Proteome-wide Prediction of the Functional Impact of Missense Variants with ProteoCast

Marina Abakarova[1,2], **Maria Ines Freiberger**[1], **Arnaud Lierhmann**[1], **Michael Rera**[2,*], **Elodie Laine**[1,3,*]

[1] Sorbonne Université, CNRS, IBPS, Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Paris, 75005, France
[2] Université Paris Cité, Institut Jacques Monod UMR7592, 75013 Paris, France
[3] Institut universitaire de France (IUF)

*corresponding authors: michael.rera@cnrs.fr, elodie.laine@sorbonne-universite.fr

Supplemental Methods

The GEMME algorithm relies on two main ingredients for estimating variant effects.

Evolutionary conservation

The first ingredient is a measure of evolutionary conservation that accounts for the topology of a phylogenetic tree relating the query sequence with the other sequences in the input alignment [1]. To estimate it, GEMME relies on the Joint Evolutionary Trees (JET) method [2] – implemented in the most recent version of the tool, JET2 (Laine and Carbone 2015) (available at www.lcqb.upmc.fr/JET2). The JET method subsamples the initial input set of sequences, generates a tree for each subset to estimate a conservation value for each residue in the query, and finally averages conservation values over the generated trees. Subsampling is performed using a Gibbs sampling strategy and hence, multiple runs may produce different conservation estimates. We chose to run 7 iterations of the algorithm and retain the maximum value over the 7 iterations for each residue. We tested the robustness of this procedure on a couple hundred protein sequences. By running GEMME twice on each of these sequences, we obtained highly similar pairs of predicted mutational landscapes, with Pearson correlation coefficients above 0.95. By comparison, relying on only one JET iteration per GEMME run yielded Pearson correlation coefficients sometimes as low as 0.78. Another parameter in JET controls the number of sequences from the input alignment that are used to estimate residue evolutionary conservation. By default, it is set to 20,000 sequences, a threshold that effectively allows us to retain all sequences for the vast majority (97%) of proteoforms. The conservation estimates are normalised for each protein between 0 and 1.

Fraction of observed substitutions

The second ingredient is the minimal evolutionary distance between the query sequence and a sequence in the input alignment displaying the mutation of interest. Mutations not observed in the alignment are assigned the maximum evolutionary distance. Hence, the resolution of the predicted mutational landscape depends on the sequence diversity of the input alignment. A popular measure for estimating this diversity is the number of effective sequences [3–5]. It is computed by summing up sequence-specific weights that reflect their overall similarity. However, we previously showed that this metric is not a good indicator of the quality of GEMME predictions [6]. Here, we relied instead on a per-residue measure reflecting the extent to which the 19 possible substitutions are observed in the alignment. Namely, the fraction of observed substitutions $f_{obs}(s^*, i)$ is expressed as, $f_{\text{obs}}(s^*, i) = \frac{1}{19} \sum_{a \in \mathcal{A} \setminus \{s_i^*\}} \mathbf{1}\{f_i(a) > 0\}$,

where $f_i(a)$ is the frequency of occurrence of the amino acid a at position i in the MSA.

Per-residue confidence scores

The confidence estimated for the variant predictions of a given residue in the query protein is expressed as,

$$c(s^*, i) = \begin{cases} 0, & \text{if } \sigma_{pred}(s^*, i) < 0.15 \text{ and } T_{JET}(s^*, i) < 0.2 \text{ and } score(s^*, i) < 0.1 \\ 1, & \text{otherwise} \end{cases},$$

where s^* is the query sequence, i is the index of the residue of interest, $\sigma_{pred}(s^*, i)$ is the standard deviation of the 19-long vector of predicted GEMME values, and $T_{JET}(s^*, i)$ is the conservation value. The term $score(s^*, i)$ is computed as, $score(s^*, i) = cov(s^*, i) \cdot f_{obs}(s^*, i)$, with $cov(s^*, i)$ the

proportion of sequences that have an amino acid at position i (as opposed to a gap) and $f_{obs}(s*, i)$ the fraction of observed substitutions at position i. We assessed the influence of changing the thresholds for $\sigma_{pred}(s*, i)$, $T_{JET}(s*, i)$, and $score(s*, i)$ on the number of segments N_{seg} detected as unreliable (low confidence score) and the proportion of segments of length one (isolated unreliable residues, f_{seg1}). We tested several combinations of threshold values in the intervals [0.1, 0.11, 0.12, 0.13, 0.14, 0.15] for $\sigma_{pred}(s*, i)$, [0.1, 0.15, 0.2, 0.25, 0.3] for $T_{JET}(s*, i)$ and [0.05, 0.1, 0.15, 0.2] for $score(s*, i)$ (**Fig. S9**). We obtained very similar results for the tested combinations, indicating that they are robust within these value ranges. We observed a slight tendency for fewer segments and relatively more single-residue detections upon increasing the thresholds (**Fig. S9**).

We further smooth the per-residue confidence scores over the sequence to mitigate the discontinuities that may arise from using deterministic cutoffs. The final confidence score for the i^{th} residue of the query sequence $s*$ is thus computed as,

$$confidence(s*, i) = \begin{cases} 1, & \text{if } \frac{\sum_{k=-2}^{2} c(s*, i+k) \cdot w[k]}{\sum_{k=-2}^{2} w} \geq 0.5 \\ 0, & \text{otherwise} \end{cases},$$

where w is the weight vector $\mathbf{w} = [2 \quad 3 \quad 4 \quad 3 \quad 2]$ (see **Fig. S10**).

Supplemental tables

	Number of genes	Number of unique proteoforms*	Number of proteoforms with predicted landscapes	Number of proteoforms with confident predicted landscapes	Number of residues with confident predictions
All	13,969	22,392 (1.6 ± 1.8 per gene)	22,169 (99%)	19,421 (86.7%)	13,943,646 (96.8%)
Subset with AF2-predicted 3D models	13,919	22,083	21,870 (99%)	19,212 (87%)	12,588,986 (97.2%)

Table S1: **Overview of the predictions for the *Drosophila melanogaster* proteome.** We took Flybase version 6.44 of the proteome. *data retrieved from FlyBase.

Dataset	Number of SNPs	Number of SNPs unique to dataset	Number of SNPs with confident predictions*	Number of proteins	Number of genes
DGRP	177,013	53,557	137,149	12,472	10,469
DEST2	331,341	207,886	269,766	13,227	10,879
Lethal	1,066	787	1,004	464	456
Hypomorphic	403	151	148	106	106

Table S2: Overview of the custom benchmark for evaluating ProteoCast predictions in the context of organismal fitness. The benchmark includes four datasets of missense mutations in *Drosophila melanogaster*: (1) inbred population polymorphisms derived from the DGRP; (2) natural population polymorphisms derived from the DEST2; (3) ethyl methanesulfonate(EMS)-induced point mutations resulting in lethality, as annotated in FlyBase; and (4) point mutations associated with hypomorphic alleles in FlyBase, which exhibit reduced gene function. *Mutations found in the Lethal dataset and also in DEST2 or DGRP are excluded to ensure a fairer calculation of performance metrics. Additionally, all mutations from the Lethal dataset are excluded from the Hypomorphic dataset.

PDB ID	UniProt ID	Organism	Identity (%)	Coverage (%)	E-value	Publication
4YUB	Q6XQN6	*Homo sapiens*	47.77	79	2e-86	Marletta *et al.* [7]
2F7F	Q830Y8	*Enterococcus faecalis V583*	44.66	67	3e-48	None
1YTD, 1YTK, 1YTE	Q9HJ28	*Thermoplasma acidophilum*	37.80	18	7e-14	Shin *et al.* [8]
2I14	Q8TZS9	*Pyrococcus furiosus*	27.45	60	6e-10	None
1VLP	P39683	*Saccharomyces cerevisiae*	22.36	32	0.002	Chappie *et al.* [9]

Table S3: Homology search for Naprt. We queried the Naprt-PH proteoform sequence against the Protein Data Bank with PSI-BLAST (2 iterations).

Gene	Proteoform Flybase Id	Position, Role	Length (in residue)	F_{obs} (%)	Number of polymorphisms (classified as impactful)	Number of DEST2 only (classified as impactful)	Number of DGRP only (classified as impactful)	Number of Lethal mutations (classified as neutral)
Snake	FBpp0082184	Upstream (protease)	435	93.0	28 (0)	15 (0)	2 (0)	0 (0)
SPE	FBpp0083832	Upstream (protease)	400	95.4	26 (2)	14 (2)	4 (0)	0 (0)
Spatzle	**FBpp0084507**	**ligand**	**326**	**66.5**	**16 (0)**	**8 (0)**	**5 (0)**	**2 (1)**
Toll	**FBpp0084431**	**Membrane receptor**	**1097**	**76.3**	**54 (0)**	**32 (0)**	**3 (0)**	**9 (2)**
Pellino	FBpp0083913	Intra (U-ligase)	424	64.2	7 (3)	5 (3)	2 (0)	0 (0)
Myd88	FBpp0087679	Intra (adaptor)	537	49.9	28 (2)	21 (2)	3 (0)	1 (0)
Tube	FBpp0291542	Intra (adaptor)	462	33.3	20 (4)	12 (2)	3 (0)	2 (1)
Pelle	FBpp0084549	Intra (kinase)	501	86.5	23 (1)	15 (1)	4 (0)	8 (3)
Cactus	FBpp0080402	Intra (inhibitory)	500	72.2	20 (1)	14 (1)	4 (0)	0 (0)

Dif	**FBpp0080561**	**Intra (target TF)**	**667**	**80.1**	**40 (0)**	**19 (0)**	**3 (0)**	**0 (0)**
Dorsal	**FBpp0080558**	**Intra (target TF)**	**677**	**79.5**	**20 (2)**	**8 (0)**	**7 (0)**	**6 (1)**
DEAF-1	FBpp0074651	Intra (TF)	576	67.8	13 (1)	10 (1)	3 (0)	1 (0)
Drs	FBpp0072935	Downstream	70	81.2	1 (0)	0 (0)	1 (0)	0 (0)
Mtk	FBpp0086518	Downstream	52	6.3	5 (1)	2 (1)	1 (0)	0 (0)
Def	FBpp0087518	Downstream	92	64	10 (0)	4 (0)	0 (0)	0 (0)
CecC	FBpp0084980	Downstream	63	36.6	5 (0)	3 (0)	0 (0)	0 (0)
AttA	FBpp0086567	Downstream	221	41.7	18 (0)	5 (0)	4 (0)	0 (0)
Dipt	FBpp0085802	Downstream	106	43.5	10 (1)	4 (1)	0 (0)	0 (0)

Table S4: Overview of the polymorphisms and referenced lethal mutations in 18 proteins from the Toll pathway. The main components, namely the Toll receptor and its ligands, and the target transcription factors (TF), are highlighted in bold. F_{obs} is the fraction of observed mutations.

Supplemental figures

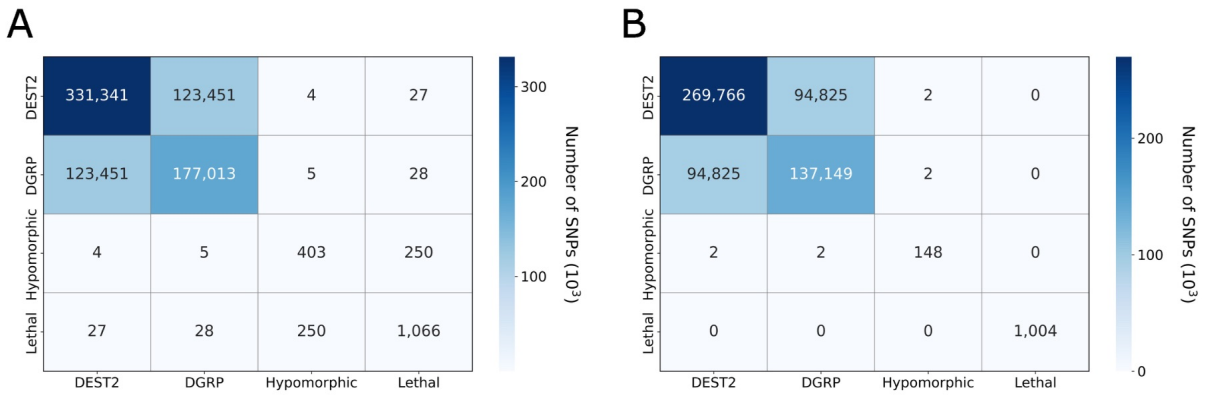

Figure S1: Overview of the custom benchmark for evaluating ProteoCast predictions in the context of organismal fitness. A. All mutations. **B.** Mutations whose effect is predicted with high global and local confidence and that are not annotated as both lethal and observed in fly populations (DEST2 or DGRP). Additionally, mutations referenced as lethal are excluded from the Hypomorphic dataset.

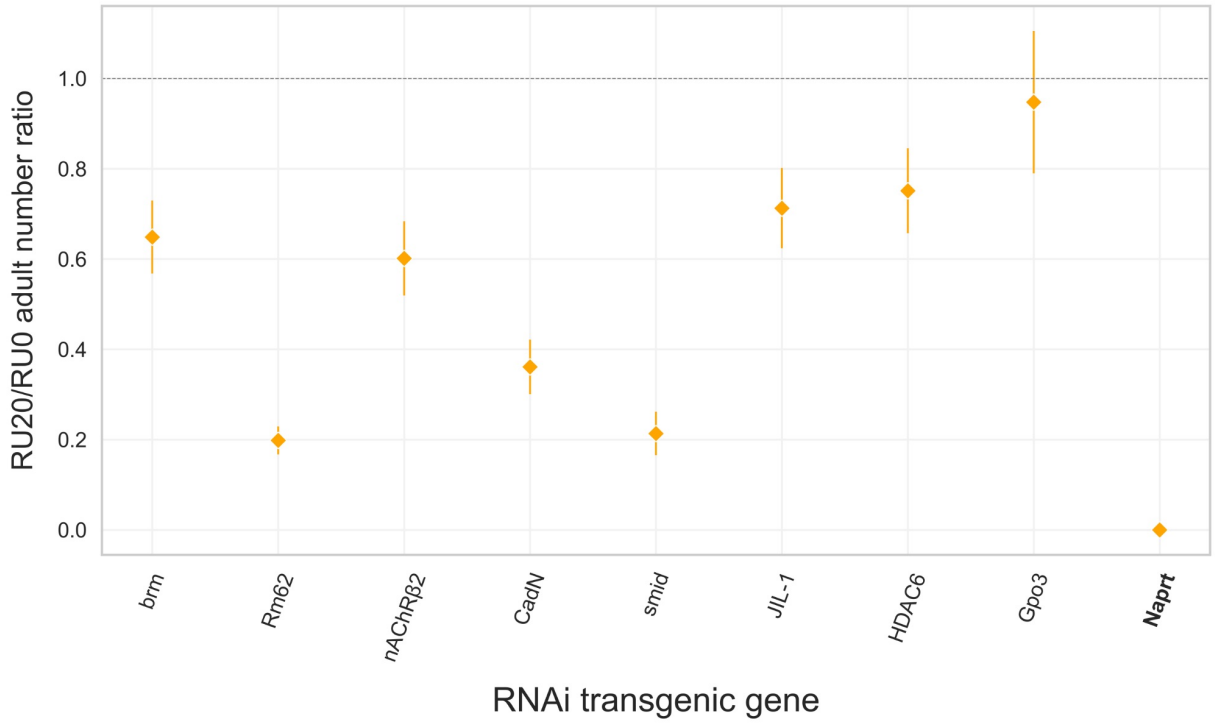

Figure S2: **Developmental lethality associated with gene-specific RNAi knockdowns.** Ratio of *da::GS*>UAS-RNAi individuals developing out of two conditions RU0 (µg/mL of RU486) - negative control, the RNAi is not induced -, and RU20 - the RNAi is expressed at the maximum minducer concentration for that system. We present here the ratio of the total number of adults eclosing from the induced condition divided by the total number of adults eclosing from the control condition. List of genes targeted with RNAi in the experiment: *nAChRβ2* (FBgn0004118), *CadN* (FBgn0015609), ***Naprt*** (FBgn0031589), *Gpo3* (FBgn0028848), *JIL-1* (FBgn0020412), *mrj* (FBgn0034091), *smid* (FBgn0016983), *Rm62* (FBgn0003261).

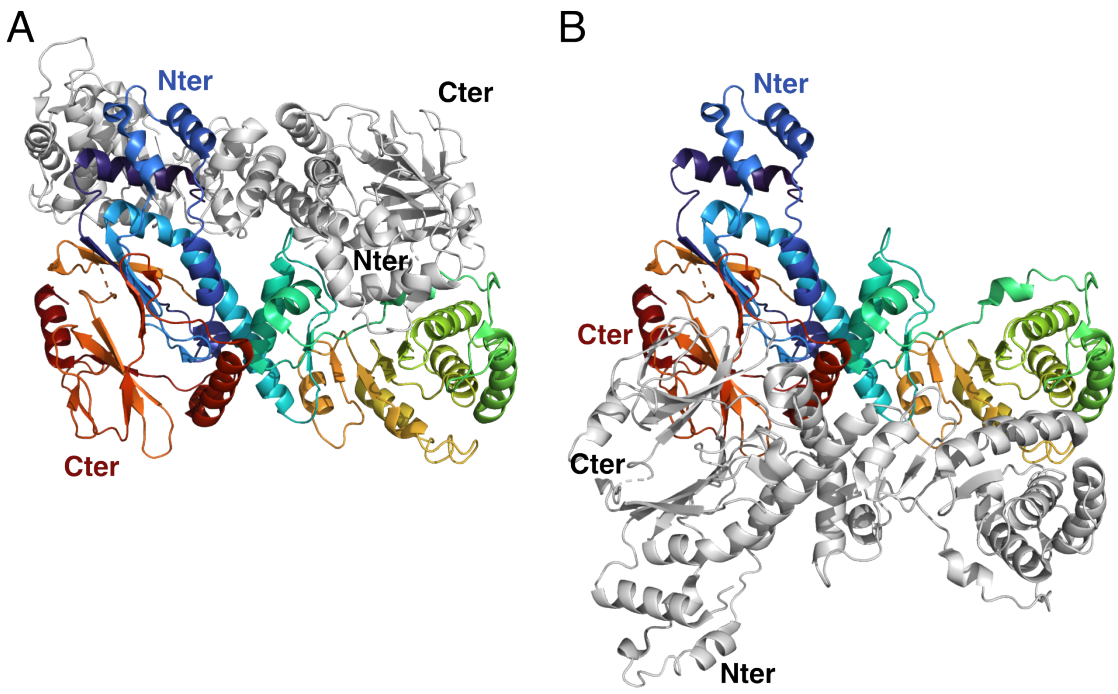

Figure S3: Human NAPRT homo-dimer. The coordinates were retrieved from the PDB entry 4YUB. One monomer is taken as reference and colored in rainbow according to the residue index in the protein sequence. The second monomer is colored in white. **A.** Biologically relevant head-to-tail homo-dimeric arrangement of human NAPRT described in Marletta and co-authors (see Fig. 2 in [7]). **B.** Head-to-head dimeric arrangement incorrectly defined as the biological assembly and asymmetric unit in the Protein Data Bank.

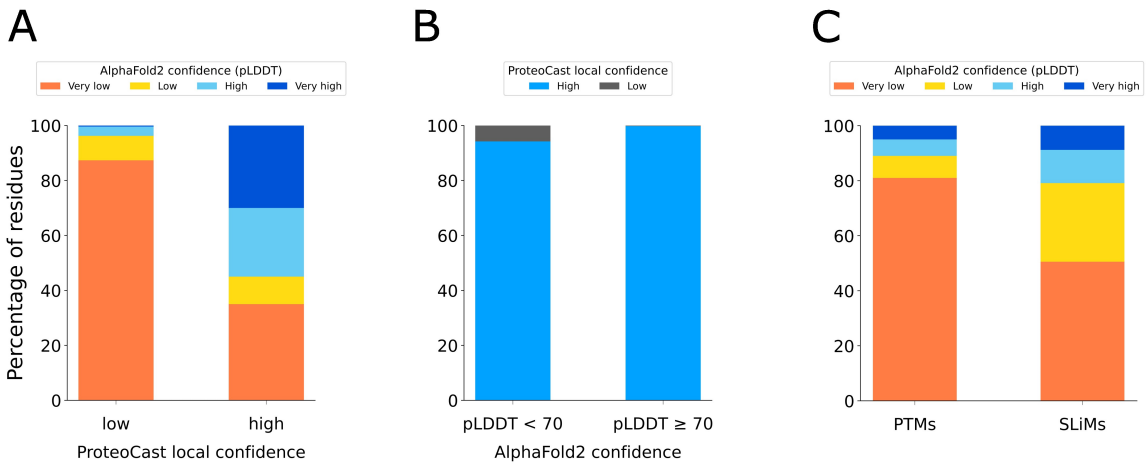

Figure S4: **Relationship between AlphaFold2 pLDDT confidence and ProteoCast local confidence. A.** Proportion of residues in different AlphaFold2 pLDDT categories (very low, low, high, very high), reflecting confidence in their predicted 3D coordinates, shown separately for low and high ProteoCast local confidence. **B.** Proportion of residues classified as high or low confidence by ProteoCast, grouped by AlphaFold2 pLDDT confidence thresholds (pLDDT < 70 vs. pLDDT ≥ 70). **C. Distribution of AlphaFold2 pLDDT scores for PTMs and SLiMs.** In total, there are ~60K PTMs and 91 SLiMs.

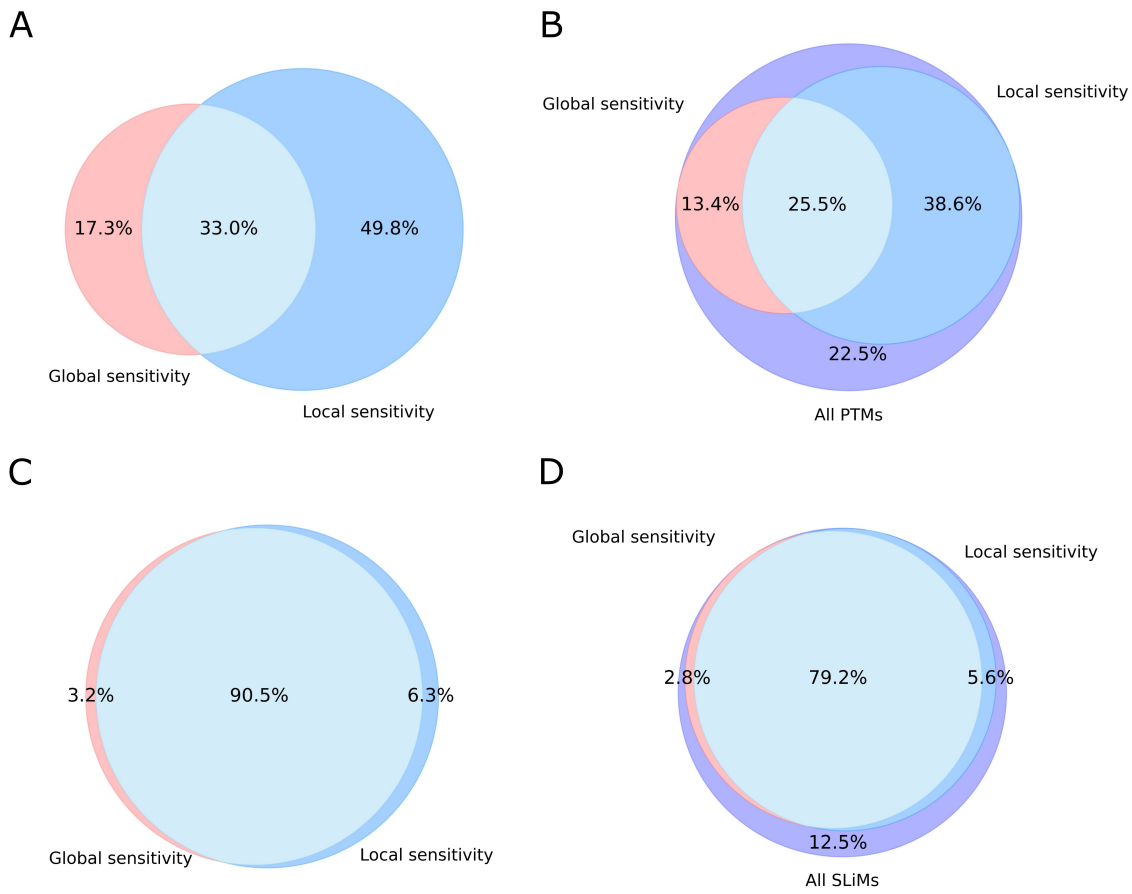

Figure S5: ProteoCast local and global mutational sensitivity-based detection of post-translational modification sites (A-B) and short linear motifs (C-D). We focus on regions with low AlphaFold2 pLDDT score (below 70). Glocal sensitivity refers to ProteoCast residue classification where sensitive residues have less than 10 neutral substitutions. Local sensitivity refers to segments whose mutational sensitivity stands out from their surroundings**.** The Venn diagrams show the complementarity between the two approaches. **A-B.** Detection of post-translational modification (PTM) sites, considering the subset of detected sites (A) or all sites (B). **C-D.** Detection of short linear motifs (SLiMS), considering the subset of detected motifs (C) or all motifs (D).

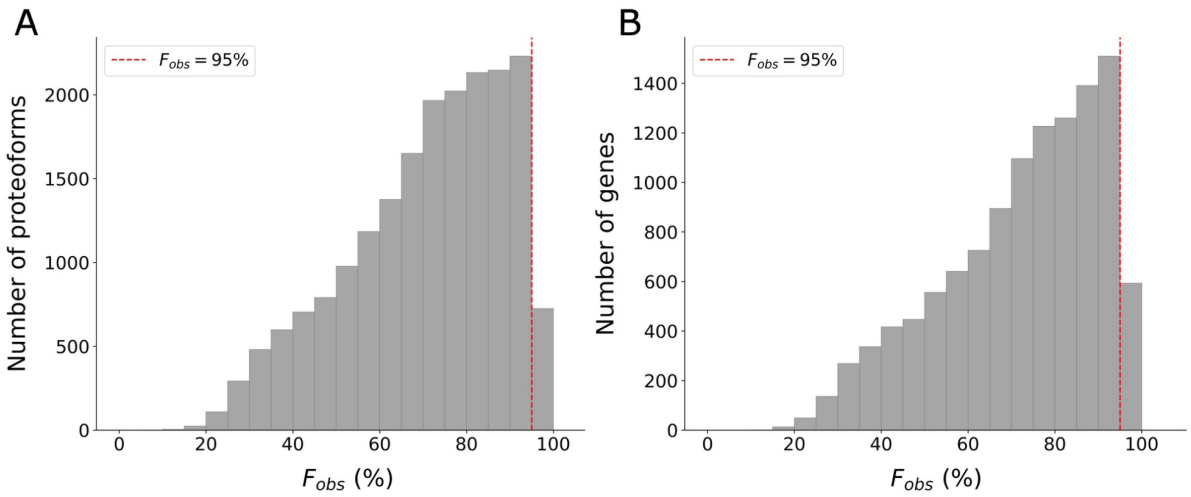

Figure S6: **Distribution of fraction of observed mutations, F_{obs}. A.** By unique proteoform. **B.** By gene, taking the maximum value over the proteoforms of a given gene. We retained only the 19,421 proteoforms, coming from 11,564 genes, with high global confidence predictions.

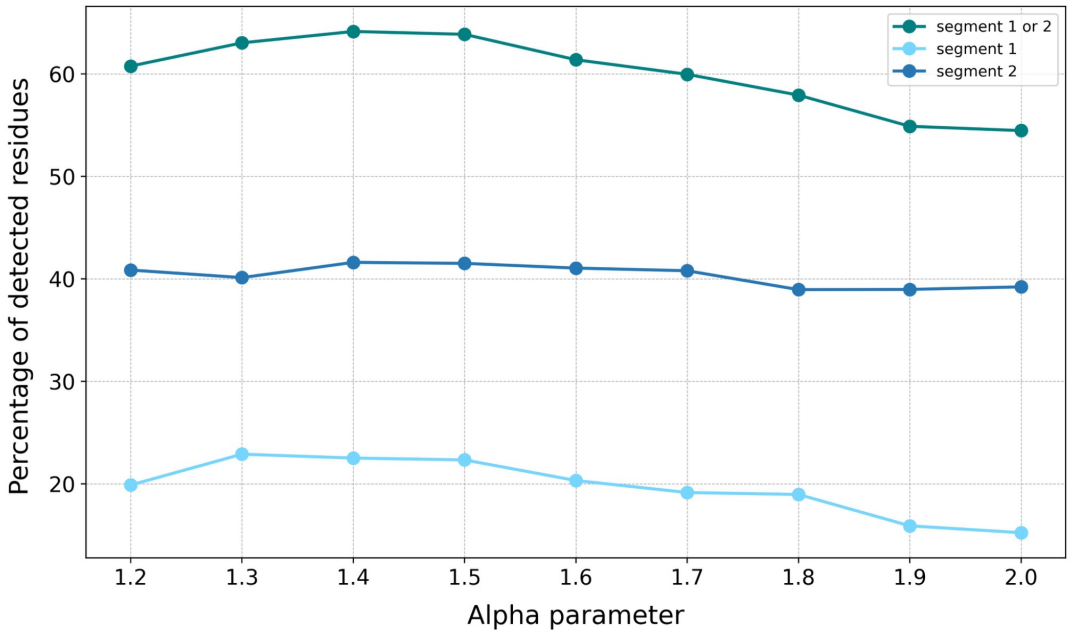

Figure S7: Assessment of PTM detection in segmented regions as a function of the hyperparameter α. The fraction of detected post-translational modifications (PTMs) within the segmented regions is plotted for segment 1, segment 2, or both. The analysis considers a subset of PTMs from confident proteoforms with an available 3D structure, focusing on regions with pLDDT < 70.

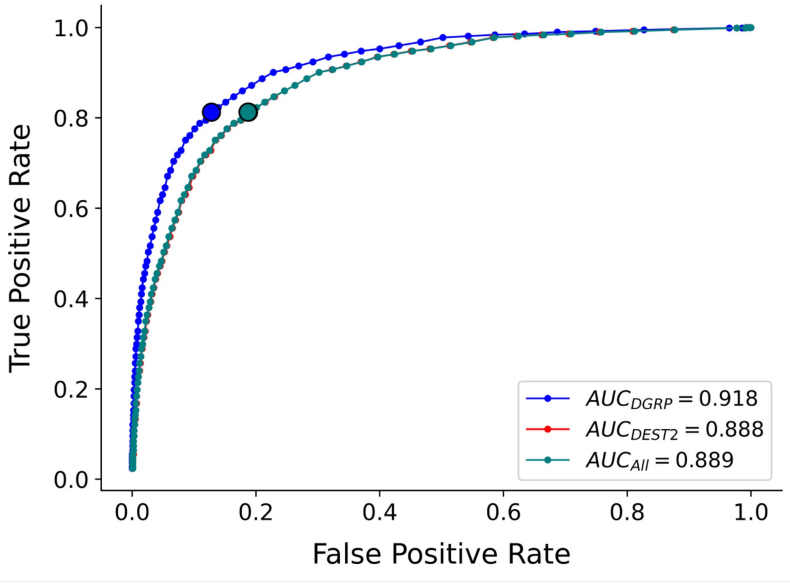

Figure S8: Receiver Operating Characteristic (ROC) curves comparing predictive performance based on ProteoCast raw scores. The True Positive Rate (sensitivity) is plotted against the False Positive Rate for three datasets differing in their selection of negative examples: DGRP-only (blue), DEST2-only (red), and All (teal), which includes the overlap between DEST2 and DGRP. The Area Under the Curve (AUC) values are provided in the legend. Large markers indicate thresholds with optimal performance. The DEST2 and All datasets exhibit nearly identical performance.

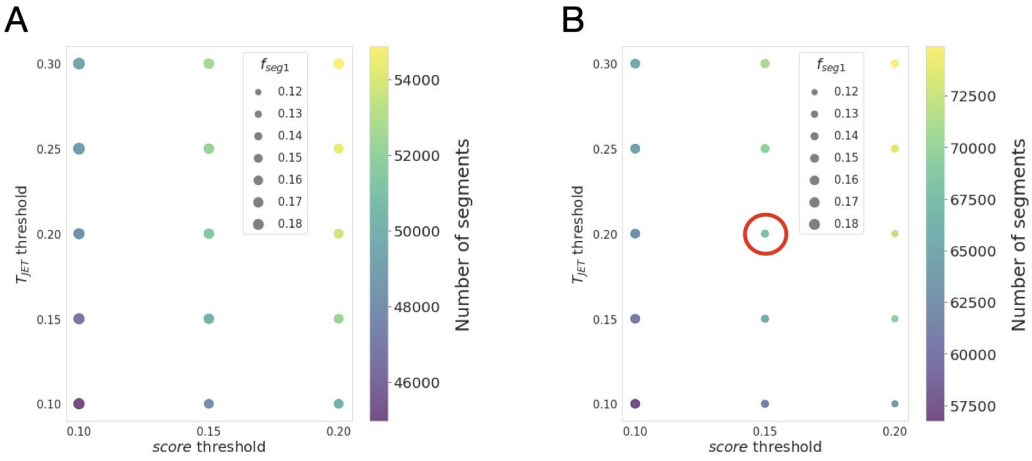

Figure S9: Assessment of different threshold configurations for detecting residues with low local confidence. Dot plots illustrate the impact of varying $score$ and T_{JET} thresholds on the number of segments (depicted by a color scale) and the fraction of segments of size one, f_{seg1}, with (A) $\sigma_{pred}(s^*, i) = 0.1$ and (B) $\sigma_{pred}(s^*, i) = 0.15$. The red circle highlights the selected threshold combination.

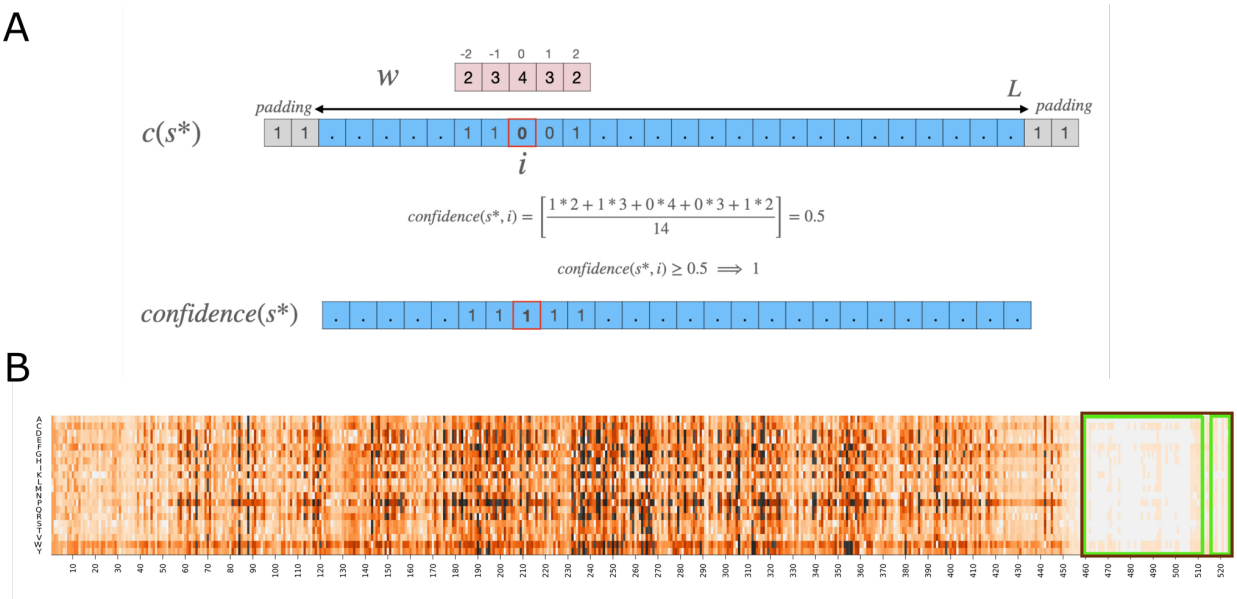

Figure S10: Overview of the smoothing process for identifying low-confidence residues. **A.** The smoothing operation involves performing a convolution over all $c(s^*)$ values in a sequence with a weight vector *W*. The sequence is scanned with overlapping windows and, within each window, the central residue receives the highest weight, while its neighbours receive progressively lower weights. The confidence values $confidence(s^*)$ are obtained by binarising the smoothed scores. **B.** The mutational landscape of the *wdb-PA* (FBpp0084577) protein. The smoothing allows the merging of two nearby low-confidence regions (in green).

References

1. Lichtarge O, Bourne HR, Cohen FE. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. J. Mol. Biol. 1996;257:342–58.
2. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. PLOS Comput. Biol. 2009;5:e1000267.
3. Notin P, Niekerk LV, Kollasch AW, Ritter D, Gal Y, Marks DS. TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. 2022; https://doi.org/10.1101/2022.12.07.519495.
4. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 2018;15:816–22.
5. Notin P, Kollasch A, Ritter D, van Niekerk L, Paul S, Spinner H, et al. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. Adv. Neural Inf. Process. Syst. 2023;36:64331–79.
6. Abakarova M, Marquet C, Rera M, Rost B, Laine E. Alignment-based Protein Mutational Landscape Prediction: Doing More with Less. Genome Biol. Evol. 2023;15:evad201.
7. Marletta AS, Massarotti A, Orsomando G, Magni G, Rizzi M, Garavaglia S. Crystal structure of human nicotinic acid phosphoribosyltransferase. FEBS Open Bio 2015;5:419–28.
8. Shin DH, Oganesyan N, Jancarik J, Yokota H, Kim R, Kim SH. Crystal Structure of a Nicotinate Phosphoribosyltransferase from Thermoplasma acidophilum*. J. Biol. Chem. 2005;280:18326–35.
9. Chappie JS, Cànaves JM, Han GW, Rife CL, Xu Q, Stevens RC. The Structure of a Eukaryotic Nicotinic Acid Phosphoribosyltransferase Reveals Structural Heterogeneity among Type II PRTases. Structure 2005;13:1385–96.