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

Abstract

Background:
Dissecting the functional impact of genetic mutations is essential to advancing our understanding of genotype-phenotype relationships and identifying new therapeutic targets. Despite the progress in sequencing and CRISPR technologies, proteome-wide mutation effect prediction remains challenging. Here, we introduce ProteoCast, a scalable and interpretable computational method for proteome-wide classification of genetic variants and functional protein site identification. It relies solely on evolutionary information, leveraging protein sequence data across organisms.
Results:
Using ProteoCast, we generated mutational landscapes for 22,169 *Drosophila melanogaster* protein isoforms, categorising over 293 million amino acid substitutions as functionally neutral, uncertain, or impactful. We validated our predictions with over 380 thousand natural polymorphisms observed in the Drosophila Genetic Reference Panel (DGRP) and Drosophila Evolution over Space and Time (DEST) datasets and with FlyBase's developmentally lethal mutations. About 86% of known lethal mutations were classified as impactful or uncertain, versus only 13% and 18% of DGRP and DEST mutations. Moreover, we performed ProteoCast-guided genome editing experiments, providing a proof-of-concept of the validity of this strategy. Beyond variant effect prediction, ProteoCast detected evolutionary conservation signals in about one-third of 40.5K annotated post-translational modification sites and 83% of ~90 known short linear motifs. These results support its usefulness for uncovering interaction and regulatory sites in unstructured protein regions.
Conclusions:
Our results demonstrate ProteoCast applicability for model organisms, contributing to basic genetic research and translational studies. This work provides a publicly available dataset, user-friendly interactive web services, and a locally deployable pipeline tool for further research into gene function and mutation effects in any organism.

Keywords
Protein mutation, variant effect prediction, developmental lethality, protein function, *Drosophila melanogaster*, functional sites discovery

Background

More than 50,000 single nucleotide mutations are associated with genetic disorders in humans [1], and correcting them at the genomic level could provide curative treatments for various diseases [2]. The advent of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology has revolutionized genetics by enabling precise and efficient DNA editing as well as establishing targeted mutational screens as a powerful forward genetic approach to identify novel druggable targets [3]. When combined with high-throughput sequencing, which enables the analysis of inter-individual and inter-tissue genomic variability, these single-point genome editing tools allow unprecedented exploration of the functional significance of genetic variants. They provide effective means for assessing their contributions to disease susceptibility, drug responses and, more generally, to complex traits.

However, the vast number of possible substitutions makes it challenging to select the optimal amino acid target(s) for genome editing. To prioritise single-point genetic modifications for CRISPR experiments, researchers often use conservation scores derived from multiple sequence alignments (MSAs), whole-exome sequencing to identify patient-specific mutations, or statistical tools like Genome-Wide Association Studies (GWAS) to link variants to specific diseases or phenotypes. While invaluable, these methods provide only a limited view of impactful mutations across the entire proteome. Computational predictive models can address these limitations, offering a more comprehensive approach to understanding the functional impact of mutations.

In the last decades, we have witnessed a profusion of protein variant effect predictors leveraging protein sequence data across species using heuristic algorithms [4–7], family-specific deep learning models [8,9], or universal protein language models (pLMs) [10–13]. Additionally, some predictors are trained on variants with known phenotypes or polymorphisms observed in the population [11,14]. These population-based predictors, however, do not typically perform better than population-free methods and tend to suffer from circularity issues and pervasive ancestry biases [15]. Furthermore, a new generation of predictors [11,16,17] systematically integrates high-quality 3D structural models [18]. However, they have shown limited improvement over purely sequence-based methods [19], and the structural contribution is sometimes unclear.

Researchers have actively engaged in efforts to assess the comparative performance of these tools by performing comprehensive evaluations using vast datasets of deep mutational scanning (DMS) experiments and clinical annotations from experts [19–21]. The results highlighted GEMME as one of the top-performing models. GEMME is a fully interpretable, unsupervised tool that uses evolutionary information from aligned protein sequences to generate a complete mutational landscape – namely, the predicted effects of all possible substitutions at every position within a protein. By focusing on how protein residues segregate along the topology of evolutionary trees, GEMME has proven instrumental in studying protein stability, function, and

disease mechanisms [22–25]. Notably, GEMME is fast: when paired with the ColabFold protocol [26,27], which provides cost-effective, small, and diverse MSAs, it enables proteome-wide predictions within just a few days.

In this study, we present a proteome-wide validation of GEMME predictions using *Drosophila melanogaster.* We propose to use these predictions for prioritising single-point genetic modifications. For this, we have developed a fully automated computational pipeline named ProteoCast, which estimates the impact of variants on protein function, determines which residues are sensitive to mutations, maps this information on protein 3D structures, and identifies putative binding and regulatory sites in unstructured regions (**Fig. 1**). Additionally, ProteoCast provides confidence metrics to evaluate the quality of its predictions and help users make informed decisions. We thoroughly tested ProteoCast on a benchmark dataset of over 386,000 single-point mutations we compiled for *Drosophila melanogaster* using genetic variation under natural or artificial selection and referenced hypomorph and lethal mutations.

While much genetic research focuses on humans, we selected *Drosophila melanogaster* as our model organism due to its rich and well-documented genetic and phenotypic data accumulated over a century of research. This extensive body of knowledge offers unique insights into natural genetic variability and the functional consequences of mutations. We used three key public resources: the Drosophila Genetic Reference Panel (DGRP) [28], a collection of 205 inbred lines, the community-generated population genomic resource Drosophila Evolution over Space and Time (DEST v2.0) [29], a collection of 530 high-quality pooled libraries from >32,000 flies collected across six continents over more than a decade, and FlyBase [30], a comprehensive database on Drosophila genetics that has aggregated the phenotypic effects of point mutations for over 30 years.

With approximately 75% of human genetic diseases being associated with genes conserved in fruit flies, the relevance of Drosophila as a model organism is obvious. Our findings, therefore, are advancing our understanding and ability to intervene in functional genomics, with a broad applicability to human health and diseases. ProteoCast and all generated and analysed data are freely available to the community at https://proteocast.ijm.fr/drosophiladb/. ProteoCast can be run on any protein sequence from any organism using the Docker image at https://hub.docker.com/r/marinaabakarova/proteocast/.

Results

We used ProteoCast to generate full single-site mutational landscapes over the entire *D. melanogaster* proteome, including all referenced protein isoforms (proteoforms) resulting from alternative splicing (**Fig. 1A,** see also ***Materials and Methods***). We estimated the functional impact of every possible substitution at every position for 22,169 unique proteoforms, covering 99% of the Drosophila proteome (**Table S1**). We provide predictions of variant effects for about 293 million missense mutations enriched with detailed information to help guide users in interpreting predictions more effectively. These metadata include confidence scores, variant functional classification as neutral, uncertain, or impactful, as well as each residue's sensitivity to mutations. Our online database https://proteocast.ijm.fr/drosophiladb/ allows for querying any Drosophila proteoform and interactively exploring the associated ProteoCast predictions. Below,

we provide a graphical summary of the prediction pipeline (**Fig. 1**) and functional validation of the predicted functional impacts.

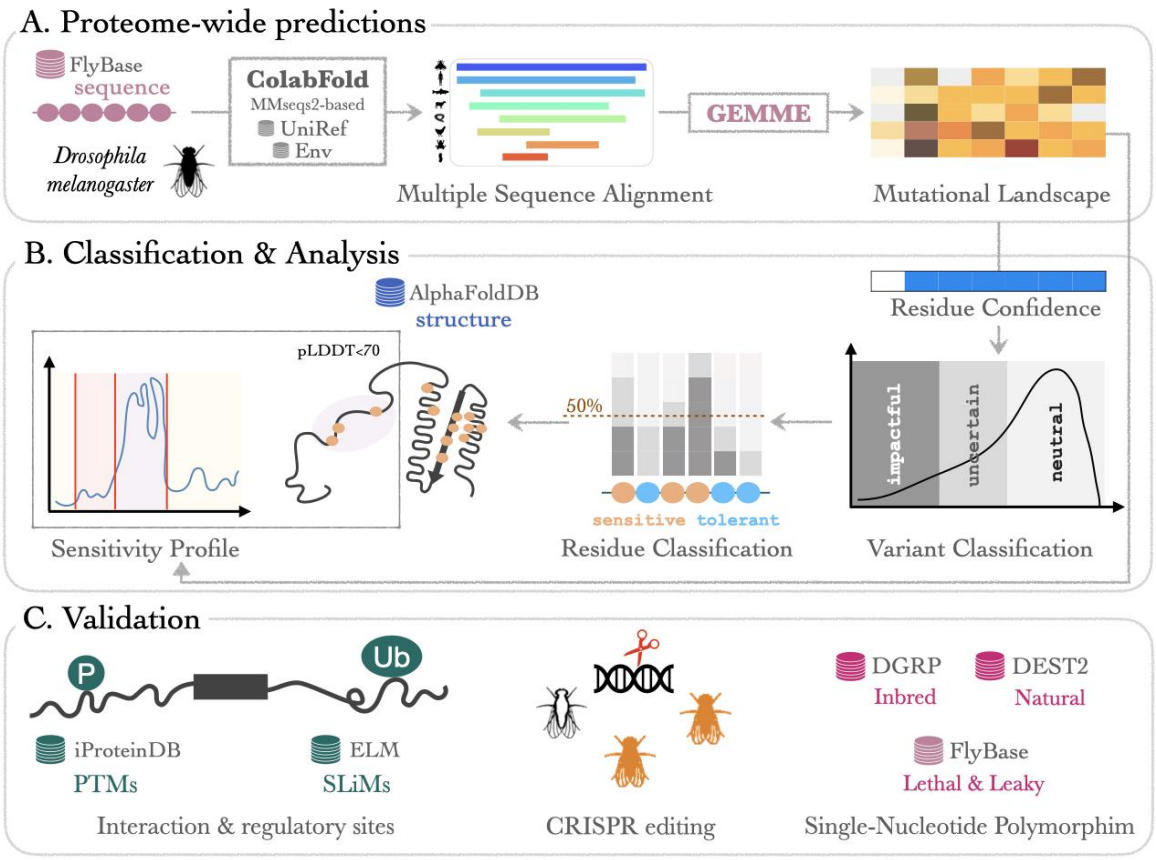

Fig. 1. Schematic representation of ProteoCast workflow. A. The workflow starts by generating a Multiple Sequence Alignment for each *Drosophila melanogaster* proteoform sequence using the ColabFold MMseqs2-based pipeline against UniRef and environmental sequence databases. GEMME then produces an Lx19 matrix representing the proteoform complete single-site Mutational Landscape, with L the number of amino acid residues. **B.** ProteoCast filters out unreliable predictions (Residue Confidence) and then classifies the remaining variants as impactful, neutral, or uncertain (Variant Classification) by fitting a Gaussian mixture model to the raw score distribution. It then determines whether each residue is sensitive or tolerant to mutations depending on its proportion of impactful variants (Residue Classification). Proteocast further maps per-residue mutational sensitivity (average raw score) to 3D models from AlphaFoldDB. It segments the mutational sensitivity profile to identify potential binding and regulatory sites in unstructured regions. **C.** As a proof-of-concept, we performed ProteoCast-guided CRISPR editing experiments. Additionally, we assessed ProteoCast predictions against known post-translational modification (PTM) sites and short linear motifs (SLiMs), as well as single-nucleotide polymorphism, either observed in inbred or wild populations, or associated with deleterious outcome such as developmental lethality or partial loss of function.

Most of the variants predicted with high confidence

Although our variant effect predictor GEMME has demonstrated strong robustness even with low alignment depth [2,19], its accuracy still depends on the quality of the input alignment. To help users assess the reliability of the predictions, ProteoCast provides binary confidence values at both the global (per-protein) and local (per-residue) levels (**Fig. 1B,** *Residue*

Confidence). We consider predictions as unreliable when evolutionary information is scarce. At the protein level, less than a couple hundred sequences in the input alignment is typically insufficient [26]. ProteoCast filtered out about 12% of all unique proteoforms from *D. melanogaster* according to this criterion (**Fig. S1**). At the residue level, ProteoCast evaluates confidence based on the dispersion of GEMME raw scores, the estimated conservation, and the amount of local information in the alignment used for input (see ***Materials and Methods*** and **Fig. S1,** *insert*). It detected unreliable predictions for ~460,000 residues (~3%) in *D. melanogaster* proteoforms (**Fig. S1**). These residues were poorly conserved, with a high proportion of gaps or a limited diversity of amino acids across species. Overall, the predictions labelled as unreliable based on the global or local confidence metrics concerned less than 10% of all residues, underlying the quality and abundance of the evolutionary information available for the fly proteome.

High correlation between predicted effects and organismal fitness

ProteoCast predicts a protein's complete single mutational landscape based solely on evolutionary information specific to that protein, without considering its biological function nor interactions with other proteins. To evaluate the relevance of these protein-centered predictions for organismal fitness, we created a custom benchmark of missense mutations resulting from single nucleotide polymorphisms (**Fig. 1C,** *Single-Nucleotide Polymorphism*). This benchmark comprises all inbred missense DGRP polymorphisms, all natural missense DEST2 polymorphism as well as all ethyl methanesulfonate (EMS) point mutations referenced in Flybase. We used Flybase's experimentally supported annotations labeled as hypomorphic – indicating reduced gene function – or as developmentally lethal in flies (**Table S2**). The benchmark is highly imbalanced, comprising approximately ~176k DGRP and ~286k DEST2 missense mutations, ~1k lethal mutations, and 139 hypomorphic non-lethal ones.

To illustrate ProteoCast analysis, we consider the Yorkie protein, a key component of the Hippo pathway involved in developmental processes including apoptosis and cell proliferation [31]. It comprises 6 missense mutations experimentally associated with developmental lethality as well as 12 inter-individual polymorphisms (**Fig. 2A-B**). We focus on two lethal mutations, namely P65L, Y259N, and two mutations observed in the DEST2, namely N169T and Q177P. Visually, the two groups of mutations are clearly set apart in the predicted mutational landscape (**Fig. 2A**). GEMME raw scores for the lethal mutations are strongly negative, all ranking in the top 5%, while the score for the viable DGRP mutation places it in the bottom 30%.

To systematically and automatically segregate variants, ProteoCast applies a thresholding strategy tailored to each protein by fitting a mixture of three Gaussians to GEMME raw score distributions using the expectation-maximization algorithm (**Fig. 1B,** *Variant Classification*, **Fig. 2B,** and see ***Materials and Methods***). We exclude unreliable predictions from this analysis to reduce the risk of bias they might introduce. At one end of the spectrum, mutations with scores close to zero extending down to the median of the middle Gaussian are classified as ***neutral*** (**Fig. 2B,** blue arrow). At the other end, mutations with strongly negative scores are classified as ***impactful*** up to the intersecting point between the leftmost and middle Gaussians (**Fig. 2B**, red arrow). Mutations falling in between are categorised as ***uncertain*** **(Fig. 2B,** purple arrow). Equipped with this classification procedure, ProteoCast identified the two lethal Yorkie mutations P65L, Y259N, as impactful and the viable mutations N169T, Q177P as neutral (**Fig. 2B**). This

high discriminative capability generalised to the full set of 18 mutations annotated in our benchmark for Yorkie. Indeed, ProteoCast classified all 6 known lethal mutations as impactful, the 2 inbred DGRP mutations and 7 out of the 10 DEST2 mutations – by definition, present in living adults – as neutral, and the remaining 3 DEST2 mutations as uncertain.

More broadly, ProteoCast classified 85% of all lethal mutations in our benchmark as impactful (2 thirds) or uncertain (1 third), 88% of the DGRP and 82% of the DEST2 missense mutations as neutral (**Fig. 2C**). The hypomorphic mutations are somewhat intermediate with still the major part (70.5%) classified as impactful or uncertain. Taken together, these results show that variant effects inferred from the evolutionary history of individual proteins correlate well with experimentally characterized organismal fitness. The vast majority of mutations observed in viable organisms, such as the DGRP lines and DEST pooled libraries, are not strongly counter-selected in evolution, while higher severity tends to be associated with higher evolutionary pressure detectable by our approach. Additionally, mutations with lower Minor Allele Frequency (MAF) tend to be assigned more negative raw scores and Z-scores (**Fig. S3**). The use of GMM helped to adapt our variant classification to the shape of each protein's score distribution and enabled across-protein evaluation of its agreement with existing functional annotations. ProteoCast variant classification strikes a better balance between recall over lethal mutations and specificity over inbred or natural polymorphisms than binary classifications based on universal thresholds (**Fig. S4** and **Table S3**).

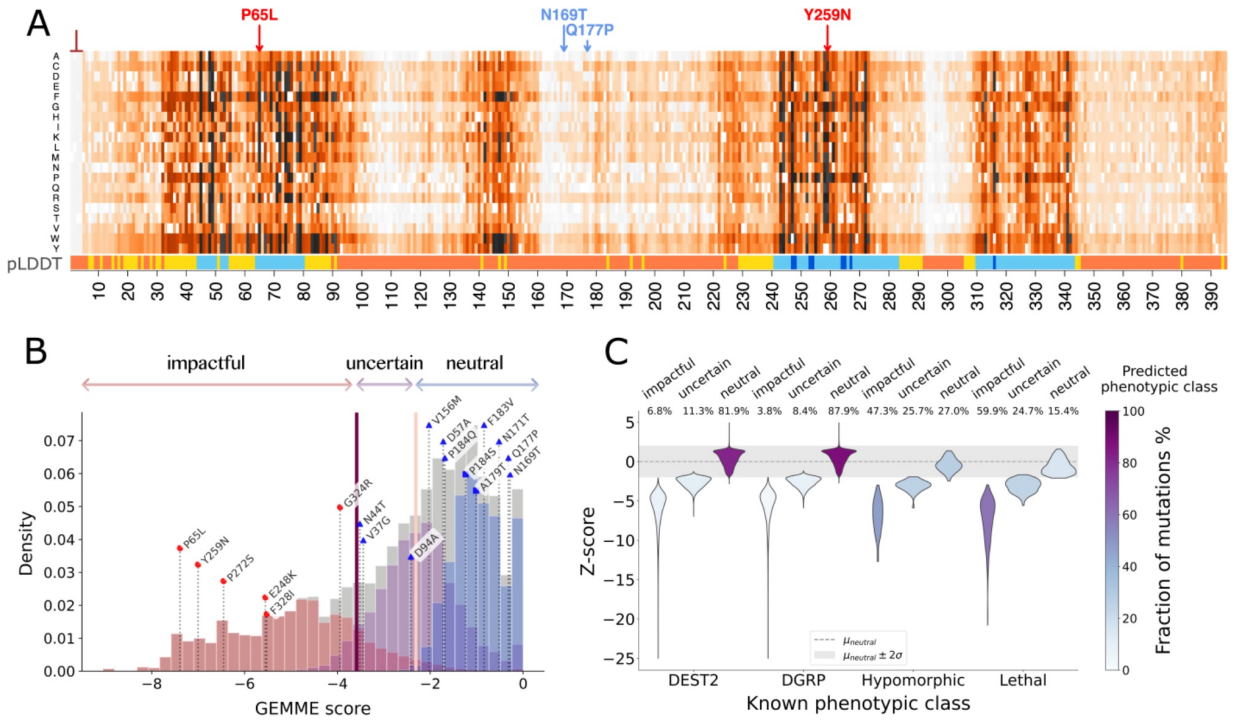

Figure 2: Variant classification using ProteoCast. A. Full single-mutational landscape computed for Yorkie protein (*yki* gene, Flybase proteoform id: FBpp0288697), a key downstream effector in the *Hippo*/*SWH* (*Sav*/*Wts*/*Hpo*) signaling pathway. Each cell of the matrix contains a numerical estimate of the functional impact of an amino acid substitution (y-axis) at a given residue position (x-axis) on the protein function. Darker colors indicate stronger effects, and white cells correspond to the wild-type amino acids. The T-shaped symbol at the N-terminus marks low-

confidence predictions. Two mutations labelled in red (*P65L*, and *Y259N*) provoke developmental lethality – as referenced in FlyBase, and two in blue (*N169T*, and *Q177P*) occur in DEST2. **B.** ProteoCast classification of Yorkie mutations based on fitting a mixture of three gaussians to the distribution of estimated effects. The pale pink vertical line represents the threshold above which mutations are considered *neutral*, the deep purple vertical line marks the threshold below which mutations are considered as *impactful*. Mutations with scores between the two threshold values are considered *uncertain*. ProteoCast accurately classified the four mutations labelled in A. **C.** Distributions of Z-scores for DGRP and DEST polymorphisms, as well as hypomorphic and lethal mutations annotated in FlyBase, depending on whether they are classified as *impactful*, *uncertain* or *neutral*. We computed the Z-scores with respect to the distribution of estimated effects for the *neutral* class within each protein. Three mutations with scores below -25 were excluded from the plot for graphical representation clarity.

ProteoCast effective in guiding genome editing strategies

As a proof-of-concept, we used ProteoCast predictions to prioritize variants for experimental genome editing in *D. melanogaster*, introducing single-point missense mutations (**Fig. 1C,** *CRISPR editing*). First, we selected nine candidate genes with high-quality input alignments, consisting of 10,000 to 30,000 sequences each. These genes were chosen based on the absence of lethal missense mutations reported in FlyBase and a broad range of predicted raw scores, including very negative values (< -8), which indicate strong effects (see ***Materials and Methods***). We tested the developmental lethality of RNAi-induced knockdowns for these genes using transgenics under control of the ubiquitous *daughterless::GeneSwitch* inducible driver. The RNAi-mediated knockdown of *Naprt* (CG3714), which encodes nicotinate phosphoribosyltransferase, an enzyme involved in NAD biosynthesis, resulted in 100% developmental lethality, identifying it as an essential gene during development (**Fig. S5**).

These results led us to prioritize *Naprt* for genome editing. Among the five unique proteoforms annotated in FlyBase for this gene, we chose Naprt-PH (662 residues) as our reference. It is the longest proteoform cross-referenced in UniProt. Based on ProteoCast predictions, we selected five variants (**Table 1**), including the two predicted to be the most impactful (most negative raw scores), namely G497P and K532G. The corresponding wild-type residues are located near the putative active sites, as inferred by homology transfer (**Fig. 3** and **Table S4**). To choose the third variant, we ordered residues according to their sensitivity to mutations, estimated by averaging their 19 raw predicted scores. We selected a residue among the 25 most sensitive ones (**Table 1**) and potentially involved in the protein's homo-dimeric interface, as inferred by homology from human Naprt 3D arrangement (**Fig. 3D**). Specifically, we chose E147 and substituted it with the amino acid predicted to have the strongest effect, namely Valine (V).

As negative controls, we looked for residues predicted to be tolerant to mutations, with more than 10 out of the 19 possible substitutions predicted as neutral (**Fig. 1B**, *Residue Classification*). To ensure a rigorous test of ProteoCast's predictions, we excluded the N-terminal extremity (residues 1-21) and an inter-domain linker (residues 301-407). These regions are poorly conserved and unstructured, and they comprise the vast majority (17 out of 20) of the naturally occurring missense mutations referenced in DGRP and DEST2. We focused on the five most tolerant residues in the structured protein fraction. The top-4 residues form a cluster in 3D space (**Fig. 3B**), from which we randomly selected a representative residue, A262. Additionally, we retained I618, the fifth-ranked residue, located in a solvent-exposed helix. To challenge the robustness of ProteoCast's neutral predictions, we substituted A262 and I618 with amino acids that exhibit starkly contrasting physico-chemical properties (**Table 1**). Specifically,

we mutated the small A262 into Phenylalanine (F), a large aromatic amino acid, and I618 into Proline (P), which has the tendency to break secondary structures. This design allowed us to test whether ProteoCast could correctly predict neutral effects despite substantial changes in amino acid properties.

Naprt Variant	Variant class	Variant raw score	Variant rank overall*	Variant rank for the residue	Residue class	Residue rank*
G497P	Impactful	-8.75	1 ↑	1 ↑	Sensitive	1 ↑
K532G	Impactful	-8.73	2 ↑	1 ↑	Sensitive	11 ↑
E147V	Impactful	-8.33	52 ↑	1 ↑	Sensitive	21 ↑
I618P	Neutral	-1.54	3,322 ↓	1 ↑	Tolerant	5 ↓
A262F	Neutral	-0.84	1,670 ↓	3 ↑	Tolerant	2 ↓

Table 1: Residues selected for CRISPR introduction of single amino acid missense mutations. For each selected variant, we report its class (among impactful, uncertain and neutral), its raw predicted score and associated overall and relative ranks. In addition, we report the mutational sensitivity class (either tolerant or sensitive) and rank of the corresponding wild-type residue.

* Residues with pLDDT<70 are excluded from the ranking. This leaves us with 533 residues and 10,127 possible variants. The arrows indicate how the values are ordered to compute the ranks. For instance, a variant ranked 1 with an up arrow has the lowest (most negative) raw score. When some values are equal, they are assigned the same (minimum) rank. Sequencing validations of the five mutants are presented in Annex 1.

The single point mutation genome editing was conducted in collaboration with the company WellGenetics following the designs presented in Annex 1 in a w^{1118} genetic background. The obtained results were perfectly in line with our predictions. Specifically, the introduction of either of the mutations G497P, K532G or E147V did not allow for homozygotization, while introducing I618P or A262F mutations led to viable homozygotes.

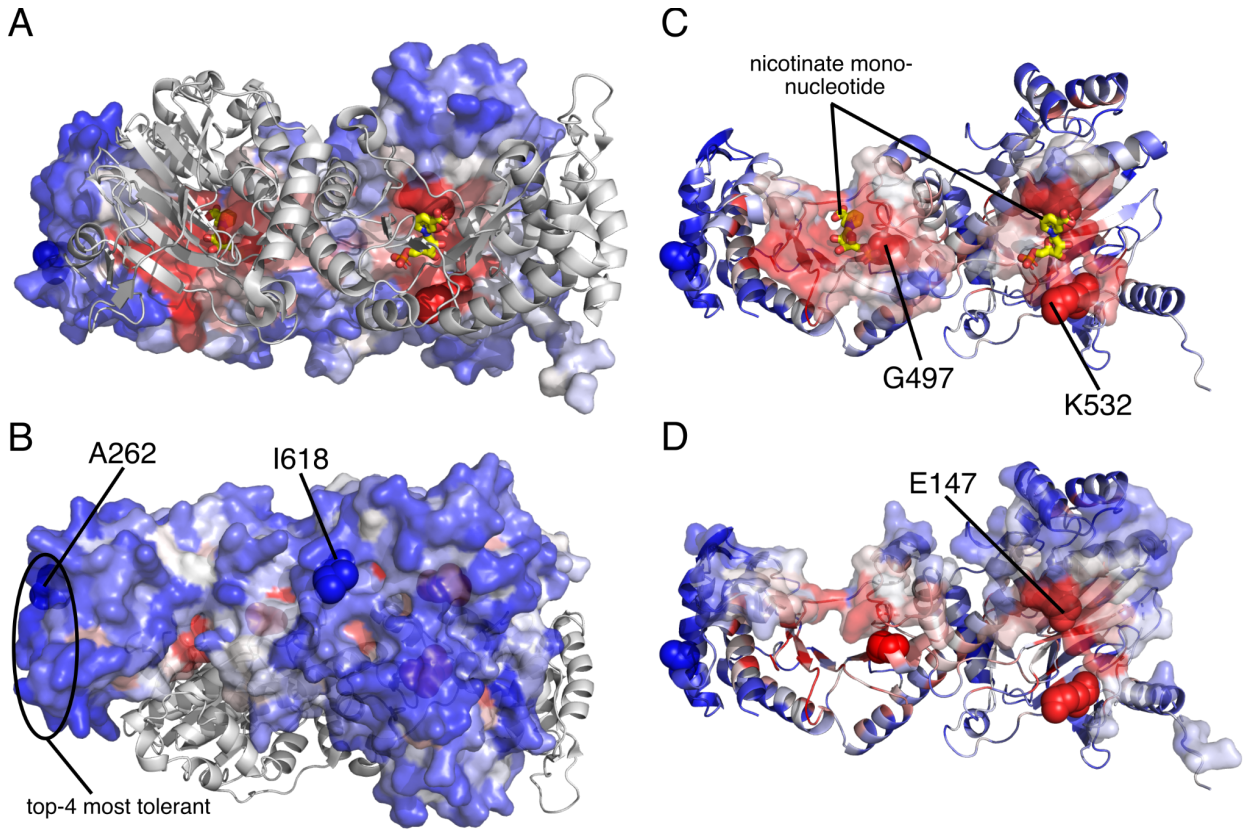

Figure 3: 3D mapping of Naprt mutational sensitivity and selected variants. The AlphaFold 3D model of the Naprt-PH proteoform (FlyBase id: FBpp0306840, 662 residues) is colored according to ProteoCast-predicted mutational sensitivity, from blue (low sensitivity) through white to red (high sensitivity). The unstructured regions 1-21 and 301-407, with pLDDT<70, are omitted for clarity. The five variants selected for CRISPR validation are highlighted in spheres and labelled. **A-B.** Complex obtained by superimposing the biologically relevant homo-dimeric arrangement of the human Naprt (PDB id: 4YUB, see also Table S4 and Fig. S6). The two yellow compounds, nicotinate mononucleotides, indicate the putative location of the active sites. They were positioned by superimposing the structure of TaNAPRTase from *Thermoplasma acidophilum* (PDB id: 1YTK). **A.** Front view. **B.** Back view. **C-D.** Transparent surfaces indicate regions potentially involved in binding the substrate or in forming the homo-dimer. **C.** Residues within 10 Å from the nicotinate mononucleotides. **D.** Residues within 5 Å from the superimposed human Naprt's second monomer.

To further demonstrate the added value of ProteoCast, we compared it with the popular tool SIFT [32] (**Fig. S7**). Predictions from the two methods showed overall good agreement (**Fig. S7A,C**). However, ProteoCast offers much finer resolution and higher dynamic range of the mutational effects (**Fig. S7B,D**). This ability of ProteoCast to discriminate between substitutions across protein sites and also at each protein site constitute a significant advantage when prioritizing variants for experimental testing. In contrast, selecting a few variants among the couple of thousands (~25% of all possible variants) to which SIFT assigned the minimal score of 0 in the structured part of the protein may prove challenging.

Discovery of interaction and regulatory sites in unstructured regions

ProteoCast systematically maps predicted variant scores to proteoform 3D models from the AlphaFold database [18,33], leveraging its extensive protein structural data. We observed that

when GEMME faces difficulty making accurate predictions due to limited evolutionary information, AlphaFold2 similarly struggles to produce reliable 3D models. Almost all residues (96%) with unreliable GEMME predictions are located within protein regions with low- or very low-confidence AlphaFold2-predicted 3D coordinates (predicted Local Distance Difference Test scores below 70, **Fig. S8A**). Conversely, however, AlphaFold2's inability to model 3D structures does not imply that the evolutionary signal is too weak for ProteoCast to predict variant effects. GEMME provided confident predictions for an overwhelming fraction (94%) of the residues with pLDDT scores below 70 (**Fig. S8B**).

Visual analysis of Yorkie's mutational landscape provides a preliminary understanding of the type of signals detected by ProteoCast in unstructured regions (**Fig. 2A**). One can easily identify several patterns distinctive of high mutational sensitivity (strongly negative average GEMME raw scores) within regions of low or very low pLDDT scores. In particular, the segment spanning residues 135 to 160, with pLDDT scores below 55, stands out prominently against its surrounding background in the mutational landscape (**Fig. 2A**). This strong mutational sensitivity suggests functional importance for this segment, going beyond the existing knowledge in UniProt – intrinsic disorder detected by MobiDB-lite [34], without any known function. Another region, namely from residues 32 to 98, displays high mutational sensitivity while being partially structured (mean pLDDT score of 63, **Fig. 2A**). Uniprot does not provide any information about this region. Nevertheless, a search for homologous sequences revealed that it corresponds to the TEAD binding domain of Yorkie's human homolog, YAP1 [35]. This domain is natively unfolded and conditionally adopts a stable conformation upon binding to its partner. The AlphaFold2 prediction for Yorkie closely mirrors the TEAD-bound conformation of YAP1 (**Fig. S9**). These pieces of evidence support the hypothesis that this region is important for the function of the protein by establishing interaction(s) with partner(s). The high concentration of known phospho-sites – 10 out of 20 annotated in iProteinDB [36], within the two identified segments (32–98 and 135-160) further underpin their functional role. For the sake of completeness, let us stress that the two additional Yorkie regions that are highly sensitive to mutations (residues 241–274 and 310–343, **Fig. 2A**) correspond to the well-characterized and well-structured WW domains (**Fig. S8**).

For a more systematic and automated detection of interaction and regulatory sites in unstructured regions, ProteoCast applies a segmentation strategy to GEMME mutational sensitivity profiles (**Fig. 1B,** *Sensitivity Profile*). The segmentation algorithm accounts for the pLDDT profile to tune granularity (see ***Materials and Methods***). To illustrate this functionality, we consider the Delta protein, an 833 residue-long ligand of the Notch signaling pathway (**Fig. 4**). ProteoCast segmentation analysis emphasised four distinct segments within the unstructured C-terminal part of the protein that stand out from their surrounding background (**Fig. 4B**, in purple). Three of these segments encompass almost all residues ProteoCast considers sensitive to mutations (marked with stars in **Fig. 4B**). The segmentation result and the high concentration of sensitive residues suggest high functional importance for these segments, in line with experimental data showing their direct implication in regulating Delta's ubiquitination and activation by Mindbomb1 and Neuralized (**Fig. 3B**, black and pink boxes) [37]. By contrast, the ubiquitination sites themselves (marked with "K" in **Fig. 4B**) show varying levels of mutational sensitivity, consistent with experimental evidence suggesting their dispensability – ubiquitin does not necessarily need to be added to these lysines for proper regulation [37]. Among the few other sensitive residues, four cysteines, namely C710, C731, C786 and C802,

are associated with sharp peaks in the mutational sensitivity profile (**Fig. 4B**). These evolutionary signals, detected without any supervision by ProteoCast, may be interpreted in light of experimental studies supporting the essential regulatory role of reactive cysteines within intrinsically disordered regions, particularly in response to Reactive Oxygen Species (ROS) [38,39]. This region may thus function in a redox-dependent manner, with conditional disorder [40].

Beyond Yorkie and Delta example proteins, we quantitatively assessed ProteoCast predictions in unstructured regions over the entire fly proteome (**Fig. 1C,** *Interaction & regulatory sites*). We relied on annotations from the iProteinDB database [36], which comprises 40.5K post-translational modification (PTM) sites, and the ELM resource [41] featuring ~90 short conserved stretches of amino acids mediating protein-protein interactions, known as short linear motifs (SLiMs). The vast majority of PTM sites and SLiMs annotated in these databases are located in regions with pLDDT scores below 70 (**Sup. Fig. S10**). ProteoCast identified about one-third of the low-pLDDT PTM sites as sensitive to mutations (**Table S5**). This proportion increases up to nearly 50% when considering regions whose mutational sensitivity stands out from their surrounding background with ProteoCast segmentation strategy (**Table S5**). Moreover, 83% of the low-pLDDT SLiMs contain at least one sensitive residue, with 68% of SLiMs entirely residing in segments with significantly higher sensitivity compared to their background. Overall, these results confirm that SLiMs, and to a lesser extent PTMs, are under selective pressure. Our estimates are in line with previous studies reporting between 35% and 65% of phosphosites under evolutionary constraints across eukaryotic proteomes [42,43]. Furthermore, our analysis demonstrates that ProteoCast residue classification (as sensitive or tolerant) and mutational sensitivity segmentation provide two complementary perspectives, the former global and absolute, the latter local and relative, on the problem of detecting functionally important sites (**Fig. S11**).

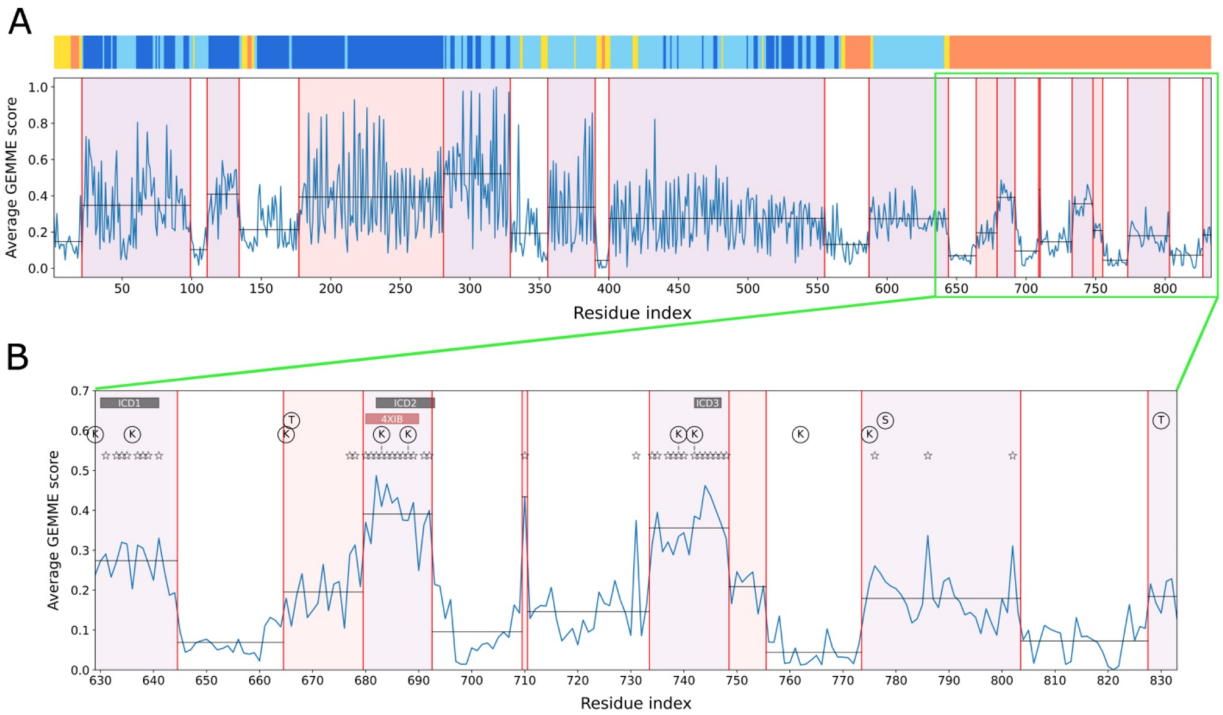

Figure 4: Structural confidence and mutational sensitivity of the Delta protein. **A. Top**. Structure confidence as estimated by AlphaFold2 pLDDT. **Bottom**. Mutational sensitivity as estimated by GEMME average score. The two curves are segmented using the FPOP algorithm (see Materials and Methods). The identified segments are colored in yellow, red or purple depending on whether their mean value (horizontal line) is lower than those of the two neighbouring segments, higher than one of them, or higher than both, respectively. **B**. Close-up view of the GEMME sensitivity profile for the C-terminal region from residues 642 to 833, where the stars highlight residues sensitive to mutations, according to ProteoCast classification. The black boxes mark experimentally validated binding motifs for ubiquitylation by Mindbomb1 and Neuralized [37]. The pink box indicates the region whose structure was experimentally resolved in complex with Mind bomb 1 (MIB1) (PDB code: 4XIB, residues 680 to 690). Ubiquitination sites are indicated by circled "K," and phosphorylations by circled "S" and "T".

Mutational landscape resolution as a proxy for gene essentiality

The resolution of ProteoCast-predicted mutational landscapes directly depends on the diversity of amino acids observed across species. Indeed, ProteoCast will not be able to distinguish between amino acid substitutions that do not appear in the input alignment. The mutational landscapes generated for the fly proteome are generally of high resolution, in line with the fact that most predictions have high local confidence. Namely, for most of the proteoforms, more than 70% of all possible amino acid substitutions (19 multiplying the protein length) are in fact observed in the input alignment. Moreover, a subset of 725 unique proteoforms from 593 genes have their mutational landscapes nearly entirely covered by the input alignment, with the fraction of observed mutations (F_{obs}) exceeding 95% (**Fig. S12**). This near-complete evolutionary sampling of substitutions at each position highlights the ubiquity of the corresponding protein families across life kingdoms. These proteoforms have many identifiable homologs (at least 4,000 sequences in the alignment) that retain detectable similarity while exhibiting extensive variation at individual sites. They are more likely to harbor mutations impairing organismal fitness compared to the rest of the proteome – 1.8 fold enrichments in lethal mutations. They tend to be more connected in the fly interactome, with a 2.8-fold higher

likelihood of having >20 protein partners. Furthermore, their genomic loci are significantly enriched for regions associated with transcription factor (TF) binding sites recognised by early development genes such as *CTCF*, *BEAF-32* or *Dref* that are also involved in chromatin domain organisation and early development (**Fig. S13**). Collectively, these findings suggest the functional importance of these proteins to the organism through their involvement in key developmental processes.

The higher mutational landscape resolution is associated with a distribution of GEMME predicted raw scores shifted toward higher values, highlighting the need for adaptive thresholds for variant classification (**Fig. S14**). As an example, let us consider the two known lethal mutations Y264F and D222N in the serine/threonine phosphatase Flapwing (Flw), essential for larval development. With raw scores of -1.38 and -0.97, ProteoCast classified these mutations as uncertain and correctly distinguished them from the DGRP mutation E314D, classified as neutral (**Fig. S15**). If we were to apply the thresholds defined for the Yorkie protein, the resulting classification would be incorrect, with the two lethal mutations appearing as neutral (**Fig. 2**). More broadly, applying a universal neutral-uncertain threshold across all proteins (-2.4, Table S3) would increase the number of false negatives from 30 to 44 (out of 128) in proteins with $F_{obs} \geq 95\,\%$.

Beyond underscoring the relevance of adaptive classification, these findings also suggest that a protein's relative importance to the organism can, to some extent, be inferred directly from its evolutionary history, without needing additional information about population, genetics or interactomics data.

Case study on Toll pathway proteins

We finally conducted a focused analysis to examine the relationship between natural polymorphism, available cross-species sequence data, and ProteoCast predictions, using a multi-protein case study from the Toll pathway (**Fig. 5**, **Table S6**). This pathway is essential to the developmental patterning of the dorso-ventral axis as well as immunity. Consistent with our proteome-wide analysis, ProteoCast classified only a small fraction of population polymorphisms (18 out of 344) as impactful in these proteins. The Toll membrane receptor, for example, exhibits 54 mutations in the DEST2 and DGRP populations, all deemed neutral by ProteoCast. The vast majority of the misclassified mutations (14 out of 18) are exclusively observed in the DEST2 pool (**Table S6**). None of the mutations exclusively observed in the DGRP are predicted as impactful. This result highlights the purifying selection operating on flies inbred in the laboratory for more than 10 years - highly homozygote - compared to wild individuals - mostly heterozygote. Moreover, mutations observed exclusively in the DEST2 pool are about four times more numerous than those exclusively observed in the DGRP, in line with proteome-wide estimates (DEST2/DRGP ratio of 3.96 for all 22,392 representative proteoforms). Larger proteins tend to harbor more polymorphisms, a tendency also confirmed at the level of the entire proteome (adjusted R^2 = 0.33 for 18,238 proteoforms spanning 40-1,000 residues).

The fraction of mutations observed in the input alignment varies widely across proteins, from less than 10% for Mtk – suggesting its specificity to a few organisms – to over 90% for the

extracellular serine proteases Snake and SPE, both cleaving the Toll ligand Spaetzle. This high fraction reflects the ubiquity of this family of enzymes in eukaryotes and prokaryotes. Intermediate fractions (50–65%) were observed for both the intra-cellular proteins MyD88, a transducer of the Toll receptor, and its negative regulator Pellino. However, the mutational landscapes of these two proteins contrast strikingly. MyD88 harbors unstructured regions that are poorly covered in the input alignment and prone to accumulating natural polymorphisms, and thus less likely to alter the protein function. Conversely, Pellino is largely structured and the alignment is dominated by homologs sharing high sequence similarity (>40%). Its low number of natural variants in DEST2 (5 mutations) likely stems from strong structural constraints. ProteoCast classified 3 of these mutations as impactful, indicating Pellino's potentially lower essentiality compared to MyD88, where 19 out of 21 mutations were classified as neutral.

Overall, the systematic application of ProteoCast can assist experimentalists and theorists in exploring the role of natural polymorphism in the function of specific proteins or entire pathways, offering complementary information for interpreting Genome Wide Association Studies (GWAS).

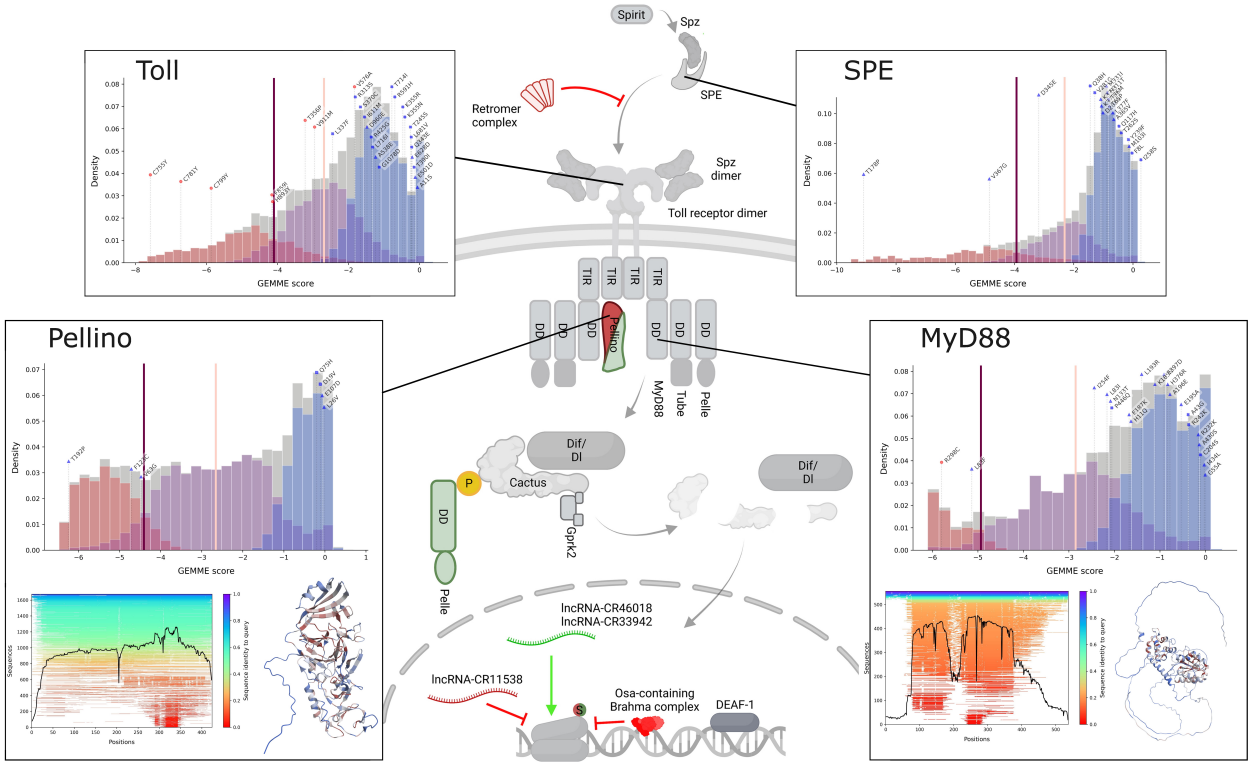

Figure 5: ProteoCast predicted mutational landscape for proteins from the Toll pathway. ProteoCast score distributions and variant classification are represented for four proteins of the Toll pathway, the membrane receptor Toll (FBpp0084431, 1,097 residues), the serine protease SPE (FBpp0083832, 400 residues), the transducer MyD88 (FBpp0087679, 537 residues) and its negative regulator Pellino (FBpp0083913, 424 residues). The three fitted Gaussians are highlighted in colors and the vertical bars indicate the variant classification thresholds, as in Fig. 2B. Sampled natural polymorphisms (inbred or wild) and reference lethal mutations are depicted as blue and red points, respectively. For Pellino and MyD88, we also show the input alignment plots and the AlphaFold-predicted 3D models colored according to ProteoCast-predicted mutational sensitivity, from blue through white to red. The schematic representation of the Toll pathway is adapted from Valanne *et al.* [44].

Discussion

Proteocast has three primary purposes for the Drosophila community and beyond. Firstly, it facilitates the identification of genome-editing targets by enabling rapid prioritisation of the hundreds of millions of possible missense mutations over the fly proteome. Secondly, it sheds light on the evolutionary conservation and 3D structural context of hundreds of thousands of missense variants found in inbred and wild fly populations and lacking annotations. Thirdly, it can aid in elucidating the role of unstructured regions of the fly proteome by pointing at potential interaction and regulatory sites within these regions. It offers freely accessible, user-friendly and interactive online web services for a detailed examination of any single-point variant of interest at https://proteocast.ijm.fr/drosophiladb/. The functional validations of our prediction using the rich datasets available for Drosophila suggest the relevance of ProteoCast as a strongly efficient predictor of missense mutations' functional impact across organisms.

We have provided a proof-of-concept for prioritising missense mutations using ProteoCast predictions. The CRISPR experiments presented in this article directly validated this strategy. The results demonstrated ProteoCast's ability to accurately predict organism-level mutational impact, specifically in terms of binary phenotypic outcomes such as developmental viability. Moreover, they allowed for discovering three recessive developmentally lethal mutations in a protein involved in the biosynthesis of a coenzyme central to metabolism. Ongoing and future work will aim at full biochemical characterisation of the mutants.

Furthermore, our analysis has revealed that a substantial fraction of variants found in inbred and wild fly populations are counter-selected in evolution, suggesting that they may alter protein or organismal fitness. Some of them occur in proteins with known lethal missense mutations, thus likely essential for the organism. For instance, in the Mi-2 protein – a nuclear ATP-dependent nucleosome remodeler of the CHD family, a set of DEST2 mutations ProteoCast classified as impactful are clustered in 3D near the putative DNA-binding interface (**Fig. S16A**). The Doa protein, which contributes to somatic sex determination, gives another example where the DEST2 mutation V364G has a much more negative predicted score than the referenced lethal mutation C275Y (**Fig. S16B**, see also https://proteocast.ijm.fr/results/results/?q=FBpp0289026). The two mutation sites are located in the protein's kinase domain, in close proximity and facing each other. These findings call for biochemical characterisation of the putative impact of these naturally occurring mutations on protein activity and interactions. In addition, ProteoCast predictions could be useful for informing the comparison between the polymorphism fixed in DGRP, inbred in the laboratory for more than 10 years, with that of natural populations in DEST2. Reciprocally, ProteoCast predictions may help refine annotations for the fly proteome. We identified 32 mutations referenced as developmentally lethal in FlyBase, yet observed in inbred or wild fly populations. Of these, ProteoCast provided high-confidence predictions for 21 mutations, classifying all but one as neutral. These findings highlight the need for further investigation for better curating the lethality annotations. For example, a closer examination of the original publication and FlyBase record for the R10K mutation in the tyrosine-protein kinase hopscotch (FBpp0073313) reveals that it is associated with developmental lethality in conjunction with a premature stop codon. This stop codon is likely responsible for the phenotype as it causes the loss of an entire kinase domain [45]. Finally, our results could help to study the relationship between seasonality, geography and natural polymorphism. Rather than examining the entire genome, where disentangling signals from noise may prove challenging, such studies

could focus on the identified DEST2 outlier variants that are predicted to have significant impacts. We explored that approach for two xenobiotics detoxifying genes, Cyp6a9 and Cyp317a1, that were shown to have geographically influenced incidence (**Fig. S17)**.

We have systematically mapped ProteoCast predictions on AlphaFold-predicted 3D models over the fly proteome. Both AlphaFold2 and ProteoCast exploit information coming from an input multiple sequence alignment. However, while ProteoCast extracts evolutionary signals from the alignment in an unsupervised fashion, AlphaFold2 directly maps the alignment to 3D coordinates. Difficulties in establishing this mapping result in low pLDDT but, as we have shown, do not rule out the possibility for ProteoCast to identify protein sites under selection and thus likely important for protein interactions and functions. Our results have highlighted ProteoCast potential for efficient and transparent screening of phosphoproteomics datasets toward identifying functionally relevant PTM sites, without requiring any expert intervention. ProteoCast adaptive detection of evolutionary signals in unstructured and variable regions, through an automated segmentation procedure, may prove more advantageous than other methods relying on absolute thresholds [46]. Beyond filtering and prioritising known PTM sites and linear binding motifs, ProteoCast can also help to discover yet uncharacterised functionally relevant sites hidden in disordered regions.

We chose to rely on a robust and extensively validated variant effect predictor that provides a clear readout of the multiple sequence alignment given as input. Yet, alignment errors might result in variant or residue misclassification. In our analysis, we observed that most of the low-confidence predictions are located at the protein extremities, likely due to a high proportion of gaps in the input alignment. As a consequence, ProteoCast often misses lethal mutations affecting the first methionine – 11 misclassified as neutral out of a total of 21. The same trend might arise for residues lying at the junction between highly covered and highly gapped regions, due to ambiguities in gap opening/closing positions. Hence, we advise users to inspect the input alignments for such potential biases. The protein language model-based predictor VespaG [13] may produce more accurate predictions in these problematic regions, or when the whole alignment is of poor quality. We previously showed that VespaG helps to regularise GEMME scores thanks to its knowledge of a universal protein representation space [13].

Future research will aim at a tighter integration of evolutionary and structural data. For instance, applying ProteoCast segmentation strategy to AlphaFold2 pLDDT profiles revealed that around 34% of the known PTM sites and 35% of the known SLiMs fall within regions modeled with higher confidence than their surroundings. This detection likely reflects AlphaFold2's ability to recapitulate and generalise experimental knowledge about conditional folding. It only partially overlaps with ProteoCast detection from GEMME mutational sensitivity profiles, suggesting avenues for combining the two types of information. Another direction will consist in extending the resource to other common model organisms, such as *X. laevi*, *C. elegans*, *S. cerevisiae*, and *E. coli*, as well as *H. sapiens*. The wealth of mutational data and associated phenotype will help further assess the quality of our predictions. The generated data will facilitate the evaluation of large-scale predictions against phenotypic knowledge at both organismal and molecular levels as well as significantly ease the identification of residue-specific genome editing. Our findings will serve as a starting point toward identifying the molecular determinants of continuous phenotypic traits controlled by multiple genes.

Material and Methods

Drosophila experiments

Lines used. *daGS* driver [47] (provided by the Monnier lab, Université de Paris Cité).

Vienna Drosophila Resource Center stock (gene name): 42571 (*JIL-1*, FBgn0020412), 31712 (*brm*, FBgn0000212), 31053 (*HDAC6*, FBgn0026428), 61875 (*Naprt*, FBgn0031589), 35704 (*smid*, FBgn0016983), 67758 (*Gpo3,* FBgn0028848), 31395 (*Rm62*, FBgn0003261), 28038 (*nAChRβ2*, FBgn0004118) and 27503 (*CadN*, FBgn0015609).

RNAi-induced developmental lethality. All the flies are kept in closed vials in incubators at controlled temperature, humidity (60%) and 12 hours light cycle. Experiments are carried at 26°C. 5 individual replicate vials were made per condition. Each RNAi was induced using two distinct concentrations of RU486 0 µg/mL (negative control) and 20 µg/mL, the maximum inducer concentration usable during development as previously described [48]. Each contains 6 *da*GS virgin females and 5 RNAi males left to cross and lay eggs for 48 hours on the following food composition: 8.33% (w/v) yeast, 8.33\% (w/v) corn, 1.13 (w/v) agar and Methyl 4-hydroxybenzoate (Moldex) at a final concentration of 5.3 g/L to prevent fungal contamination. Eclosing flies were then counted daily for 10 days.

Datasets

FlyBase. We retrieved the *Drosophila melanogaster* proteome version 6.44 as a FASTA file (*dmel-all-translation-r6.44.fasta*), available at FlyBase (https://ftp.flybase.net/genomes/). This transcriptome contains a total of 30,738 sequences, of which 22,392 are unique. We focused our attention on single point mutations generated by ethyl methylsulfonate (EMS) and made a list of those associated with developmental lethality - using the Phenotype QuickSearch tool provided by Flybase - as well as a second list of hypomorphic mutations. We filtered out the 250 mutations overlapping between the two.

Drosophila Genetic Reference Panel. We obtained all polymorphisms from the 205 inbred lines of the Drosophila Genetic Reference Panel (DGRP) [28] in a text file (*dgrp.fb557.annot.txt*). This dataset contains approximately 4.5 million genomic variations, including deletions, insertions, single nucleotide polymorphisms (SNPs), and multi-nucleotide polymorphisms (MNPs). SNPs are the most prevalent type of variations (89% of all detected variants), ~4.5% of which result in missense mutations leading to amino acid changes in proteins. Mapping these missense mutations to the fly proteome sequences resulted in 178,264 unique variants affecting 20,137 (89.9%) proteoforms from 12,682 genes (90.8%).

Drosophila Evolution over Space and Time. The Drosophila Evolution over Space and Time (DEST) [29] resource is a large-scale dataset that compiles genomic polymorphism data from natural *Drosophila melanogaster* populations sampled across diverse geographic locations and time periods. We obtained all annotated SNPs from the August 2024 version of the dataset in VCF format (*dest.all.PoolSNP.001.50.3May2024.ann.vcf.gz*). It includes approximately 4.8 million SNPs classified into 27 different annotation types. For this study, we focused specifically on missense SNPs, identifying 332,848 variants that affect 94% of Drosophila genes and 92.4%

of proteoforms. After resolving mismatches in gene and transcript identifiers, we retained 285,771 unique missense SNPs for further analysis.

Post translational modifications and short linear motifs. We retrieved all PTM sites annotated for the *Drosophila melanogaster* proteome in the iProteinDB [36] database at https://www.flyrnai.org/tools/iproteindb/ through its API. We identified PTM sites in 8,530 representative proteoforms (38%) from 4,276 genes (31%). The vast majority of these sites are targeted by phosphorylation (~92%), the remaining sites being oxidized or acetylated.

Short linear motifs. We obtained 113 short linear motifs (SLiMs) ranging from 3 to 20 amino acids, from the Eukaryotic Linear Motif (ELM) resource (http://elm.eu.org/) [41]. These motifs were retrieved from a downloadable file, *elm_instances.tsv*, using the following search path: ELM DB → ELM instances → Filter by organism (*Drosophila melanogaster*).

ProteoCast workflow

The ProteoCast workflow unfolds in four main steps.

1. **Multiple sequence alignment generation (Fig. 1A).** Given an input query protein sequence, it retrieves related sequences using the MMseqs2-based protocol implemented in ColabFold [49].
2. **Full single-mutational landscape prediction (Fig. 1A).** It uses the variant effect predictor GEMME [7] to estimate the impact of all possible amino acid substitutions at every position in the query. Then, it assigns binary confidence scores that reflect the amount and quality of evolutionary information in the input alignment.
3. **Variant and residue classification (Fig. 1B).** It classifies variants as impactful, uncertain or neutral by fitting a Gaussian mixture model [50] to the predicted raw scores. It further classifies residues as sensitive to mutations if at least 10 out of the 19 possible substitutions are impactful. Otherwise, residues are considered tolerant to mutations.
4. **3D structural mapping and segmentation (Fig. 1B).** It maps residue mutational sensitivity scores onto the 3D model predicted by AlphaFold2 for the query protein. It decomposes the mutational sensitivity profile into segments using the FPOP algorithm [51,52]. The granularity of the segmentation accounts for AlphaFold pLDDT profile.

In the following, we give details about the different steps, illustrating their application over the entire fly proteome.

Step 1. Multiple sequence alignment generation.
For every unique proteoform from the fly proteome (FlyBase version 6.44), we generated a multiple sequence alignment (MSA) using the ColabFold (CF) protocol [49]. We previously showed that this protocol compares favourably with workflows relying on profile hidden Markov models for the prediction of mutational effects [26]. It is much faster, thanks to its usage of the highly efficient sequence MMseqs2 search algorithm [53], and it yields similar or better predictive accuracy. CF protocol produced two alignments for each input sequence by searching against UniRef30, a clustered version at 30% sequence identity of UniRef100 [54], and a curated database comprising various environmental sequence sets [49] (ColabFold

environmental database, ColabFold_Env). We used UniRef30 v2103 and ColabFold_Env v202108. We concatenated the alignments and gave the resulting one as input for GEMME. The CF protocol did not manage to find any homolog for 9 proteoforms.

Step 2. Full single-mutational landscape prediction.
We utilised the *reformat.pl* tool from the HH-suite [55] to convert A3M alignment files into FASTA format. We handled sequences containing undefined residues (represented by symbols "X" or "U") by employing the following strategy: when the undefined residue was located at the beginning of the sequence, the rest of the corresponding column in the alignment was always filled with gaps, and thus we removed that column. For other instances of undefined residues, we replaced them with the most frequently occurring amino acid at the respective position(s) within the alignment. We ran GEMME with default parameters except that (1) we increased the memory (RAM) to handle very long proteins (*Xmx9126m*), (2) we increased the number of JET [56] iterations to 7 to ensure robustness (see Supplemental Methods for more details). In total, GEMME produced some predictions for 22,169 unique proteoforms (**Table S1**). ProteoCast then estimated the confidence in each of these predictions.

On a per-protein basis, we consider that mutational landscapes predicted from less than a couple hundred sequences may be of low quality. This cutoff is based on our previous findings [26] and is suited for the MMseqs2-based MSA generation protocol implemented in ColabFold. On a per-residue basis, we use GEMME predicted scores' standard deviation over the 19 possible substitutions as a proxy for the MSA information content. A small standard deviation indicates that GEMME is not able to distinguish substitutions from one another. On the one hand, this situation typically occurs upon strong selection pressure, when a position is highly conserved in evolution and thus does not tolerate any substitution. A high conservation level implies a limited subset of amino acids observed in the input MSA and thus strong supporting evidence for lack of diversity. On the other hand, a lack of evidence may warrant a narrow distribution of predicted scores, for instance in the case of highly gapped or highly divergent positions. In this situation, the conservation level will be low (**Fig. S1**). We consider positions displaying a narrow GEMME score distribution, *i.e.*, a low evolutionary conservation, and a low coverage or diversity in the input MSA as lacking supporting evidence. To detect these positions, we devised a binary-valued per-residue confidence score detailed in Supplemental Methods and illustrated in **Fig. S1**.

Step 3. Variant and residue classification.
The range of variant predicted raw scores may vary substantially from one protein to another, making direct comparisons over the entire proteome difficult. This motivated us to devise an adaptive variant classification strategy tailored to each protein-specific raw score distribution. Namely, ProteoCast uses a Gaussian Mixture Model (GMM), a probabilistic approach that assumes the data are generated from a mixture of Gaussian distributions [57]. The GMM parameters are estimated through the Expectation-Maximization (EM) algorithm [58]. Similar strategies with two or three Gaussians were previously proposed to classify experimental measurements [23] or scores predicted by a variational auto-encoder [8]. In our analysis, we obtained the best performance and highest interpretability by fitting three Gaussians G_1, G_2 and G_3, with median values η_{G1}, η_{G2}, η_{G2} , and η_{G3}, respectively, and $\eta_{G1} < \eta_{G2} < \eta_{G3}$. We used the *mixture* module from python *scikit-learn* library [59] with component-specific general

covariance matrices ("full" type), 30 initialisations, a convergence threshold of 10e-5, and a maximum number of EM iterations of 1,00. Moreover, to account for the difficulty in sharply distinguishing variants that would have a significant impact on the protein function from neutral variants, we chose to classify the variants into three categories, namely neutral, uncertain, and impactful. Let $x_{a_i \to b_i}$ be the raw score predicted for substituting the amino acid *a* by *b* at position i,

$$\text{if } x_{a_i \to b_i} > \eta_{G_2}, \text{ then } a_i \to b_i \text{ is neutral,}$$
$$\text{if } P_{G_1}(x_{a_i \to b_i}) > P_{G_2}(x_{a_i \to b_i}), \text{ then } a_i \to b_i \text{ is impactful,}$$
$$\text{otherwise, } a_i \to b_i \text{ is uncertain.}$$

For each variant, ProteoCast computes a z-score relative to the distribution of neutral mutations, Z-score$(a_i \to b_i) = (x_{a_i \to b_i} - \mu_{neutral})/\sigma_{neutral}$, where μ and σ stand for the mean and standard deviation, respectively.

Step 4. 3D structural mapping and segmentation.
For every unique proteoform from the fly proteome (Flybase version 6.44), we retrieved its 3D model either by batch downloading (https://www.uniprot.org/proteomes/UP000000803) from the AlphaFold Database (13,475 structures), or by using Python requests library [60] (5,119 structures). For the missing structures, we generated all those with fewer than 4,000 amino acids (3,345 structures) by running ColabFold with default parameters. ProteoCast then mapped the computed per-residue mutational sensitivity values (average of raw scores) and classes onto each protein's 3D model. It further used AlphaFold pLDDT scores to segment the mutational sensitivity profile, as described in the following.

The problem of segmenting a one-dimensional continuous measure along a protein sequence can be formulated as follows. We consider D changepoints $\tau_1 < \ldots < \tau_D$ within the range 1 and $n-1$. These changepoints correspond to unknown positions along a protein of length n, where a shift in the mean of the measure is observed. We adopt the convention that $\tau_0 = 0$ and $\tau_{|\tau|} = n$. These changepoints define $|\tau| = D + 1$distinct segments. The j-th segment includes the data $\{\tau_{j-1} + 1, \ldots, \tau_j\}$. Each segment is premised on the assumption that the Y_i therein are independent and follow the same Gaussian distribution, with a mean μ_j specific to that segment and a common variance σ^2. Expressed mathematically, we have:
$\forall i \in \{\tau_{j-1} + 1, \ldots, \tau_j\},\ Y_i \sim \mathcal{N}(\mu_j, \sigma^2)$

The parameters of the model (eq: 1), including $\tau_1 < \ldots < \tau_D$, can be estimated using penalized maximum likelihood inference. To achieve this, we used the FPOP algorithm [51,52] which solves the inference problem exactly with dynamic programming. Informally, the FPOP algorithm works by considering the penalized maximum likelihood of the data from observation 1 to t as a function of the parameter (the mean) of the last segment. This concept is referred to as 'functional pruning'. In the Gaussian case, the resulting function is piecewise quadratic. When a new observation at time $t+1$ is made, this function can be efficiently updated (i.e., compute the

penalized maximum likelihood function from observation 1 to t) using a formula similar to that of the Viterbi algorithm. This formula is applied piece by piece or interval by interval. At each step, the algorithm searches for the optimal value of the parameter for the last segment that maximizes the penalized likelihood.

The number of changepoints D estimated by FPOP is a decreasing function of the penalty $\alpha\ \sigma^2\ log(n)$. The constant α is a hyperparameter set to 1.4, determined after assessing the fraction of detected PTMs in the resulting segmentations while ensuring the visual goodness-of-fit (**Fig. S18**). The variance σ^2 is estimated on the data using the Hall estimator [61]. We observed that the variations in mutational sensitivity were higher in protein regions modelled with high confidence by AlphaFold2 (pLDDT > 0.7) compared to other regions. To avoid over-segmenting these areas (by introducing artificial changepoints), we applied lower weights to the observations in these regions, with w set to 0.1, while keeping it at 1 elsewhere. After estimating the segments, we assign discrete scores to each of them. A score of 0 indicates that the means of both adjacent segments are greater than that of the current segment; a score of 1 that one adjacent segment has a lower mean and the other a higher mean than the current segment; and a score of 2 that the means of both adjacent segments are lower than that of the current segment.

Implementation and computational details

To generate the input we run ColabFold protocol, multiple sequence alignment generation part, through the following command: *colabfold_search --threads=8 --db-load-mode=2 $fasta_file $ {COLABFOLD_DB} $out_folder*. For a set of 3,000 sequences comprising a total of 1,185,789 amino acids it took 1.5 hours to generate MSAs on the supercomputer “MeSU” of Sorbonne University (16 CPUs from AMD EPYC Milan processors, 400GB shared RAM memory).

We implemented ProteoCast in Python and R. ProteoCast relies on *scikit-learn* library [59] to fit a Gaussian mixture model, the *gemmi* library [62] to parse PDB files and manipulate the B-factors associated to the structure, the *re* module to parse annotation data files, and pandas for data manipulation. For the same set of the protein sequences the generation of mutational landscapes, classification and analysis (**Fig. 1A** *Mutational Landscape* and **1B**) took 4 hours on 40CPUs.

The online database (https://proteocast.ijm.fr/drosophiladb/) is a Django-based project running on a virtual machine hosted by *Institut Jacques Monod* (Dell PowerEdge R740 server). The backend was developed in Python using Django, while the frontend was built with HTML, CSS, and JavaScript to provide a user-friendly interface. The application was deployed on an Apache2 web server with Django’s WSGI integration. It incorporates *plotly* for interactive data visualization and *Mol*(Molstar)* [63] for 3D structure visualisation. Data handling and processing were facilitated using pandas.

Evaluation

We estimated the performance with the balanced accuracy metric, suited to unbalanced datasets and is defined as $\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$, where $\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$ in our study the fraction of DGRP or DEST2 polymorphisms predicted as neutral, and $\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ in our study the fraction of lethality inducing mutations predicted as impactful. If several proteoforms are affected by a SNP, we take the one with the highest F_{obs} score.

Analysis of case studies

We used iCisTarget to search for enrichment in transcription factor (TF) binding sites [64,65]. In brief, i-CisTarget searches existing databases of known TF binding sites as well as experimental chromatin immunoprecipitation experiments. We considered two categories of genes defined based on the fraction of observed mutations, either $F_{obs} \geq 95$ or $85 < F_{obs} < 90$. We extracted a subset of 635 genes for the $85 < F_{obs} < 90$ category to have roughly the same cardinality as the $F_{obs} \geq 95$ set (593). We considered the enrichment to be significant with a number of enriched features above 7 for a threshold set at Normalised Enrichment Score (NES) threshold of 5.

We visualised 3D structures with Pymol [66] and we used the *align* command to superimpose structures resolved for homologs. For the Naprt example, after careful examination of the publication accompanying the PDB entry 4YUB [67], we realised that the dimeric arrangement annotated as *Biological Assembly* did not match the one described as biologically relevant in the publication. Hence, we generate symmetry mates with Pymol to identify and visualise the correct arrangement (**Fig. S14**).

Database description

For each proteoform, we provide its sequence and the corresponding multiple sequence alignment (downloadable in FASTA format), along with a visual representation, inspired by ColabFold [50]. The mutational landscape can be interactively browsed. It includes raw predicted scores, variant predicted classification, known annotations, and local confidence scores (downloadable in CSV format). Additionally, we provide a visualization of the raw score distribution with variant classification thresholds, and the lethal, inbred, and natural missense mutations mapped on it (downloadable as a CSV file). The segmented mutational sensitivity profile can be interactively zoomed in and out and is accompanied with a color bar for pLDDT values (if a structure is available). Finally, we provide access to the mapping of AF2 pLDDT scores, ProteoCast sensitivity scores and residue classification on the 3D structure (if available, all data downloadable as PDB files with values on the B-factor column).

Availability of data and materials

ProteoCast is an open software publicly available under a GNU license on Github (https://github.com/abakarovaMarina/ProteoCast) [68]. For the analyses presented in this study,

- the pre-computed gene annotations TSV files from Flybase, https://ftp.flybase.net/releases/FB2022_01/precomputed_files/genes/,
- the variant annotation text file based on FB5.57 from DGRP2, http://dgrp2.gnets.ncsu.edu/data/website/dgrp.fb557.annot.txt,
- the variant annotation VCF file from DEST2.0, https://dest.bio/data-files/SNP-tables/dest.all.PoolSNP.001.50.24Aug2024.ann.vcf.gz,
- and the *Drosophila Melanogaster* proteome archive from the AlphaFold DB, https://ftp.ebi.ac.uk/pub/databases/alphafold/latest/UP000000803_7227_DROME_v4.tar

were downloaded. In addition, we manually retrieved 3D models that were missing from the AlphaFold DB *Drosophila Melanogaster* proteome archive.
All these files and ProteoCast predictions are available in the following Zenodo repository [27].

Acknowledgments

This work was granted access to the HPC resources of the SACADO MeSU platform at Sorbonne Université. Further, thanks to all contributing to open-source programming libraries. Last, not least, thanks to all who deposit experimental data in public databases, to those who maintain these databases, and those who make methods available enriching experimental data. This work has been funded by ADAGIO – Ageing and Natural DeAth GenetIc cOntrollers (ANR-20-CE44-0010) and co-funded by the European Union (ERC, PROMISE, 101087830). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank Joël Marchand for providing support with the online database.

References

1. Abbasi J. DNA Base Editing Could Reverse Most Disease-Causing Point Mutations. JAMA. 2017;318:2173.
2. Vertex and CRISPR Therapeutics Announce Authorization of the First CRISPR/Cas9 Gene-Edited Therapy, CASGEVY™ (exagamglogene autotemcel), by the United Kingdom MHRA for the Treatment of Sickle Cell Disease and Transfusion-Dependent Beta Thalassemia | Vertex Pharmaceuticals Newsroom.
3. Ruetz TJ, Pogson AN, Kashiwagi CM, Gagnon SD, Morton B, Sun ED, et al. CRISPR–Cas9 screens reveal regulators of ageing in neural stem cells. Nature. 2024;1–10.
4. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.
5. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol Biol Evol. 2016;33:268–80.
6. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35:128–35.
7. Laine E, Karami Y, Carbone A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. Mol Biol Evol. 2019;36:2604–19.

8. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91–5.
9. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018;15:816–22.
10. Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. Hum Genet. 2022;141:1629–47.
11. Marquet C, Schlensok J, Abakarova M, Rost B, Laine E. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. Bioinformatics. 2024;40:btae621
12. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492.
13. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. SaProt: Protein Language Modeling with Structure-aware Vocabulary. bioRxiv; 2024.
14. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
15. Pathak AK, Bora N, Badonyi M, Livesey BJ, Consortium S, Ngeow J, et al. Pervasive ancestry bias in variant effect predictors. bioRxiv; 2024.
16. Tekpinar M, David L, Henry T, Carbone A. PRESCOTT: a population aware, epistatic and structural model accurately predicts missense effect. medRxiv; 2024.
17. Jagota M, Ye C, Albors C, Rastogi R, Koehl A, Ioannidis N, et al. Cross-protein transfer learning substantially improves disease variant prediction. Genome Biol. 2023;24:182.
18. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024;52:D368–75.
19. Livesey BJ, Marsh JA. Variant effect predictor correlation with functional assays is reflective of clinical classification performance. bioRxiv; 2024.
20. Livesey BJ, Marsh JA. Updated benchmarking of variant effect predictors using deep mutational scanning. Mol Syst Biol. 2023;19:e11474.
21. Notin P, Kollasch AW, Ritter D, Niekerk L van, Paul S, Spinner H, et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. bioRxiv; 2023.
22. Abildgaard AB, Nielsen SV, Bernstein I, Stein A, Lindorff-Larsen K, Hartmann-Petersen R. Lynch syndrome, molecular mechanisms and variant classification. Br J Cancer. 2023;128:726–34.
23. Cagiada M, Bottaro S, Lindemose S, Schenstrøm SM, Stein A, Hartmann-Petersen R, et al. Discovering functionally important sites in proteins. Nat Commun. 2023;14:4175.
24. Gersing S, Cagiada M, Gebbia M, Gjesing AP, Coté AG, Seesankar G, et al. A comprehensive map of human glucokinase variant activity. Genome Biol. 2023;24:97.
25. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, et al. Mega-scale experimental analysis of protein folding stability in biology and design. Nature. 2023;620:434–44.
26. Abakarova M, Marquet C, Rera M, Rost B, Laine E. Alignment-based protein mutational landscape prediction: doing more with less. Genome Biol Evol. 2023;evad201.
27. Abakarova M, Freiberger MI, Lierhmann A, Rera M, Laine E. Drosophila melanogaster proteome-wide mutational landscape predictions with GEMME. 2024;
28. Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482:173–8.
29. Nunez JCB, Coronado-Zamora M, Gautier M, Kapun M, Steindl S, Ometto L, et al. Footprints of worldwide adaptation in structured populations of D. melanogaster through the expanded DEST 2.0 genomic resource. bioRxiv; 2024.
30. Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al. FlyBase: updates to the Drosophila genes and genomes database. Genetics. 2024;227:iyad211.
31. Oh H, Irvine KD. Yorkie: the final destination of Hippo signaling. Trends Cell Biol. 2010;20:410–7.
32. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting

effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40:W452.
33. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
34. Necci M, Piovesan D, Clementel D, Dosztányi Z, Tosatto SCE. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. Bioinforma Oxf Engl. 2021;36:5533–4.
35. Mesrouze Y, Bokhovchuk F, Meyerhofer M, Fontana P, Zimmermann C, Martin T, et al. Dissection of the interaction between the intrinsically disordered YAP protein and the transcription factor TEAD. Clarke J, editor. eLife. 2017;6:e25068.
36. Hu Y, Sopko R, Chung V, Foos M, Studer RA, Landry SD, et al. iProteinDB: An Integrative Database of Drosophila Post-translational Modifications. G3 Bethesda Md. 2019;9:1–11.
37. Kalodimou K, Stapountzi M, Vüllings N, Seib E, Klein T, Delidakis C. Separable Roles for Neur and Ubiquitin in Delta Signalling in the Drosophila CNS Lineages. Cells. 2023;12:2833.
38. Montero L, Okraine YV, Orlowski J, Matzkin S, Scarponi I, Miranda MV, et al. Conserved cysteine-switches for redox sensing operate in the cyclin-dependent kinase inhibitor p21(CIP/KIP) protein family. Free Radic Biol Med. 2024;224:494–505.
39. Carroll B, Otten EG, Manni D, Stefanatos R, Menzies FM, Smith GR, et al. Oxidation of SQSTM1/p62 mediates the link between redox state and protein homeostasis. Nat Commun. 2018;9:256.
40. Erdős G, Mészáros B, Reichmann D, Dosztányi Z. Large-Scale Analysis of Redox-Sensitive Conditionally Disordered Protein Regions Reveals Their Widespread Nature and Key Roles in High-Level Eukaryotic Processes. Proteomics. 2019;19:e1800070.
41. Kumar M, Michael S, Alvarado-Valverde J, Zeke A, Lazar T, Glavina J, et al. ELM—the Eukaryotic Linear Motif resource—2024 update. Nucleic Acids Res. 2024;52:D442–55.
42. Gray VE, Kumar S. Rampant Purifying Selection Conserves Positions with Posttranslational Modifications in Human Proteins. Mol Biol Evol. 2011;28:1565–8.
43. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. Trends Genet TIG. 2009;25:193–7.
44. Valanne S, Vesala L, Maasdorp MK, Salminen TS, Rämet M. The Drosophila Toll Pathway in Innate Immunity: from the Core Pathway toward Effector Functions. J Immunol. 2022;209:1817–25.
45. Luo H, Asha H, Kockel L, Parke T, Mlodzik M, Dearolf CR. The *Drosophila* Jak Kinase Hopscotch Is Required for Multiple Developmental Processes in the Eye. Dev Biol. 1999;213:432–41.
46. Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. The functional landscape of the human phosphoproteome. Nat Biotechnol. 2020;38:365–73.
47. Tricoire H, Battisti V, Trannoy S, Lasbleiz C, Pret A-M, Monnier V. The steroid hormone receptor EcR finely modulates Drosophila lifespan during adulthood in a sex-specific manner. Mech Ageing Dev. 2009;130:547–52.
48. Rera M, Monnier V, Tricoire H. Mitochondrial electron transport chain dysfunction during development does not extend lifespan in Drosophila melanogaster. Mech Ageing Dev. 2010;
49. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19:679–82.
50. GaussianMixture. Scikit-Learn.https://scikit-learn/stable/modules/generated/sklearn.mixture.GaussianMixture.html
51. Maidstone R, Hocking T, Rigaill G, Fearnhead P. On optimal multiple changepoint algorithms for large data. Stat Comput. 2017;27:519–33.
52. Rigaill G. fpopw: Weighted Segmentation using Functional Pruning and Optimal Partioning. 2022. https://CRAN.R-project.org/package=fpopw
53. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
54. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015; 31(6):926–932.

55. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20:473.
56. 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.
57. Hand DJ. Review of Mixture Models: Inference and Applications to Clustering. J R Stat Soc Ser C Appl Stat. 1989;38:384–5.
58. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol. 1977;39:1–38.
59. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
60. Requests: HTTP for Humans™ — Requests 2.32.3 documentation.
61. HALL P, KAY JW, TITTERINGTON DM. Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika. 1990;77:521–8.
62. Wojdyr M. GEMMI: A library for structural biology. J Open Source Softw. 2022;7:4200.
63. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures | Nucleic Acids Research | Oxford Academic.
64. Herrmann C, Van de Sande B, Potier D, Aerts S. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Res. 2012;40:e114.
65. Imrichová H, Hulselmans G, Kalender Atak Z, Potier D, Aerts S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. Nucleic Acids Res. 2015;43:W57–64.
66. DeLano WL, Scientific D, Carlos S. PyMOL: An Open-Source Molecular Graphics Tool.
67. 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.
68. Abakarova M, Freiberger MI, Lierhmann A, Rera M, Laine E. ProteoCast: https://proteocast.ijm.fr. 2025