

# DEGRONOPEDIA - a web server for proteome-wide inspection of degrons

Natalia A. Szulc<sup>1\*</sup>, Filip Stefaniak<sup>2</sup>, Małgorzata Piechota<sup>1</sup>, Andrea Cappannini<sup>2</sup>, Janusz M. Bujnicki<sup>2</sup>, Wojciech Pokrzywa<sup>1\*</sup>

<sup>1</sup> Laboratory of Protein Metabolism, International Institute of Molecular and Cell Biology in Warsaw, 4 Ks. Trojdena Str., 02-109 Warsaw, Poland

<sup>2</sup> Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, 4 Ks. Trojdena Str., 02-109 Warsaw, Poland

\* Correspondence should be directed to:

WP: [wpokrzywa@iimcb.gov.pl](mailto:wpokrzywa@iimcb.gov.pl)

NAS: [nszulc@iimcb.gov.pl](mailto:nszulc@iimcb.gov.pl)

## ABSTRACT

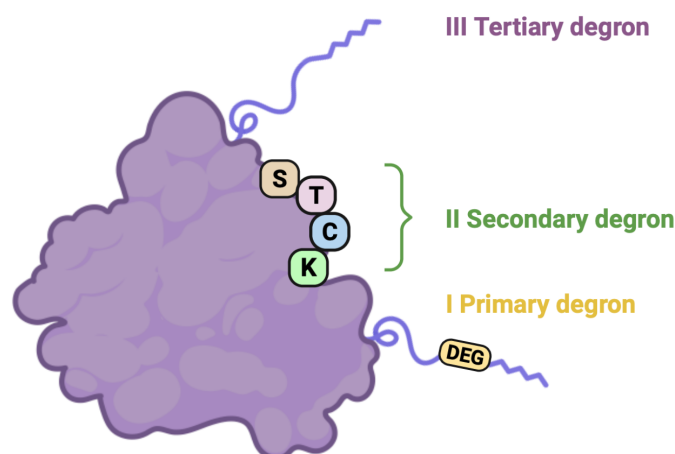
The ubiquitin-proteasome system is a proteolytic pathway that removes damaged and unwanted proteins. Their selective turnover is initiated by ubiquitin (Ub) attachment, mainly by Ub ligases that recognize substrates through their short linear motifs termed degrons. A degradation-targeting degron comprises a nearby Ub-modified residue and an intrinsically disordered region (IDR) involved in interaction with the proteasome. Degron-signaling has been studied over the last decades, yet there are no resources for systematic screening of degron sites to facilitate studies on their biological significance, such as targeted protein degradation approaches. To bridge this gap, we developed DEGRONOPEDIA, a web server that allows exploration of degron motifs in the proteomes of seven model organisms and maps these data to Lys, Cys, Thr, and Ser residues that can undergo ubiquitination and to IDRs proximal to them, both in sequence and structure. The server also reports the post-translational modifications and pathogenic mutations within the degron and its flanking regions, as these can modulate the degron's accessibility. Degrons often occur at the amino or carboxyl end of a protein substrate, acting as initiators of the N-/C-degron pathway, respectively. Therefore, since they may appear following the protease cleavage, DEGRONOPEDIA simulate sequence nicking based on experimental data and theoretical predictions and screen for emerging degron motifs. Moreover, we implemented machine learning to predict the stability of the N-/C-termini, facilitating the identification of substrates of the N-/C-degron pathways. We are confident that our tool will stimulate research on degron-signaling providing output information in a ready-to-validate context. DEGRONOPEDIA can be freely accessed at [degronopedia.com](https://degronopedia.com).

## INTRODUCTION

Cellular differentiation and development, stress conditions, and environmental factors constantly challenge the integrity of the proteome in every eukaryotic cell. Maintaining protein homeostasis (proteostasis) requires the degradation of damaged or unwanted proteins and plays a crucial role in cellular function, organismal growth, and, ultimately, cell and organism viability (Douglas & Dillin, 2010; Morimoto & Cuervo, 2014). The ubiquitin-proteasome system (UPS) is a principal proteolytic component of the cellular proteostasis network (Glickman & Ciechanover, 2002). Enzymes operating within the UPS recognize protein substrates destined for degradation and label them by attaching a small, evolutionarily conserved protein ubiquitin, primarily to internal lysine residues (Kerscher et al., 2006). It is noteworthy that a growing number of evidence indicates that cysteine and serine/threonine residues can also function as ubiquitination sites forming thioester or hydroxyester bonds with ubiquitin, respectively (Kravtsova-Ivantsiv & Ciechanover, 2012; McClellan et al., 2019). Ubiquitination is mediated by an enzymatic cascade involving a ubiquitin-activating enzyme (E1), which activates the C-terminal glycine of ubiquitin and then transfers it to a ubiquitin-conjugating enzyme (E2). Subsequently, with ubiquitin ligase (E3) participation, Ub is transferred on the lysine residue of the substrate protein, forming an isopeptide bond with it (Glickman & Ciechanover, 2002). The proteasome complex recognizes ubiquitinated proteins and, through proteolysis, degrades them into short peptides that can be further processed (Komander & Rape, 2012).

E3 substrates can be targeted for degradation by exposing peptide signal motifs, called degrons, involving a lysine (or multiple lysines) acting as a ubiquitination site (Ravid & Hochstrasser, 2008; Varshavsky, 2019). Degrons comprise mainly short linear motifs, several amino acids long, and are thought to occur preferentially in disordered regions of proteins. Degrons can be constitutive, promoting continuous protein degradation, or conditional, emerging after post-translational modifications such as phosphorylation (Holt, 2012) or after protease cleavage (Dissmeyer et al., 2018; Varshavsky, 2019). While degrons can be located anywhere in the protein sequence, those at the amino or carboxyl end, which are the initiators of the N- or C-degron pathway, respectively, have been the focus of extensive research over the last three decades (Bachmair et al., 1986; S.-J. Chen et al., 2017; Gonda et al., 1989; Hwang et al., 2010; Koren et al., 2018; Román-Hernández et al., 2009; Tasaki et al., 2005; Timms et al., 2019; Timms & Koren, 2020; Varshavsky, 2019; Yeh et al., 2021).

It is important to keep in mind that the recognition of a short linear motif by an E3 enzyme, followed by ubiquitination, may not be sufficient to lead to protein degradation. Guharoy *et al.* suggested that the short linear degron motif acts as a primary degron in the postulated tripartite degron architecture (Fig 1) (Guharoy et al., 2016). In this tripartite degron model, the secondary degron refers to lysine residues to which ubiquitin may be attached, and the tertiary degron indicates the flexible, intrinsically disordered region (IDR) in close proximity to the secondary degron, acting as a site to initiate protein unfolding prior to the entry into the proteasome. The secondary and tertiary degrons are suggested to play subsidiary roles that affect ubiquitin-signaling; the lack of a component of the tripartite degron model, e.g., an IDR near a ubiquitinated lysine, can result in non-proteolytic ubiquitination functions. Mutations leading to substitutions in degron motifs, as well as their secondary and tertiary degron sites, can therefore alter protein stability, contributing to diseases such as cancer and neurodegeneration (Eldeeb et al., 2022; Mészáros et al., 2017; Tokheim et al., 2021).



**Figure 1. The tripartite degron model.** The primary degron is a short linear motif recognized by the E3 ligase, localized preferentially within an IDR region of the protein. The secondary degron is a residue nearby the primary degron onto which ubiquitin transfer can occur (in our implementation, it is not only lysine (K) since ubiquitination can occur on cysteine (C), serine (S), or threonine (T)). The tertiary degron is an IDR close to the secondary degron, which acts as an unfolding seed initiating proteasome-dependent protein degradation. Modified from Guharoy et al., 2016, created in BioRender.com.

To better understand the specificity of the UPS and its deregulation in disease, degron sequences need to be discovered and matched with their respective degradation pathways. Recently, systematic approaches such as high-throughput experimental techniques, state-of-the-art proteomics technologies, and computational tools have been developed to understand the UPS selectivity (Coyaud et al., 2015; De Cesare et al., 2021; Kats et al., 2018; Koren et al., 2018; Nie et al., 2020; Timms et al., 2019; Wang et al., 2022; Yoshida et al., 2015). Biochemical and structural approaches complement the former to broaden our understanding of molecular mechanisms underpinning substrate selection by dedicated E3 ligases (X. Chen et al., 2021; Z. Chen et al., 2020; Chrustowicz et al., 2022; Yan et al., 2021). However, as the research on degron motifs and their physiological function is accelerating, there are few bioinformatics tools that would allow for degron motifs screening or analyzing. The anaphase-promoting complex/cyclosome (APC/C) degron repository (He et al., 2013) provides data on the sequence determinants of the three major classes of APC/C degron and overlays it with, e.g., disordered regions and post-translational modifications (PTMs). The eukaryotic linear motif (ELM) resource (Kumar et al., 2019) enables, among other functional sites, to detect 29 different degron motifs in the query protein and provides their structural context. A list of degron motifs, based on the data derived from the ELM resource and supplemented by manual curation, was released in 2017 as an interactive web table ([dosztanyi.web.elte.hu/CANCER/DEGRON/TP.html](https://dosztanyi.web.elte.hu/CANCER/DEGRON/TP.html), described in Mészáros et al., 2017). Finally, a deep learning model, deepDegron, was developed to predict degron disruption by mutations (Tokheim et al., 2021). However, deepDegron works as a standalone tool with specific input requirements - it accepts a list of mutations in a Mutation Annotation Format (MAF) file, containing particular columns with information on the gene of interest and its variants. To the best of our knowledge, no resource collects all known degron motifs and enables their systematic screening in query proteins, providing data on the degron site context, such as the postulated tripartite degron model.

Here we present DEGRONOPEDIA, the first web server designed to screen for known degron motifs in protein sequence or structure and provide their comprehensive sequence and spatial context complying with the tripartite degron mode. If the query protein comes from one of the selected model organisms - *A. thaliana*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, or *H. sapiens*, information on PTMs and known pathogenic mutations within the degron site and its neighborhood is also included. In addition, our tool allows the user to simulate the sequence cleavage by a number of different proteases to report N-/C-degron motifs in the newly emerged protein termini that may be of physiological interest. Our tool also allows users to examine their sequences and structures of interest for degron motifs. Moreover, DEGRONOPEDIA uses pre-trained machine learning (ML) models to predict the N-/C-terminal stability of the query protein. DEGRONOPEDIA is available free of charge in the form of a user-friendly web server at [degronopedia.com](https://degronopedia.com).

## MATERIALS AND METHODS

### Inputs

Three input types can be used for querying DEGRONOPEDIA: (i) a UniProt ID of a protein from the reference proteome (according to the UniProt database (UniProt Consortium, 2021)) of one of the selected model organisms - *A. thaliana*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, or *H. sapiens*, (ii) a protein sequence in the FASTA format, and (iii) a protein structure in the PDB format. The query protein must have between 50 and 8000 canonical amino acids regardless of input type.

**Query by PDB.** The submitted PDB file must not exceed 5MB size. Importantly, its B-factor column must carry either pLLDT (predicted Local Distance Difference Test; ranges 0-100; default for the AlphaFold models (Jumper et al., 2021)) or LLDT scores (Local Distance Difference Test; ranges 0-1; possible to obtain when predicting the model with RoseTTAFold (Baek et al., 2021), but requires further processing). The server does not support experimental PDB files since they lack long IDRs, which are crucial from the tripartite degron perspective. The detailed guide on the structure input type is available on the web server's Tutorial web page.

### Customizable parameters

Screening for known degron motifs is dynamically performed on the web server, considering the eight user-customizable parameters. In short, they describe thresholds related to the distinct regions of the tripartite degron architecture and the structural models regarding the minimum values to reckon residues as buried or disordered (Table 1). A detailed description of the parameters and the appropriate visualizations are available on the web server's Tutorial web page.

**Table 1.** Description of the eight user-customizable parameters available in the DEGRONOPEDIA.

Parameter	Unit	Default value	Allowed values	Description
<b>Primary degron-related</b>				
1. Degron flanking region in sequence	aa	20	5-40	Maximum sequence distance to regions upstream and downstream of a degron site to be considered flanking
2. Degron flanking region in structure	Å	20	5-40	Maximum structural distance to residues near a degron site to be considered flanking (such residues are not necessarily close in sequence to the degron site)
3. Region length to calculate degron disorder	aa	10	1-20	Maximum sequence distance to regions upstream and downstream of a degron site to be included in the degron mean disorder score based on pLLDT/LDDT values
<b>Secondary degron-related</b>				
4. Region length to calculate K/C/T/S disorder	aa	3	1-15	Maximum sequence distance to regions upstream and downstream of secondary degron (K/C/T/S) to be included in the secondary degron mean disorder score based on pLLDT/LDDT values
<b>Tertiary degron-related</b>				
5. Minimum IDR distance from K/C/T/S	aa	10	5-40	Minimum sequence distance of the secondary degron (K/C/T/S) to the continuous IDR region of defined length (see parameter 7) to consider it as a tertiary degron
<b>Structure-related</b>				
6. Minimum continuous IDR length	aa	10	5-40	Minimum number of subsequent (in sequence) disordered residues to be considered as IDR
7. pLDDT/LDDT disorder threshold	%	70	40-90	Minimum value to recognize a residue as disordered based on its pLLDT/LDDT score
8. Buried residue threshold	%	20	5-60	Minimum value to recognize the residue as buried based on its Relative Solvent Accessibility (RSA)

## Implementation

We describe the calculations workflow for the query by UniProt ID since it provides the most exhaustive result information. The other query types are processed identically, with the exception that they do not cover certain analyses available when querying by UniProt ID as they do not access any additional data, i.e., structure model if submitting a sequence, PTMs or mutations.

**General.** The queried protein from the selected reference proteome is screened for the presence of each degron motif (see Degron motifs in the Datasets section), considering the defined degron position separately; N-termini and C-termini degrons are matched to the beginning or end of the sequence, respectively, whereas degron motifs classified as internal are searched in the entire sequence. The Gravy hydrophobicity index (Kyte & Doolittle, 1982) of the first/last 15 amino acids of the queried protein is also calculated (Hickey et al., 2021; Kats et al., 2018). Upon data availability, the server provides the experimentally measured Protein Stability Index (PSI) of the N-/C-terminus and E3 ubiquitin ligases known to interact with the query protein (see N-/C-termini stability data and E3 interactome data in the Datasets section).

**Structural data.** Solvent-accessibility, location within an IDR region, and secondary structure are derived from the corresponding AlphaFold model. Of note, the AlphaFold models cover in its B-factor column pLDDT scores on a scale from 0 to 100, which estimate the accuracy of the modeled residues. Those with pLDDT above 70 are generally expected to be modeled well, while pLDDT below 70 correlates with disordered regions (Tunyasuvunakool et al., 2021). The server calculates (i) the IDR regions' positions (based on the pLDDT threshold as defined in parameter 7 (Table 1) and the IDR minimum continuous length as defined in parameter 6 (Table 1)), (ii) the secondary structure, and (iii) Relative Solvent Accessibility (RSA) of each residue (based on the threshold as defined in parameter 8 (Table 1)). The secondary structure and solvent accessibility are calculated using the mkdssp software (Joosten et al., 2011; Kabsch & Sander, 1983), with the latter being normalized to the RSA based on the Sander method (Rost & Sander, 1994). The server further maps these data on each found degron motif and calculates the degron's mean disorder (see parameter 3 in Table 1).

**PTMs and mutations.** As PTMs provide valuable information on potential degron modulation, the server maps the positions of known PTMs to each found degron motif and its flanking regions (regarding proximity in both sequence and structure, as defined in parameters 1 and 2 (Table 1), respectively). Moreover, the server also maps pathogenic missense mutations (only for the human proteins) to each degron motif and PTMs within its flanking regions.

**Tripartite degron model context.** The DEGRONOPEDIA server calculates the tripartite degron model for each of the found degron motifs. It searches for all potentially-ubiquitinated residues (in our implementation, these are not only lysines but also cysteines, serines, or threonines; K/C/T/S) within the degron flanking regions (regarding proximity in both sequence and structure, as defined in parameters 1 and 2 (Table 1), respectively) and maps their positions to the solvent accessibility and secondary structure data, location within an IDR, PTMs and pathogenic mutations; the server also calculates their mean disorder (see



parameter 4 in Table 1). Finally, our tool reports the closest IDR (as defined in parameter 5 in Table 1) and its distance (both in sequence and structure) to each of the aforementioned secondary degrons.

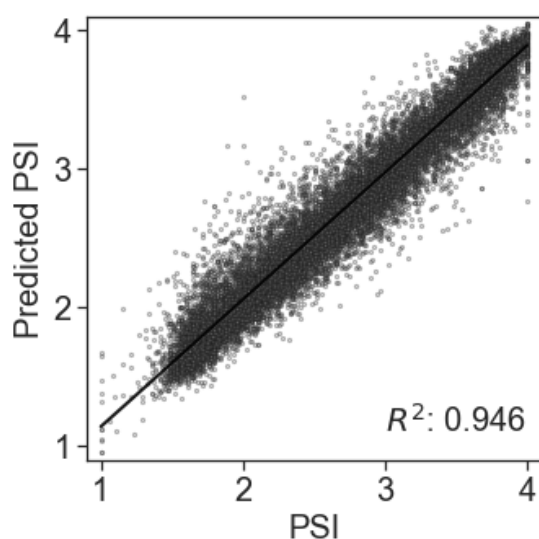
**Protein cleavage.** The DEGRONOPEDIA server simulates the cleavage of queried protein based on the experimentally-validated proteolytic sites derived from the MEROPS database as well as predictions of cleavage sites for 35 different proteolytic enzymes using the Pyteomics Python module (Goloborodko et al., 2013; Levitsky et al., 2019), which implements the cleavage prediction rules of the PeptideCutter ExPASy web server (Gasteiger et al., 2005). Next, our tool screens each newly emerged N-/C-terminus for known degron motifs as described before.

## N-/C-terminus stability predictions

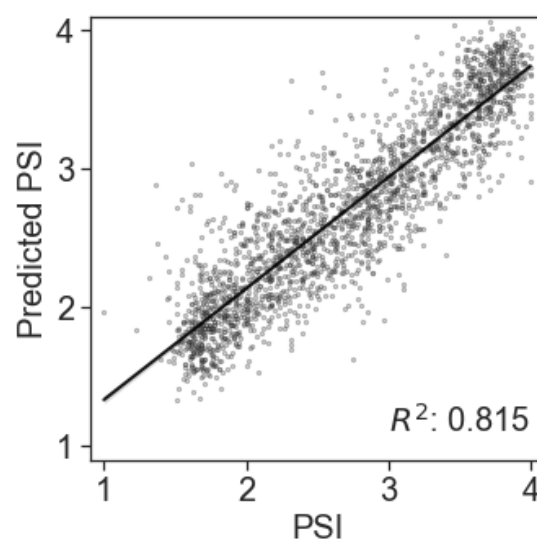
For a protein sequence or structure submission, the user may optionally request predictions of its N-/C-terminus stability using our pre-trained ML models.

**ML models development.** The experimental N-/C-termini stability data, expressed as Protein Stability Index (PSI) values measured for 23-mers covering N-/C-termini of the nearly complete human proteome (Koren et al., 2018; Timms et al., 2019), were used to develop the ML models. Of note, from the N-termini dataset, we considered only peptides' variants without the first methionine residue. Although methionine is the first amino acid incorporated into each new protein, it is not always the first amino acid in mature proteins undergoing post-translational removal (Varland et al., 2015; Yeom et al., 2017). Therefore, we decided to train an ML model to predict N-terminal stability without the initial methionine. The datasets were split into the training and testing set (in the ratio of 90:10), and each testing set remained untouched until the final testing of the models. Predictive models were built using the CatBoost regressor (Prokhorenkova et al., 2018) and trained on the aforementioned datasets, separately for each terminus. Hyperparameters of the CatBoost were optimized with the Optuna framework (Akiba et al., 2019) with five-fold cross-validation (using random permutations cross-validation implemented in scikit-learn Python library, with 20% validation set; (Pedregosa et al., 2012)). Descriptors used for building the models include the sequence of the peptide, RDKit descriptors (RDKit: Open-source cheminformatics; <http://www.rdkit.org>), Gravy hydrophobicity index (Kyte & Doolittle, 1982), and Peptides module (its Python version; [github.com/althonos/peptides.py](https://github.com/althonos/peptides.py); (Osorio et al., 2015)). All descriptors were calculated for the whole sequence and the first (excluding the N-terminal methionine) or the last (for the C-terminus) ten, eight, six, four, and two amino acids. The performance of the final models was evaluated using the testing set and an  $R^2$  coefficient, reaching the values of 0.815 for the C-terminus and 0.796 for the N-terminus (Fig 2). To visualize the predicted PSI N-/C-terminus value, the server maps it to the distribution of the corresponding experimental N-/C-termini stability dataset and classifies it as unstable/moderately unstable/average/moderately stable/stable, denoted by quantile thresholds of 0.2/0.4/0.6/0.8/1.0.

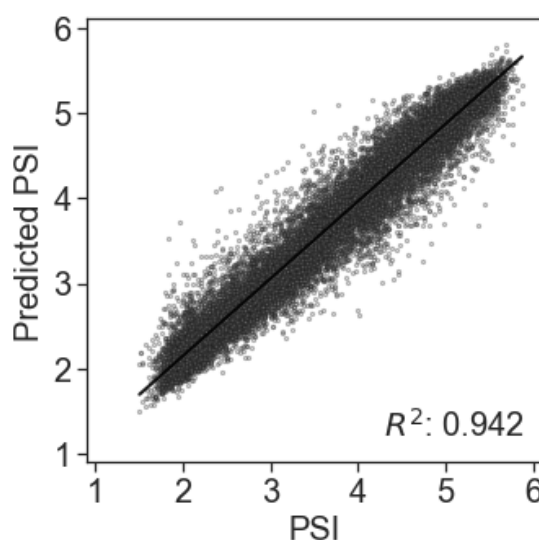
**A. . C-termini, training set**



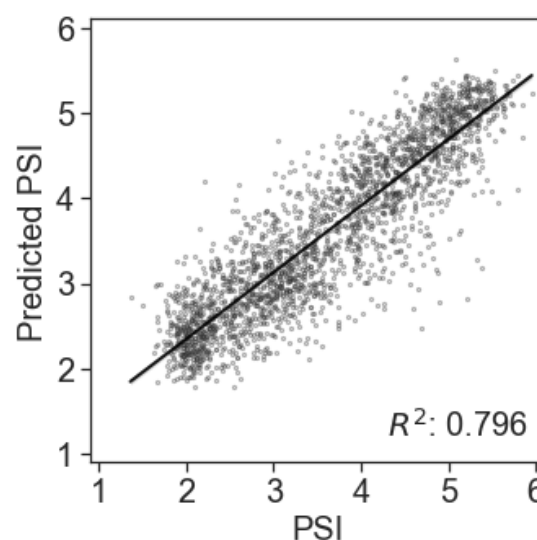
**B. . C-termini, testing set**



**C. . N-termini, training set**



**D. . N-termini, testing set**



**Figure 2. Predictions of the PSI using ML CatBoost regression models for C-termini (A, B) and N-termini (C, D); for the training (A, C) and testing set (B, D). Scatter plots show a regression line with a 95% confidence.**

## Outputs

The output information may be downloaded as a xlsx file, with corresponding data saved to separate sheets.

## Visualization

The Feature-Viewer tool (Paladin et al., 2020) was used for visualizations.



## Datasets

All the described datasets, except for the Degron motifs dataset, are applicable only when querying by UniProt ID.

**Degron motifs.** Over 400 degron motifs were obtained from the literature (X. Chen et al., 2021; Guharoy et al., 2016; Koren et al., 2018; Maurer et al., 2016; Timms et al., 2019; Varshavsky, 2019; Yan et al., 2021). Each motif was defined as either N-terminus, C-terminus, or internal, regarding its occurrence location. Moreover, additional data, including organisms in which the degron motif was found, degron type, known E3 ligases recognizing it, and subsidiary information, were added upon availability. N-terminus (Timms et al., 2019) and C-terminus (Koren et al., 2018) degron motifs derived from Global Protein Stability assays were selected based on their delta PSI (Protein Stability Index) value defined as >0.7 and >0.3, respectively.

**Reference proteome data.** Reference proteomes for *A. thaliana*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H. sapiens* were obtained from the UniProt database (IDs: UP000006548, UP000002311, UP000001940, UP000000803, UP000000437, UP000000589, UP0000005640, respectively). Sequences shorter than 50 and longer than 8000 amino acids were excluded from further analysis as very short peptides may not contain all the components of the tripartite degron model, and extremely long sequences are sporadic and require computationally exhaustive calculations. Where literature data were available, relevant information about each protein, i.e., its validated degron motifs or ubiquitinated lysines leading to proteasome-dependent degradation, was added.

**Structural data.** For each proteome as described above, its appropriate structural models were downloaded from the AlphaFold Protein Structure database (Varadi et al., 2022), excluding models for proteins exceeding the above-mentioned sequence length thresholds. As AlphaFold models of proteins longer than 2700 amino acids are split to separate files (containing overlapping fragments of the model), an in-house script utilizing the Biopython module (Cock et al., 2009) was used in each such case to superimpose overlapping residues, merge the structures' parts and save them to a single pdb file.

**N-/C-termini stability data.** The N-/C-termini stability data were obtained from the Global Protein Stability assays (Koren et al., 2018; Timms et al., 2019). These data covered the stability of N-/C-terminal 23-mers of 48251 (variants with N-terminal methionine and without it) and 22564 human proteins, respectively, represented as the Protein Stability Index (PSI), where its highest values (6 for the N-termini and 4 for the C-termini) indicate the most stable, thus possibly deprived of degron motifs, peptides. Consecutive mapping of N-/C-terminus PSI is performed by the server only for human proteins and based on the N-/C-terminal 23-mers identity, not the protein IDs. Similarly as in our ML models implementation (see N-/C-terminus stability predictions in Materials and methods), the N-termini peptides' variants without the first methionine residue were considered, and thus, the reported N-terminal PSI value refers to the stability of 23-mer N-terminus with absent methionine.

**Post-translational modification data.** The post-translational modification datasets were obtained from the iPTMNet (Huang et al., 2018) (phosphorylation, acetylation, ubiquitination,

methylation, N-Glycosylation, O-Glycosylation, C-Glycosylation, S-Glycosylation, sumoylation, myristoylation, S-Nitrosylation), PhosphoSitePlus (Hornbeck et al., 2015) (phosphorylation, acetylation, ubiquitination, methylation, O-Glycosylation, sumoylation) and the PLMD (Z. Liu et al., 2011, 2014; Xu et al., 2017) (neddylation and formylation) databases as well as from the literature, where we manually compiled datasets of non-canonical ubiquitination (Carvalho et al., 2007; Chua et al., 2019; Hensel et al., 2011; Ishikura et al., 2010; X. Liu & Subramani, 2013; Williams et al., 2007) and arginylation (Wong et al., 2007).








**E3 interactome data.** The complete interactomes were acquired from the BioGRID (Oughtred et al., 2021), IntAct (Orchard et al., 2014) and UbiNet 2.0 (Li et al., 2021) databases and from the literature, where we manually compiled a dataset of E3-substrate interactions absent in the aforementioned databases from (Koren et al., 2018; Oh et al., 2017; Weaver et al., 2017). As we provide information about the known interactions of the query protein with various E3 ligases (including components of their complexes), it was necessary to filter proteins from the selected model organisms annotated as E3 ligases. These annotations were obtained from the AMIGO web server (Ashburner et al., 2000; Carbon et al., 2009; Gene Ontology Consortium, 2021) by submitting the GO:0061630 query for each model organism and mapping the unique hits to their UniProt IDs. In the case of human E3 ligases, an additional annotation dataset was manually created, tabulating information from the AMIGO web server with ESBL (Medvar et al., 2016) and the UbiNet 2.0 data. Finally, the BioGRID and IntAct interactomes were accordingly filtered to derive only the E3-substrate interactions' sub-datasets (the UbiNet 2.0 already contained the E3-substrate interactions only).

**Mutation data.** A dataset of human mutations classified as pathogenic was obtained from the COSMIC database (Tate et al., 2019). Among the available mutation types, only 'Substitution - Missense' mutations were considered, as they have the least disruptive effect on the entire protein compared, e.g., to the frameshift or nonstop mutations, and may most severely impact the degron motif itself.

**Proteolytic cleavage sites data.** The experimental proteolytic cleavage sites with adherent information about the involved proteolytic enzymes were derived from the MEROPS database (Rawlings et al., 2018). The MEROPS dataset was subsequently filtered to contain only cleavage sites classified as physiologically relevant with present information about their exact position in the sequence.

## RESULTS

The DEGRONOPEDIA web server accepts three input types: (i) UniProt ID of a protein from the reference proteome of one of the selected model organisms - *A. thaliana*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, or *H. sapiens*, (ii) protein sequence in the FASTA format, and (iii) protein structure in the PDB format. Depending on the input type, different granularity of degron-related information is provided (Fig 3), with the most comprehensive data available for the UniProt ID query. However, regardless of the submitted query type, the protein sequence is always screened for the presence of over 400 degron motifs, which were obtained from the literature (X. Chen et al., 2021; Guharoy et al., 2016; Koren et al., 2018; Maurer et al., 2016; Timms et al., 2019; Varshavsky, 2019; Yan et al., 2021).

	Degron motifs occurrence 	Degron flanking regions 	Structural context 			PTMs 	Pathogenic mutations 	Cleavage simulations 		N-/C-terminus stability predictions 
			Solvent accessibility	Secondary structure	IDRs			Upon experimental sites	Upon predicted sites	
Query by UniProt ID	✓	Defined in sequence and structure	✓	✓	✓	✓	✓	✓	✓	✗
Query by sequence	✓	Defined in sequence	✗	✗	✗	✗	✗	✗	✓	✓
Query by structure	✓	Defined in sequence and structure	✓	✓	✓	✗	✗	✗	✓	✓

**Figure 3. Comparison of the result information obtained upon different query types in the DEGRONOPEDIA server.** Created in BioRender.com.

## Query by UniProt ID

The UniProt ID input type provides the most exhaustive information about the putative and experimentally-validated degron sites overlaid with additional information, i.e., solvent-accessibility, occurring PTMs or pathological mutations nearby. Of note, it is the only input type available in the DEGRONOPEDIA that provides experimentally-derived information on PTMs, mutations, physiologically-relevant proteolytic cleavage sites, and E3 interactors.

The queried protein is screened for the presence of degron motifs from the dataset prepared as described before. The Gravy hydrophobicity index (Kyte & Doolittle, 1982) of the N-/C-termini is also calculated, as hydrophobicity was shown to play an essential role in the N-/C-degron recognition by different E3 ubiquitin ligases (Hickey et al., 2021; Kats et al., 2018). The server also provides the information on N-/C-terminus stability (defined as Protein Stability Index (PSI); currently only for human proteins), collected from Global Protein Stability studies (Koren et al., 2018; Timms et al., 2019), and reports on E3 ubiquitin ligases interacting with the queried protein to provide further insights into possible degron-receptor interactions.

Since solvent-accessibility, location within an IDR region, or lack of secondary structure are important premises of a site acting as an actual degron, our tool derives such information from the corresponding AlphaFold model and overlays it on degron sites. The server also introduces degron flanking regions in sequence and structure to further extend the degron local context and maps the aforementioned features to these. PTMs are involved in multiple cellular processes, i.a., molecular interactions, protein folding, solubility, or signaling (Ramazi & Zahiri, 2021). Degron sites undergo various PTMs, with the primary role of phosphorylation (such degrons are often referred to as phosphodegrons), which can modulate their exposure (Holt, 2012). Hence, the server reports PTMs (up to 14 types) occurring within each found degron motif and its flanking regions. Amino acid substitutions in degron motifs can lead to altered protein stability, contributing to severe diseases such as cancer (Mészáros et al., 2017) or neurodegeneration (Eldeeb et al., 2022), indicating critical sites for proper protein function. Therefore, the server provides information about known

pathogenic missense mutations within the degron motifs and PTMs in their flanking regions.

For each of the found degron motifs (primary degron), its context to the secondary and tertiary degron is calculated according to the tripartite degron model postulated by Guharoy and colleagues (Fig 1) (Guharoy et al., 2016). In particular, the server provides information on solvent-accessibility, secondary structure, location within an IDR, mean disorder, PTMs, and pathogenic mutations for each potentially-ubiquitinated residue (secondary degron) located within the degron flanking regions. Finally, the server reports the closest IDR to each of the aforementioned secondary degrons.

It has been shown that protein turnover may be regulated by different proteolytic enzymes that cleave the protein, leading to new N- and C-termini which may act as degrons (Dissmeyer et al., 2018; Varshavsky, 2019). Therefore, the server simulates protein cleavage, after which it analyzes the newly emerged N- and C-termini for degron motifs.

All the results are reported in the form of comprehensive tables, where each column is described in detail on the web server's Tutorial web page. Localization of degron motifs, coils, buried residues, IDRs, PTMs, and pathogenic mutations are mapped to the query sequence and visualized using the Feature-Viewer tool (Paladin et al., 2020). All the output information may be downloaded as a xlsx file, with corresponding data appropriately sorted to separate sheets.

## Query by sequence

Submitting a sequence provides the most limited output data compared to two other input types available in the DEGRONOPEDIA since it carries the least information about the protein. Briefly, the degron motifs screening and cleavage simulations (upon predicted proteolytic sites) are performed identically as described for the UniProt ID query. For the tripartite degron model representation, only the secondary degrons (located within the degron flanking region in sequence) are reported since no structural data is provided.

However, an additional feature of this query type is the possibility to run ML models to predict the stability, expressed as the PSI, of the N-/C-terminus of the submitted protein sequence. The predicted PSI is visualized as a publication-ready figure, since the server maps each predicted PSI to the distribution of its corresponding experimental N-/C-termini stability dataset and classifies it as unstable/moderately unstable/average/moderately stable/stable. We recommend running the N-/C-termini stability predictions only on proteins from higher mammals, as our ML models were trained on stability datasets of human proteins (see N-/C-terminus stability predictions in Materials and methods).

As sequence input yields restricted calculations and, consequently, limited output information, we encourage users to query by structure and submit a model of their protein of interest from state-of-the-art structure prediction tools such as AlphaFold or RoseTTAFold.

## Query by structure

The structure input type was designed as a compromise between the granularity of the output information and the possibility of analyzing proteins other than from the reference proteomes of selected model organisms. All the calculations are performed identically as when querying by the UniProt ID, except no data on PTMs, mutations, physiologically-relevant proteolytic cleavage sites, or E3 interactors are provided. The user may run our ML models to predict the stability of the N-/C-terminus, identically as when querying by sequence.

## Example application

As a case study, we chose the human p53 tumor suppressor protein (UniProt ID P04637), for which numerous experimental data on turnover are available. Proteasomal degradation of p53 is based on degron sequences and post-translational modifications (Asher et al., 2005; Melvin et al., 2016; Yang et al., 2006). Within the p53 protein, DEGRONOPEDIA annotated six degrons (we ran the calculations with the default parameters), one as of the N-terminus pathway and the rest as internal degrons. These include the FSDLWKLL motif (positions 19-26; also defined more broadly as F<sup>[^P]{3}</sup>W<sup>[^P]{2,3}</sup>[VIL]), which is recognized by the E3 ubiquitin ligase MDM2 (Böttger et al., 1997; Kussie et al., 1996; Schon et al., 2002). In addition, the server reported the presence of two sites (positions 248-256 and 338-346) corresponding to motifs recognized by the anaphase-promoting complex (APC/C) E3 ubiquitin ligase, which, to our knowledge, has not been previously described as engaged in p53 degradation. Interestingly, the server annotated that the putative degron recognized by APC/C (positions 338-346) is frequently mutated, which may indicate its functionality. Its neighboring lysines (positions 291-292, 351 (<20Å) and positions 319-321, 357 (<20 aa)) are subject to several PTMs, including ubiquitination, and thus, could represent a secondary part of the degron. Fig 4 shows how DEGRONOPEDIA summarizes found degron sites, plots PSI values of N-/C-termini on the experimental data distribution, and assigns secondary/tertiary degrons in the p53 sequence and structure.





**Figure 4. Overview of the DEGRONOPEDIA server on the example of p53 tumor suppressor protein. (A)** overview panel with general information on the p53 protein (left) and degrons' summary panel with information on the number of found degnon motifs (right); **(B)** experimental PSI values of N-/C-terminus of p53 plotted on the distribution of experimental stability datasets; **(C)** visualization of the degnon motifs, structural data, PTMs and mutations on the protein sequence; **(D)** table summarizing found degnon motifs; **(E)** extract of table denoting the secondary and tertiary degrons associated with a putative degnon motif at positions 338-346.



## DISCUSSION

DEGRONOPEDIA is a web server that integrates multifaceted degron screening in a query protein, complying with the postulated tripartite degron model by Guharoy and colleagues (Guharoy et al., 2016), with an intuitive interface and convenient visualization of the analysis results. It accepts three input types: UniProt ID, the sequence in the FASTA format, and structure in the PDB format, providing different levels of the output information depending on the submitted query type, with the most comprehensive results available for the UniProt ID query. The operability of our web server allows intuitive checking for the presence of known degron motifs in combination with overlapping PTMs and pathogenic mutations (if available) and detection of new N-/C- pathway degron motifs after simulated proteolysis. Additionally, DEGRONOPEDIA can predict protein N-/C-termini stability based on the submitted sequence or structure, utilizing the pre-trained ML models.

However, DEGRONOPEDIA has some limitations. Regardless of the input type, the protein sequence must be between 50 and 8000 amino acids long (which roughly corresponds to the maximum 5MB file size when querying by structure) and should contain the 20 canonical residues only. Of note, the query by UniProt ID option is currently available only for proteins belonging to the reference proteomes of the seven most studied model organisms. Thus, users who would like to analyze a protein from another taxon, or an isoform not included in the reference proteome, have to choose between the query by sequence or query by structure options. Another limitation is that the submitted structure must be a monomer and should contain valid pLLDT or LDDT scores, as this information is used to extract the position of IDRs. Therefore, since experimentally obtained structures do not hold these values, they cannot serve as inputs to DEGRONOPEDIA. We reasoned that their lack of long IDRs does not make them the input of choice from the tripartite degron model perspective, especially in the light of the currently available state-of-the-art tools such as AlphaFold or RoseTTAFold which provide high-quality complete protein models. However, one must keep in mind that deriving information on the disorder residues based on the pLLDT-defined threshold may be inaccurate. Aderinwale and colleagues recently analyzed the correlation between disorder predictions obtained from different disorder prediction software tools and the regions from AlphaFold models with pLLDT scores below 0.5 and 0.7 (Aderinwale et al., 2022). Their results showed that only 30-50% of residues with low pLLDT scores correspond to disordered regions predicted using other methods, while the rest would adopt folded structures. Therefore, one must not treat the pLLDT scores as an absolute metric for defining IDRs. For this reason, we provided a customizable parameter so the user may manipulate the pLLDT threshold below which the residue is considered disordered to partially circumvent this issue.

In future versions of the DEGRONOPEDIA, we plan to further enhance the accuracy of the disorder region predictions by allowing the user to upload their own list of disordered residues as well as we will provide the option to predict IDRs using a state-of-the-art software tool to derive the disorder consensus in concert with the pLLDT/LDDT scores. In addition, we aim to extend the functionality and customizability of the web server by providing an option to define own degron motifs to screen for in the query protein. In addition, we will implement a new analysis type allowing for consecutive N-/C-terminus depletion, which would serve as a guide in the site-directed mutagenesis studies aiming to delete the degron site and, simultaneously, not create a novel one. As visualizations provide the most understandable and user-friendly way to report the result data, we also plan to map the elements of the degron tripartite model on a structure viewer, allowing for its

comprehensive inspection on the native protein surface. Last but not least, the detailed output provided by the DEGRONOPEDIA heavily depends on the literature data. Thus, we plan to integrate new datasets relevant to the degron-signaling upon newly published research on a rolling basis.

## DATA AVAILABILITY

The web server is available at [degronopedia.com](https://degronopedia.com). This website is free and open to all users, and there is no login required.

## ACKNOWLEDGMENTS

We would like to acknowledge Martina Bevilacqua for her invaluable support regarding the requested enhancements of the Feature-Viewer tool. We would like to express our gratitude to Dr. Natalia Gumińska for preparing the visual identification of the DEGRONOPEDIA. We would like to thank Dr. Neil Rawlings for providing exhaustive explanations on the MEROPS database usage. We would also like to acknowledge current members of the Pokrzywa group, in particular, Lilla Biriczová and Dr. Abhishek Dubey.

This research was carried out in part with the support of the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) at the University of Warsaw under computational allocation no G88-1177 to F.S.

## FUNDING

This research was supported by the National Science Centre, Poland (grant PRELUDIUM number 2021/41/N/NZ1/03473 to N.A.S; F.S. was supported by the National Science Center, Poland OPUS grant number 2020/39/B/NZ2/03127). A.C. was supported by the ROPES ITN grant from the European Commission [H2020-MSCA-ITN-2020, GA No. 956810, CA16120]; J.M.B. was supported by the National Science Center, Poland MAESTRO grant number 2017/26/A/NZ1/01083. W.P. was supported by the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund (grant POIR.04.04.00-00-5EAB/18-00).

## REFERENCES

- Aderinwale, T., Bharadwaj, V., Christoffer, C., Terashi, G., Zhang, Z., Jahandideh, R., Kagaya, Y., & Kihara, D. (2022). Real-time structure search and structure classification for AlphaFold protein models. *Communications Biology*, 5(1), 316. <https://doi.org/10.1038/s42003-022-03261-8>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Masanori, K. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, 2623–2631. <https://doi.org/10.48550/arXiv.1907.10902>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., &

- Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>
- Asher, G., Tsvetkov, P., Kahana, C., & Shaul, Y. (2005). A mechanism of ubiquitin-independent proteasomal degradation of the tumor suppressors p53 and p73. *Genes & Development*, 19(3), 316–321. <https://doi.org/10.1101/gad.319905>
- Bachmair, A., Finley, D., & Varshavsky, A. (1986). In Vivo Half-Life of a Protein Is a Function of Its Amino-Terminal Residue. *Science* (Vol. 234, Issue 4773, pp. 179–186). <https://doi.org/10.1126/science.3018930>
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Böttger, A., Böttger, V., Garcia-Echeverria, C., Chène, P., Hochkeppel, H. K., Sampson, W., Ang, K., Howard, S. F., Picksley, S. M., & Lane, D. P. (1997). Molecular characterization of the hdm2-p53 interaction. *Journal of Molecular Biology*, 269(5), 744–756. <https://doi.org/10.1006/jmbi.1997.1078>
- Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub, & Web Presence Working Group. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2), 288–289. <https://doi.org/10.1093/bioinformatics/btn615>
- Carvalho, A. F., Pinto, M. P., Grou, C. P., Alencastre, I. S., Fransen, M., Sá-Miranda, C., & Azevedo, J. E. (2007). Ubiquitination of mammalian Pex5p, the peroxisomal import receptor. *The Journal of Biological Chemistry*, 282(43), 31267–31272. <https://doi.org/10.1074/jbc.M706325200>
- Chen, S.-J., Wu, X., Wadas, B., Oh, J.-H., & Varshavsky, A. (2017). An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes. *Science*, 355(6323). <https://doi.org/10.1126/science.aal3655>
- Chen, X., Liao, S., Makaros, Y., Guo, Q., Zhu, Z., Krizelman, R., Dahan, K., Tu, X., Yao, X., Koren, I., & Xu, C. (2021). Molecular basis for arginine C-terminal degron recognition by Cul2 E3 ligase. *Nature Chemical Biology*, 17(3), 254–262. <https://doi.org/10.1038/s41589-020-00704-3>
- Chen, Z., Wasney, G. A., Picaud, S., Filippakopoulos, P., Vedadi, M., D'Angiolella, V., & Bullock, A. N. (2020). Identification of a PGXPP degron motif in dishevelled and structural basis for its binding to the E3 ligase KLHL12. *Open Biology* (Vol. 10, Issue 6, p. 200041). <https://doi.org/10.1098/rsob.200041>
- Chrutowicz, J., Sherpa, D., Teyra, J., Loke, M. S., Popowicz, G. M., Basquin, J., Sattler, M., Prabu, J. R., Sidhu, S. S., & Schulman, B. A. (2022). Multifaceted N-Degron Recognition and Ubiquitylation by GID/CTLH E3 Ligases. *Journal of Molecular Biology*, 434(2), 167347. <https://doi.org/10.1016/j.jmb.2021.167347>
- Chua, N. K., Hart-Smith, G., & Brown, A. J. (2019). Non-canonical ubiquitination of the cholesterol-regulated degron of squalene monooxygenase. *The Journal of Biological Chemistry*, 294(20), 8134–8147. <https://doi.org/10.1074/jbc.RA119.007798>
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>
- Coyaud, E., Mis, M., Laurent, E. M. N., Dunham, W. H., Couzens, A. L., Robitaille, M.,

- Gingras, A.-C., Angers, S., & Raught, B. (2015). BioID-based Identification of Skp Cullin F-box (SCF) $\beta$ -TrCP1/2 E3 Ligase Substrates. *Molecular & Cellular Proteomics: MCP*, 14(7), 1781–1795. <https://doi.org/10.1074/mcp.M114.045658>
- De Cesare, V., Carbajo Lopez, D., Mabbitt, P. D., Fletcher, A. J., Soetens, M., Antico, O., Wood, N. T., & Virdee, S. (2021). Deubiquitinating enzyme amino acid profiling reveals a class of ubiquitin esterases. *Proceedings of the National Academy of Sciences of the United States of America*, 118(4). <https://doi.org/10.1073/pnas.2006947118>
- Dissmeyer, N., Rivas, S., & Graciet, E. (2018). Life and death of proteins after protease cleavage: protein degradation by the N-end rule pathway. *The New Phytologist*, 218(3), 929–935. <https://doi.org/10.1111/nph.14619>
- Douglas, P. M., & Dillin, A. (2010). Protein homeostasis and aging in neurodegeneration. *Journal of Cell Biology* (Vol. 190, Issue 5, pp. 719–729). <https://doi.org/10.1083/jcb.201005144>
- Eldeeb, M. A., Ragheb, M. A., Soliman, M. H., & Fahlman, R. P. (2022). Regulation of Neurodegeneration-associated Protein Fragments by the N-degron Pathways. *Neurotoxicity Research*, 40(1), 298–318. <https://doi.org/10.1007/s12640-021-00396-0>
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., 'everine, Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. *The Proteomics Protocols Handbook* (pp. 571–607). <https://doi.org/10.1385/1-59259-890-0:571>
- Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1), D325–D334. <https://doi.org/10.1093/nar/gkaa1113>
- Glickman, M. H., & Ciechanover, A. (2002). The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction. *Physiological Reviews* (Vol. 82, Issue 2, pp. 373–428). <https://doi.org/10.1152/physrev.00027.2001>
- Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V., & Gorshkov, M. V. (2013). Pyteomics--a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *Journal of the American Society for Mass Spectrometry*, 24(2), 301–304. <https://doi.org/10.1007/s13361-012-0516-6>
- Gonda, D. K., Bachmair, A., Wüning, I., Tobias, J. W., Lane, W. S., & Varshavsky, A. (1989). Universality and structure of the N-end rule. *The Journal of Biological Chemistry*, 264(28), 16700–16712. <https://www.ncbi.nlm.nih.gov/pubmed/2506181>
- Guharoy, M., Bhowmick, P., Sallam, M., & Tompa, P. (2016). Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. *Nature Communications*, 7, 10239. <https://doi.org/10.1038/ncomms10239>
- He, J., Chao, W. C. H., Zhang, Z., Yang, J., Cronin, N., & Barford, D. (2013). Insights into degron recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. *Molecular Cell*, 50(5), 649–660. <https://doi.org/10.1016/j.molcel.2013.04.024>
- Hensel, A., Beck, S., El Magraoui, F., Platta, H. W., Girzalsky, W., & Erdmann, R. (2011). Cysteine-dependent ubiquitination of Pex18p is linked to cargo translocation across the peroxisomal membrane. *The Journal of Biological Chemistry*, 286(50), 43495–43505. <https://doi.org/10.1074/jbc.M111.286104>
- Hickey, C. M., Breckel, C., Zhang, M., Theune, W. C., & Hochstrasser, M. (2021). Protein quality control degron-containing substrates are differentially targeted in the cytoplasm and nucleus by ubiquitin ligases. *Genetics*, 217(1), 1–19. <https://doi.org/10.1093/genetics/iyaa031>
- Holt, L. J. (2012). Regulatory modules: Coupling protein stability to phopshoregulation during cell division. *FEBS Letters*, 586(17), 2773–2777.



- <https://doi.org/10.1016/j.febslet.2012.05.045>
- Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(Database issue), D512–D520. <https://doi.org/10.1093/nar/gku1267>
- Huang, H., Arighi, C. N., Ross, K. E., Ren, J., Li, G., Chen, S.-C., Wang, Q., Cowart, J., Vijay-Shanker, K., & Wu, C. H. (2018). iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Research*, 46(D1), D542–D550. <https://doi.org/10.1093/nar/gkx1104>
- Hwang, C.-S., Shemorry, A., & Varshavsky, A. (2010). N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, 327(5968), 973–977. <https://doi.org/10.1126/science.1183147>
- Ishikura, S., Weissman, A. M., & Bonifacino, J. S. (2010). Serine Residues in the Cytosolic Tail of the T-cell Antigen Receptor  $\alpha$ -Chain Mediate Ubiquitination and Endoplasmic Reticulum-associated Degradation of the Unassembled Protein. *Journal of Biological Chemistry* (Vol. 285, Issue 31, pp. 23916–23924). <https://doi.org/10.1074/jbc.m110.127936>
- Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39(Database issue), D411–D419. <https://doi.org/10.1093/nar/gkq1105>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637. <https://doi.org/10.1002/bip.360221211>
- Kats, I., Khmelinskii, A., Kschonsak, M., Huber, F., Knieß, R. A., Bartosik, A., & Knop, M. (2018). Mapping Degradation Signals and Pathways in a Eukaryotic N-terminome. *Molecular Cell*, 70(3), 488–501.e5. <https://doi.org/10.1016/j.molcel.2018.03.033>
- Kerscher, O., Felberbaum, R., & Hochstrasser, M. (2006). Modification of Proteins by Ubiquitin and Ubiquitin-Like Proteins. *Annual Review of Cell and Developmental Biology* (Vol. 22, Issue 1, pp. 159–180). <https://doi.org/10.1146/annurev.cellbio.22.010605.093503>
- Komander, D., & Rape, M. (2012). The Ubiquitin Code. *Annual Review of Biochemistry* (Vol. 81, Issue 1, pp. 203–229). <https://doi.org/10.1146/annurev-biochem-060310-170328>
- Koren, I., Timms, R. T., Kula, T., Xu, Q., Li, M. Z., & Elledge, S. J. (2018). The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. *Cell*, 173(7), 1622–1635.e14. <https://doi.org/10.1016/j.cell.2018.04.028>
- Kravtsova-Ivantsiv, Y., & Ciechanover, A. (2012). Non-canonical ubiquitin-based signals for proteasomal degradation. *Journal of Cell Science* (Vol. 125, Issue 3, pp. 539–548). <https://doi.org/10.1242/jcs.093567>
- Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J. A., Bukirova, D., Čalyševa, J., Palopoli, N., Davey, N. E., Chemes, L. B., & Gibson, T. J. (2019). ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz1030>
- Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., & Pavletich, N.

- N. P. (1996). Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* (Vol. 274, Issue 5289, pp. 948–953).  
<https://doi.org/10.1126/science.274.5289.948>
- Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1), 105–132.  
[https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
- Levitsky, L. I., Klein, J. A., Ivanov, M. V., & Gorshkov, M. V. (2019). Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *Journal of Proteome Research* (Vol. 18, Issue 2, pp. 709–714).  
<https://doi.org/10.1021/acs.jproteome.8b00717>
- Liu, X., & Subramani, S. (2013). Unique requirements for mono- and polyubiquitination of the peroxisomal targeting signal co-receptor, Pex20. *The Journal of Biological Chemistry*, 288(10), 7230–7240. <https://doi.org/10.1074/jbc.M112.424911>
- Liu, Z., Cao, J., Gao, X., Zhou, Y., Wen, L., Yang, X., Yao, X., Ren, J., & Xue, Y. (2011). CPLA 1.0: an integrated database of protein lysine acetylation. *Nucleic Acids Research* (Vol. 39, Issue suppl\_1, pp. D1029–D1034). <https://doi.org/10.1093/nar/gkq939>
- Liu, Z., Wang, Y., Gao, T., Pan, Z., Cheng, H., Yang, Q., Cheng, Z., Guo, A., Ren, J., & Xue, Y. (2014). CPLM: a database of protein lysine modifications. *Nucleic Acids Research*, 42(Database issue), D531–D536. <https://doi.org/10.1093/nar/gkt1093>
- Li, Z., Chen, S., Jhong, J.-H., Pang, Y., Huang, K.-Y., Li, S., & Lee, T.-Y. (2021). UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions. *Database* (Vol. 2021). <https://doi.org/10.1093/database/baab010>
- Maurer, M. J., Spear, E. D., Yu, A. T., Lee, E. J., Shahzad, S., & Michaelis, S. (2016). Degradation Signals for Ubiquitin-Proteasome Dependent Cytosolic Protein Quality Control (CytoQC) in Yeast. *G3*, 6(7), 1853–1866. <https://doi.org/10.1534/g3.116.027953>
- McClellan, A. J., Laugesen, S. H., & Ellgaard, L. (2019). Cellular functions and molecular mechanisms of non-lysine ubiquitination. *Open Biology* (Vol. 9, Issue 9, p. 190147).  
<https://doi.org/10.1098/rsob.190147>
- Medvar, B., Raghuram, V., Pisitkun, T., Sarkar, A., & Knepper, M. A. (2016). Comprehensive database of human E3 ubiquitin ligases: application to aquaporin-2 regulation. *Physiological Genomics*, 48(7), 502–512.  
<https://doi.org/10.1152/physiolgenomics.00031.2016>
- Melvin, A. T., Dumberger, L. D., Woss, G. S., Waters, M. L., & Allbritton, N. L. (2016). Identification of a p53-based portable degron based on the MDM2-p53 binding region. *The Analyst*, 141(2), 570–578. <https://doi.org/10.1039/c5an01429h>
- Mészáros, B., Kumar, M., Gibson, T. J., Uyar, B., & Dosztányi, Z. (2017). Degrons in cancer. *Science Signaling*, 10(470). <https://doi.org/10.1126/scisignal.aak9982>
- Morimoto, R. I., & Cuervo, A. M. (2014). Proteostasis and the Aging Proteome in Health and Disease. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* (Vol. 69, Issue Suppl 1, pp. S33–S38). <https://doi.org/10.1093/gerona/glu049>
- Nie, L., Wang, C., Li, N., Feng, X., Lee, N., Su, D., Tang, M., Yao, F., & Chen, J. (2020). Proteome-wide Analysis Reveals Substrates of E3 Ligase RNF146 Targeted for Degradation. *Molecular & Cellular Proteomics: MCP*, 19(12), 2015–2030.  
<https://doi.org/10.1074/mcp.RA120.002290>
- Oh, J.-H., Hyun, J.-Y., & Varshavsky, A. (2017). Control of Hsp90 chaperone and its clients by N-terminal acetylation and the N-end rule pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22), E4370–E4379.  
<https://doi.org/10.1073/pnas.1705898114>



- Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., ... Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(Database issue), D358–D363. <https://doi.org/10.1093/nar/gkt1115>
- Osorio, D., Rondón-Villarreal, P., & Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. *The R Journal* (Vol. 7, Issue 1, p. 4). <https://doi.org/10.32614/rj-2015-001>
- Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science: A Publication of the Protein Society*, 30(1), 187–200. <https://doi.org/10.1002/pro.3978>
- Paladin, L., Schaeffer, M., Gaudet, P., Zahn-Zabal, M., Michel, P.-A., Piovesan, D., Tosatto, S. C. E., & Bairoch, A. (2020). The Feature-Viewer: a visualization tool for positional annotations on a sequence. *Bioinformatics*, 36(10), 3244–3245. <https://doi.org/10.1093/bioinformatics/btaa055>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *arXiv*. <https://doi.org/10.48550/arxiv.1201.0490>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Andrey, G. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31. <https://doi.org/10.48550/arXiv.1706.09516>
- Ramazi, S., & Zahiri, J. (2021). Posttranslational modifications in proteins: resources, tools and prediction methods. *Database: The Journal of Biological Databases and Curation*, 2021. <https://doi.org/10.1093/database/baab012>
- Ravid, T., & Hochstrasser, M. (2008). Diversity of degradation signals in the ubiquitin–proteasome system. *Nature Reviews Molecular Cell Biology* (Vol. 9, Issue 9, pp. 679–689). <https://doi.org/10.1038/nrm2468>
- Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* (Vol. 46, Issue D1, pp. D624–D632). <https://doi.org/10.1093/nar/gkx1134>
- Román-Hernández, G., Grant, R. A., Sauer, R. T., & Baker, T. A. (2009). Molecular basis of substrate selection by the N-end rule adaptor protein ClpS. *Proceedings of the National Academy of Sciences of the United States of America*, 106(22), 8888–8893. <https://doi.org/10.1073/pnas.0903614106>
- Rost, B., & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20(3), 216–226. <https://doi.org/10.1002/prot.340200303>
- Schon, O., Friedler, A., Bycroft, M., Freund, S. M. V., & Fersht, A. R. (2002). Molecular Mechanism of the Interaction between MDM2 and p53. *Journal of Molecular Biology* (Vol. 323, Issue 3, pp. 491–501). [https://doi.org/10.1016/s0022-2836\(02\)00852-5](https://doi.org/10.1016/s0022-2836(02)00852-5)
- Tasaki, T., Mulder, L. C. F., Iwamatsu, A., Lee, M. J., Davydov, I. V., Varshavsky, A., Muesing, M., & Kwon, Y. T. (2005). A family of mammalian E3 ubiquitin ligases that contain the

- UBR box motif and recognize N-degrons. *Molecular and Cellular Biology*, 25(16), 7120–7136. <https://doi.org/10.1128/MCB.25.16.7120-7136.2005>
- Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., ... Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1), D941–D947. <https://doi.org/10.1093/nar/gky1015>
- Timms, R. T., & Koren, I. (2020). Tying up loose ends: the N-degron and C-degron pathways of protein degradation. *Biochemical Society Transactions*, 48(4), 1557–1567. <https://doi.org/10.1042/BST20191094>
- Timms, R. T., Zhang, Z., Rhee, D. Y., Harper, J. W., Koren, I., & Elledge, S. J. (2019). A glycine-specific N-degron pathway mediates the quality control of protein -myristoylation. *Science*, 365(6448). <https://doi.org/10.1126/science.aaw4912>
- Tokheim, C., Wang, X., Timms, R. T., Zhang, B., Mena, E. L., Wang, B., Chen, C., Ge, J., Chu, J., Zhang, W., Elledge, S. J., Brown, M., & Shirley Liu, X. (2021). Systematic characterization of mutations altering protein degradation in human cancers. *Molecular Cell* (Vol. 81, Issue 6, pp. 1292–1308.e11). <https://doi.org/10.1016/j.molcel.2021.01.020>
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., ... Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590–596. <https://doi.org/10.1038/s41586-021-03828-1>
- UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Židek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1), D439–D444. <https://doi.org/10.1093/nar/gkab1061>
- Varland, S., Osberg, C., & Arnesen, T. (2015). N-terminal modifications of cellular proteins: The enzymes involved, their substrate specificities and biological effects. *PROTEOMICS* (Vol. 15, Issue 14, pp. 2385–2401). <https://doi.org/10.1002/pmic.201400619>
- Varshavsky, A. (2019). N-degron and C-degron pathways of protein degradation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(2), 358–366. <https://doi.org/10.1073/pnas.1816596116>
- Wang, X., Li, Y., He, M., Kong, X., Jiang, P., Liu, X., Diao, L., Zhang, X., Li, H., Ling, X., Xia, S., Liu, Z., Liu, Y., Cui, C.-P., Wang, Y., Tang, L., Zhang, L., He, F., & Li, D. (2022). UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase–substrate interactions in eukaryotic species. *Nucleic Acids Research* (Vol. 50, Issue D1, pp. D719–D728). <https://doi.org/10.1093/nar/gkab962>
- Weaver, B. P., Weaver, Y. M., Mitani, S., & Han, M. (2017). Coupled Caspase and N-End Rule Ligase Activities Allow Recognition and Degradation of Pluripotency Factor LIN-28 during Non-Apoptotic Development. *Developmental Cell*, 41(6), 665–673.e6. <https://doi.org/10.1016/j.devcel.2017.05.013>
- Williams, C., van den Berg, M., Sprenger, R. R., & Distel, B. (2007). A conserved cysteine is

- essential for Pex4p-dependent ubiquitination of the peroxisomal import receptor Pex5p. *The Journal of Biological Chemistry*, 282(31), 22534–22543. <https://doi.org/10.1074/jbc.M702038200>
- Wong, C. C. L., Xu, T., Rai, R., Bailey, A. O., Yates, J. R., 3rd, Wolf, Y. I., Zebroski, H., & Kashina, A. (2007). Global analysis of posttranslational protein arginylation. *PLoS Biology*, 5(10), e258. <https://doi.org/10.1371/journal.pbio.0050258>
- Xu, H., Zhou, J., Lin, S., Deng, W., Zhang, Y., & Xue, Y. (2017). PLMD: An updated data resource of protein lysine modifications. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, 44(5), 243–250. <https://doi.org/10.1016/j.jgg.2017.03.007>
- Yang, W. H., Kim, J. E., Nam, H. W., Ju, J. W., Kim, H. S., Kim, Y. S., & Cho, J. W. (2006). Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. *Nature Cell Biology*, 8(10), 1074–1083. <https://doi.org/10.1038/ncb1470>
- Yan, X., Wang, X., Li, Y., Zhou, M., Li, Y., Song, L., Mi, W., Min, J., & Dong, C. (2021). Molecular basis for ubiquitin ligase CRL2-mediated recognition of C-degron. *Nature Chemical Biology*, 17(3), 263–271. <https://doi.org/10.1038/s41589-020-00703-4>
- Yeh, C.-W., Huang, W.-C., Hsu, P.-H., Yeh, K.-H., Wang, L.-C., Hsu, P. W.-C., Lin, H.-C., Chen, Y.-N., Chen, S.-C., Yeang, C.-H., & Yen, H.-C. S. (2021). The C-degron pathway eliminates mislocalized proteins and products of deubiquitinating enzymes. *The EMBO Journal*, 40(7), e105846. <https://doi.org/10.15252/embj.2020105846>
- Yeom, J., Ju, S., Choi, Y., Paek, E., & Lee, C. (2017). Comprehensive analysis of human protein N-termini enables assessment of various protein forms. *Scientific Reports*, 7(1), 6599. <https://doi.org/10.1038/s41598-017-06314-9>
- Yoshida, Y., Saeki, Y., Murakami, A., Kawawaki, J., Tsuchiya, H., Yoshihara, H., Shindo, M., & Tanaka, K. (2015). A comprehensive method for detecting ubiquitinated substrates using TR-TUBE. *Proceedings of the National Academy of Sciences* (Vol. 112, Issue 15, pp. 4630–4635). <https://doi.org/10.1073/pnas.1422313112>