective residues, puts 99% of the mass on ten of them, and reaches the clamp on 71% of proposals, at which point the pooled summary is no longer a well-defined weighted average. This supplies a mechanism for the loss of accuracy reported in Appendix D.

D Co-training the site loss on EC

Every run reported elsewhere in this paper optimizes the site objective only before function training, never alongside it. To test the remaining option we continued training the released EC checkpoint in two arms that differ in one configuration flag: one keeps the site loss active throughout and the other does not. Both arms use the same warm start, seed, optimizer settings, epoch budget, patience and data, and both select their own checkpoint and threshold on validation. The site-supervised arm triggered early stopping well before its budget was spent, and its best checkpoint came early in the continuation; the control ran its full budget.

Table 8: Continuing EC training with and without the site loss active.

	95% identity	
Continued training	$F_{\max}\uparrow$	AUPR $\uparrow$
Site loss co-trained	0.871	0.862
Site loss off (control)	**0.881**	**0.885**

Co-training the site loss is worse on both metrics, by -0.010 in $F_{\max}$ and -0.023 in AUPR. The diagnostics identify a mechanism rather than noise. The two arms converge to different gate regimes: the mean ROI gate settles at 0.094 with the site loss active and 0.154 without it, and mean proposal objectness differs by a factor of 2.5 (0.26 versus 0.65). The geometry probe of Appendix C shows what that does to the regions. The co-trained arm concentrates on 2.4 effective residues with 99%

of the mass on ten of them, and it reaches the 10^{-8} denominator clamp on 2,900 of 4,096 proposals where its control reaches it on none. Once the clamp binds, `GraphROIPool` is no longer computing a weighted average, so the ROI token is corrupted rather than merely uninformative. Supervising objectness against residue annotations drives it toward the empirical site rate. Most residues are not annotated sites, so the masks sharpen past the point where the pooling operator is numerically well posed. On this task the objective that makes proposals faithful to annotations and the objective that makes them usable by the classifier are in direct conflict.

Two limits bound what this shows. It is a continuation of an already converged model rather than joint training from initialization, so it speaks to adding the site objective late rather than to the two-stage design as a whole. More important, it is a single label space, and EC is the one where the ROI pathway contributes nothing measurable: disabling that pathway costs 0.0001 AUPR, with a bootstrap interval spanning zero. Strengthening supervision of a pathway that is already inert may simply add optimization burden. The argument does not transfer to BP, where the same pathway is worth 0.020 AUPR. We report this as a negative result on EC, not as a general verdict on co-training.

E Implementation and provenance details

The gates of Equation 7 are initialized with negative bias so that training starts close to the global predictor and the proposal pathway has to earn its contribution; the ROI logit head is initialized near zero for the same reason.

Every reported evaluation was produced under a single frozen code revision. Training and evaluation artifacts record SHA-256 digests for label order, split membership, annotations, sequences, checkpoints, ontology files, and evaluation code, and each reported run refuses to load against a mismatched digest.

F Site data after validation and filtering

Table 9: Accepted VenusX records used by Locus.

Task	Train	Validation	Test
Act	2,034	345	366
BindI	1,701	293	296
BindB	3,807	632	646
Motif	1,810	308	335
Total	9,352	1,578	1,643

The filtering pipeline removed 5,964 training records with close homology to protected GO validation or test sequences under the original audit and removed exact split overlaps before constructing the accepted manifests.