MAVISp: A Modular Structure-Based Framework for Protein Variant Effects

Matteo Arnaudi[1,2†], Mattia Utichi[1,2†], Kristine Degn[1,2†], Ludovica Beltrame[1,2†], Simone Scrima[2], Karolina Krzesińska[1,2], Pablo Sánchez-Izquierdo Besora[1,2], Katrine Meldgård[1,2], Laura Bauer[2], Terézia Dorčaková[2], Anna Melidi[1,2], Lorenzo Favaro[1], Eleni Kiachaki[1,2], Anu Oswal[1,2], Alberte Heering Estad[1,2], Joachim Breitenstein[1,2], Jordan Safer[4], Paraskevi Saridaki[1], Francesca Maselli[2,3], Burcu Aykac Fas[5], Guglielmo Tedeschi[6], Philipp Becker[1], Jérémy Vinhas[2], Alberto Pettenella[2], Peter Wad Sackett[1], Claudia Cava[7], Anna Rohlin[8,9], Mef Nilbert[10,11], Sumaiya Iqbal[4], Matteo Lambrughi[2], Matteo Tiberti[2], Elena Papaleo[1,2]*

[1]Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, 2800, Lyngby, Denmark
[2]Cancer Structural Biology, Danish Cancer Institute, 2100, Copenhagen, Denmark
[3]Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Via F. Cervi 93, Segrate-Milan, 20054, Milan, Italy
[4]Bioinformatic and Computational Biology group, the Center for Development of Therapeutics, Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142
[5]Laboratoire de Biochimie Théorique, CNRS (UPR9080), Université Paris Cité, F-75005 Paris, France
[6]Department of Biochemistry and Microbiology, University of Chemistry and Technology Prague, Prague 6 166 28, Czech Republic
[7]Department of Science, Technology and Society, Scuola Universitaria IUSS, Istituto Universitario di Studio Superiori, Piazza della Vittoria 15, 27100 Pavia, Italy
[8]Department of Clinical Genetics and Genomics, Sahlgrenska university hospital, Gothenburg, Sweden
[9] Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[10]Institute of Clinical Medicine, Dept. Oncology and Pathology, Lund University, Sweden
[11] Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark

† These authors equally contributed to the work

* Corresponding author: Elena Papaleo, elpap@dtu.dk, elenap@cancer.dk

The role of genomic variants in disease has expanded significantly with the advent of advanced sequencing techniques. The rapid increase in identified genomic variants has led to many variants being classified as Variants of Uncertain Significance or as having conflicting evidence, posing challenges for their interpretation and characterization. Additionally, current methods for predicting pathogenic variants often lack insights into the underlying molecular mechanisms. Here, we introduce MAVISp (Multi-layered Assessment of Varlants by Structure for proteins), a modular structural framework for variant effects, accompanied by a web server (https://services.healthtech.dtu.dk/services/MAVISp-1.0/) to enhance data accessibility, consultation, and reusability. MAVISp currently provides data for 500 proteins, encompassing over four million variants. A team of biocurators regularly analyze and update protein entries using standardized workflows, incorporating free energy calculations or biomolecular simulations. We illustrate the utility of MAVISp through selected case studies. The framework facilitates the analysis of variant effects at the protein level and has the potential to advance the understanding and application of mutational data in disease research.

Keywords: variant effects, cancer genomics, protein structures, free energy calculations, protein stability, protein function, long-range structural communication

Introduction

We are witnessing unprecedented advances in cancer genomics, sequencing[1], structural biology[2], and high-throughput multiplex-based assays[3,4]. Consequently, Variants of Uncertain Significance (VUS) or conflicting classifications are continuously identified[5]. On the other hand, many variants with pathogenic or benign effects are reported in different repositories and compendia[6–9]. While sequencing approaches can identify alterations, understanding the molecular mechanisms and variant effects remains a challenge. The bioinformatics community has developed various approaches for predicting variant impact, many of which are benchmarked against experimental data and cellular readouts[10–14]. Furthermore, the advent of AlphaFold2[2,15,16] has enabled the prediction of accurate three-dimensional (3D) protein structures and complexes, often approaching experimental quality. In addition, the capability to perform long and accurate biomolecular simulations using GPU-based infrastructures and robust physical models allows the exploration of conformational changes and dynamics in protein structures across different timescales[17].

In previous pilot projects[18–20], we explored structure-based methods to analyze the impact of variants in coding regions of cancer-related genes, focusing on their consequences on the protein product. We propose that these methodologies could be widely applied to study disease-associated variants. When formalized and standardized, this approach can complement existing methods for predicting pathogenic variants, such as AlphaMissense[21]. Available pathogenicity predictors estimate the likelihood of variant pathogenicity but do not provide evidence of variant effects in relation to altered protein functions at the cellular level. The overall goal presented here is to link the effects of variants to specific underlying molecular mechanisms[20]. A mechanistic understanding of variant effects is essential knowledge for designing strategies in disease prevention, genetic counseling, clinical care, and drug treatments. From a fundamental research perspective, mechanistic knowledge is also essential for designing and prioritizing experiments to investigate underlying molecular details.

Considering this, we developed MAVISp (Multi-layered Assessment of VarIants by Structure for proteins) to enable high-throughput variant analysis within standardized workflows. MAVISp integrates results from pathogenicity predictors, gain/loss-of-function predictors, and structure-based analyses on variant effects on various protein properties. The data are accessible through a Streamlit-based website for consultation and download (https://services.healthtech.dtu.dk/services/MAVISp-1.0/). Additionally, we maintain a Gitbook resource with detailed reports for individual proteins (https://elelab.gitbook.io/mavisp/).

With this publication, we release the results of in silico saturation mutagenesis for all possible variants at each protein position for the first 500 proteins and over four million+ variants. New data and updates of existing entries will be released continuously. Currently, we are capable of process up to 20 new proteins weekly, which are deposited in a local version of the database. The public database is updated every two months. Based on recent statistics (https://elelab.gitbook.io/mavisp/documentation/coverage-and-statistics), we expect to provide 80-100 new proteins with each update, along with additional modules for existing entries. In this manuscript, we provide an overview of the methodology and show examples of data analysis and application for illustrative purposes.

Results

Overview of MAVISp and its database

MAVISp perform a set of independent predictions, each assessing the effect of a specific amino acid substitution on a different aspect of protein function and structural stability, starting from one or more protein structures. These independent predictions are executed by the so-called MAVISp modules (**Fig. 1a**). MAVISp can be applied to individual 3D protein structures and their complexes (*simple mode*) or to an ensemble of structures generated through various approaches (*ensemble mode*). The framework is modular, allowing all the modules or only a selected subset to be applied, depending on the case study. Each module relies on Snakemake, Dask workflows, or Python scripts, supported by specific virtual environments. The modules are divided into two main categories: (i) modules to retrieve and select structures for analyses (shown in orange in **Fig. 1a**) and (ii) modules to perform analyses related to variant assessment or annotations (shown in blue in **Fig. 1a**). Each module includes a strictly defined protocol for computational analysis that can be carried out either step by step or automatically as embedded in more comprehensive pipelines (Methods). They are designed to ensure consistency across all the proteins under investigation, along with to enhance reproducibility and repeatability.

The MAVISp workflow (**Fig. 1b**) is designed to enable multiple biocurators to work concurrently and independently on distinct proteins. A biocurator retrieves structural and functional information on a protein of interest, along with the basic input data for the workflows (e.g., gene name, UniProt AC, RefSeq identifier). Additionally, the biocurator identifies the initial structures for the calculations and simulations to be performed on the protein target in its free or bound state with other biomolecules. Subsequently, biocurators adhere to strict workflows for data based on the strict set of procedures that are codified in each module. Furthermore, they analyze the resulting aggregated csv files from the data managers (see below) and provide reports in Gitbook (https://elelab.gitbook.io/mavisp/) that can be exported as PDF files. Biocurators also act as reviewers for reports created by their peers. A review status is assigned to each GitBook entry to guide users regarding the quality and integrity of the curated data. To achieve this, we defined four review status levels (i.e. stars) for each protein entry(https://elelab.gitbook.io/mavisp/documentation/mavisp-review-status).

As mentioned above, after the data have been collected by the biocurators, the team of data manager will import and aggregate the data to a local version of the MAVISp database. The process results in a set of text files containing the final output for each protein target. The Python package mavisp.py parses the output data provided by the biocurators, processes them further, performs sanity checks and writes the results as human-readable CSV format, constituting the MAVISp database.

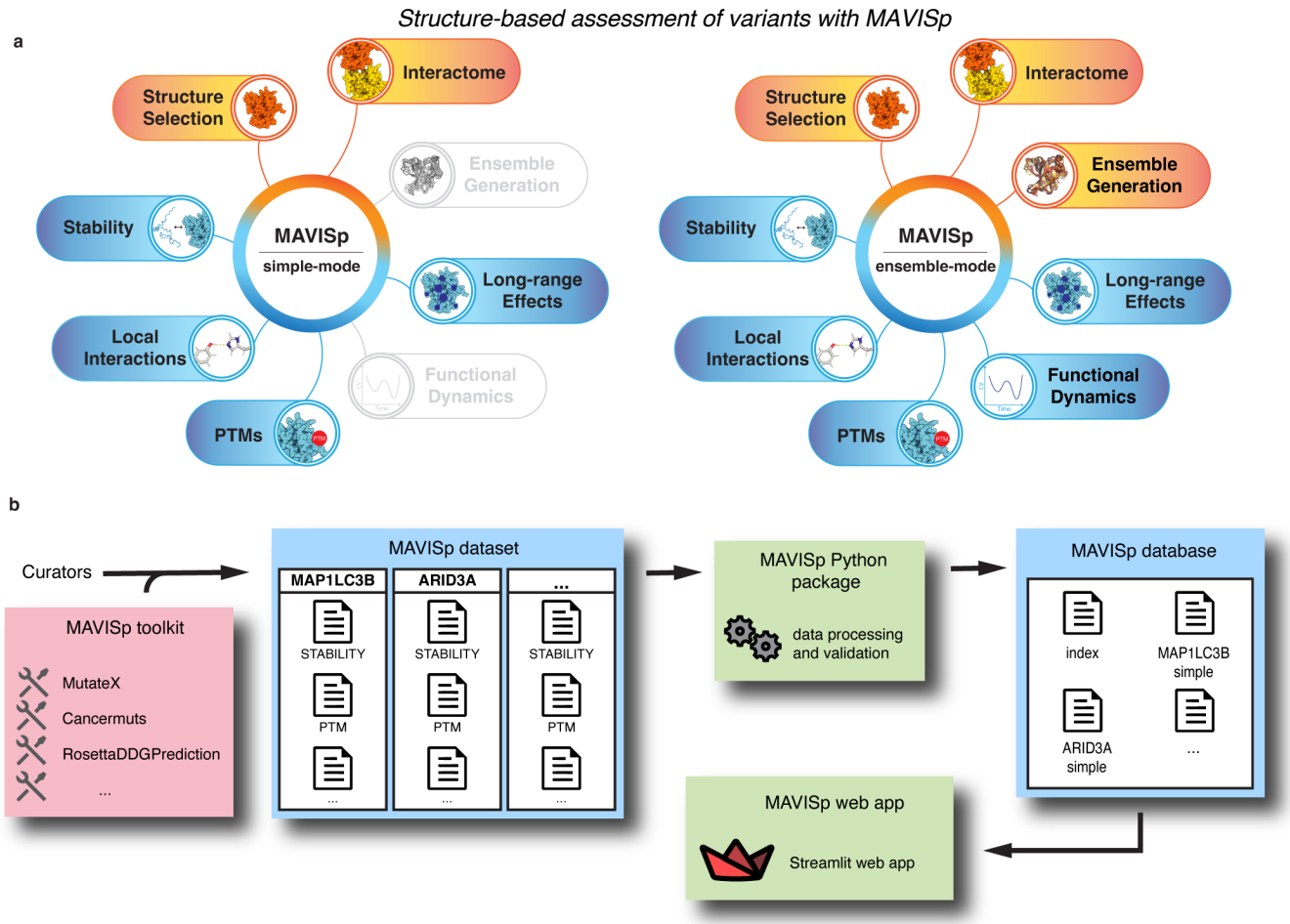

Fig. 1. Overview of MAVISp components. (a) MAVISp includes different modules, each managed by workflow engines or dedicated tools. The modules highlighted in orange handle the selection and collection of protein structures, while the modules in blue are dedicated to structural analyses of variant effects. The procedure begins with a gene name, its UniProt and RefSeq identifiers and the desired structural coverage. For each gene, all the steps can be conducted on a standard server with 32-64 CPUs. The only exceptions are: i) the ENSEMBLE GENERATION module, which includes all-atom MD simulations, and ii) Rosetta-based calculations on binding free energies and folding/unfolding free energy calculations. Depending on the simulation length and system size, these might require access to HPC facilities. On the left, the *simple mode* for the assessment is illustrated, which uses single experimental structures or models from AlphaFold2 or AlphaFold3. On the right, the *ensemble mode* is schematized in which a conformational ensemble for the target protein or its complexes is applied. Hereby, we consider a conformational ensemble a collection of 3D conformations of the protein generated by a sampling method such as molecular dynamics or provided by NMR structures in the PDB (b) Scheme of the current workflow for the MAVISp database and app. Biocurators apply specific workflows and protocols within each MAVISp module to generate structure-based predictions of changes linked to variants in each protein target. The results are gathered into a local text-based database. The data are further processed by the MAVISp Python package, which performs consistency checks, aggregate the data and outputs human-readable CSV table files, comprising the MAVISp database. These CSV files are imported by the Streamlit web app, powering the MAVISp webserver (https://services.healthtech.dtu.dk/services/MAVISp-1.0/), where the data are available for interactive visualization and download. In addition, we provide Gitbook reports for each protein to facilitate the interpretation of the results: https://elelab.gitbook.io/mavisp/.

Finally, the MAVISp web app interfaces with the database and present the data through a user-friendly Streamlit-based website (https://services.healthtech.dtu.dk/services/MAVISp-1.0/), where all the analytical results are accessible. The web app includes various visualizations to aid the interpretation of MAVISp results: (a) a dot plot displaying classifications for each mutation across MAVISp modules and collected pathogenicity predictors, (b) a lollipop plot aggregating MAVISp-identified effects for mutations categorized as 'pathogenic' by AlphaMissense, and (c) an interactive representation on the 3D structure, showing the localization of mutation sites identified in (b). These features are designed to simplify result interpretation and facilitate identification of mutations with specific mechanisms and multiple effects. The source code for the MAVISp Python package and the web application is available on GitHub (https://github.com/ELELAB/MAVISp), while the complete dataset can be downloaded from an OSF repository (https://osf.io/ufpzm/). Both source code and data are freely available and released under open-source or free licenses.

We invite requests on targets or variants that are not yet available in MAVISp or scheduled for curation. We also welcome contributors as biocurators or developers, pending training and adherence to our guidelines (https://elelab.gitbook.io/mavisp). Additionally, a full update, incorporating new versions of external tools or resources used by MAVISp, will be conducted annually to ensure that resources remain current. Moreover, we are continuously expanding the toolkit and developing new modules to enable even more comprehensive

assessments. The criteria for including new methods and approaches in the framework are detailed in the Gitbook documentation.

MAVISp modules for structure collection and selection

MAVISp includes three modules for structure collection and selection in the *ensemble mode,* and two for the *simple mode* (**Fig. 1a**).

The STRUCTURE_SELECTION module enables the user to identify the starting structure for their study. This module includes structure retrieval from the Protein Data Bank (PDB)[22], the AlphaFold Protein Structure Database[16], or through generation of initial models with AlphaFold3[23], AlphaFold2[2] and AlphaFold-multimer[15]. In addition, it streamlines the selection of the structures in terms of structural quality, experimental resolution, missing residues, amino acidic substitutions with respect to the UniProt reference sequence, and the AlphaFold per-residue confidence score (pLDDT), integrating tools such as PDBminer[24]. A programmatic access to AlphaFill[25] further assists in identifying cofactors to be included in the model structure or to identify mutation sites that should be flagged if in the proximity of a missing cofactor in the starting model. When necessary, a workflow is available to reconstruct missing residues or design linker regions to replace large, disordered linkers between protein domains (Methods).

The INTERACTOME module aids the identification of protein interactors for the target protein and their complex structures by querying the Mentha database[26], the PDB, and proteome-wide AlphaFold models validated experimentally[27].For predicted complexes, it retains models with reasonable accuracy based on metrics, such as pLDDT, Predicted Aligned Error (PAE), and pDOCKQ2 scores[28]. In addition, the module includes protocols based on AlphaFold [15,29] or comparative modeling[30,31] when the complex between the protein target and the interactor involved Short Linear Motifs (SLiMs).

The ENSEMBLE_GENERATION module allows using ensembles from different sources, such as NMR structures as deposited in PDB, coarse-grained models for protein flexibility (with CABS-flex[32]) or all-atom Molecular Dynamics (MD) simulations (with GROMACS[33] and PLUMED[34,35]) of the structure of the protein (or its complexes). The selection is based on the required accuracy for the generated ensemble and available computational resources. Once individual structures or structural ensembles for the protein candidate are selected – either alone or with interactors - the analysis modules can be used.

MAVISp modules for structural analysis

MAVISp integrates five main analyses modules for the ensemble mode and four for the simple mode (**Fig.1a**). The minimal set of data to import a protein target and its variants in the MAVISp database is the availability of the results from the STABILITY and PTM modules, along with the predictions for pathogenicity and loss/gain of function. The STABILITY module is devoted to estimate the effects of the variants on the protein structural stability using folding free energy calculations (Methods). This module takes advantage of recently developed workflows for high-throughput in-silico mutagenesis scans[36,37] and a newly implemented protocol for RaSP[38] (Methods). Once the free energy predictions have been collected, MAVISp applies a consensus approach to classify them. Evidence shows that variants with changes in stability below 3 kcal/mol are not featuring marked decreased stability at the cellular level [39,40]. Thus, MAVISp defines the following classes for changes in stability: stabilizing ($\Delta\Delta G <= -3$ kcal/mol with both methods, FoldX and Rosetta or RaSP), destabilizing ($\Delta\Delta G >= 3$ kcal/mol), neutral ($-2 =<= \Delta\Delta G <= 2$ kcal/mol), and uncertain ($2 <= abs(\Delta\Delta G) < 3$ kcal/mol). A mutation is also classified as uncertain if the two methods would classify the effect of the mutation differently. Since March 2024, we have adopted as a default for data collection the consensus between RaSP and FoldX, after performing a benchmark using the MAVISp datasets (**Supplementary Text S1** and https://github.com/ELELAB/MAVISp_RaSP_benchmark). RaSP provides a suitable solution for high-throughput data collection compared to the CPU-intensive scans based on Rosetta. In low-throughput studies, where we focus in detail on a target protein, we can include Rosetta data, which are more accurate but computationally demanding.

The LOCAL_INTERACTION module can be applied if the STRUCTURE_SELECTION and INTERACTOME modules identify at least a suitable structure of the complex between the target protein and another biomolecule. The LOCAL_INTERACTION module is based on the estimate of changes in binding free energy for variants at protein sites within 10 Å from the interaction interface using protocols and consensus strategies mirroring the ones for STABILITY (Methods).

Binding free energy thresholds are set based on the expected error margins of the predictors, approximately ±1 kcal/mol, as originally outlined by the developers. This approach addresses the scarcity of experimental

datasets on mutations impacting protein-protein interactions[41–43], which are often constrained by system heterogeneity, limited mutation numbers, or both, complicating reliable benchmarking. We classify a mutation as stabilizing (both methods predict ΔΔG <= -1 kcal/mol), neutral (-1 kcal/mol < ΔΔG < 1 kcal/mol) or destabilizing (ΔΔG >=1 kcal/mol). Cases in which the two methods disagree on the classification, or for which we do not have a prediction for both methods and the side chain relative solvent accessible area of the residue is >= 20%, are classified as uncertain. We also included support for LOCAL_INTERACTION for protein and DNA interactions, as well as for homodimers. Of note, a strength of our approach is to provide annotations for the effects of protein variants on different biological interfaces for the same target protein.

In the *ensemble mode*, the STABILITY and LOCAL_INTERACTION modules are used on ensembles of at least 20 structures from the simulations or on the three main representative structures upon clustering, depending on the free energy calculation scheme to apply. This approach is used to mitigate limitations due to lack of backbone flexibility when these free energy methods are applied to just one single 3D structure[20,36,44,45].

The LONG_RANGE module applies coarse-grained models to estimate allosteric free energy changes upon mutation based on Allosigma2[46]. Details on the parameters and steps for analysis are provided in the Methods and **Supplementary Text S2**. Variants are annotated as destabilizing (positive changes in allosteric free energy), stabilizing (negative changes in allosteric free energy), mixed effects (both conditions occur), or neutral if the mutation does not cause any significant change. Additionally, variants that do not cause a significant change in residue side-chain volume are annotated as uncertain. In the *ensemble mode*, we applied graph theory metrics based on changes in the shortest communication paths using atomic contact-based Protein Structure Network[47] . This analysis, combined with the AlloSigma2 data, allows pinpointing variants with long-range effects to functional sites or protein pockets that could serve as interfaces to recruit interactors or ligands.

The FUNCTIONAL_DYNAMICS includes the design and production of enhanced sampling simulations to further assess the local or long-range effects of a variant. As a first example, we applied this class of methods to validate the long-range effects predicted for p53 variants on the DNA-binding loops[20], which results are also deposited in the MAVISp database.

The PTM module currently supports phosphorylations, annotating the effect of variants at phosphorylatable sites. The PTM module evaluates how the loss or change of phosphorylation sites may impact protein regulation, stability, or interactions. To this goal, the module collects analyses and annotations such as solvent accessibility of the mutation site, inclusion of the site in phosphorylatable linear motif, comparison between predicted changes in folding/unfolding or binding free energy upon mutation or upon phosphorylation at the site of interest. In the module, we applied decision logic (**Supplementary Text S3**) to derive the classification for each variant as neutral, damaging, unknown effect, potentially damaging or uncertain. The identification of the phosphorylation sites in the PTM module is based on known experimental phosphosites and SLiMs, as retrieved by Cancermuts[48]. These data are complemented by a manually curated selection of phospho-modulated SLiMs (https://github.com/ELELAB/MAVISp/blob/main/mavisp/data/phospho-SLiMs_09062023.csv). For solvent-inaccessible phosphorylatable residues, the effects are classified as uncertain in the *simple mode*. In these cases, the *ensemble mode* is required to investigate wheatear a cryptic phosphorylated site may become accessible upon conformational changes[49,50]. Of note, the current version of the PTM module has been designed based on fundamental principles on how phosphorylation can affect the protein structure and should be used to identify variants for further investigation, particularly for experimental research, given the limited capacity to benchmark these computational predictions until we include more proteins in the database. We have in the queue for processing with MAVISp proteins from datasets relevant to the benchmark of the PTM module. These datasets include experimental data on protein stability and protein-protein interactions upon phosphorylation [51 52].

Additionally, MAVISp provides annotations for the variant interpretation reported in ClinVar[9], or calculated with REVEL[53], DeMask[54] , GEMME[11], EVE (Evolutionary model of variant effect)[55] , and AlphaMissense[21]. The results of the pathogenicity predictors (i.e., EVE, REVEL and AlphaMissense) or the predictors for loss/gain of function (DeMask and GEMME) can be combined with the results from the MAVISp modules to understand variant effects and to prioritize variants for other studies. MAVISp also includes annotations on pLDDT score, secondary structure, and side-chain solvent accessibility, which can assist in the interpretation of the results. Finally, we incorporated predictions on regions involved in early folding events.[56]

Sources of variants supported by MAVISp

We apply in silico saturation mutagenesis, which means that we provide predicted effects for each variant of a target protein at any position that has a structural coverage. Additionally, all variants reported for the target protein in COSMIC, cBioPortal, and ClinVar are annotated within MAVISp. We routinely update and maintain the entries in the MAVISp database to include up-to-date annotations using Cancermuts[48]. All Cancermuts annotations for MAVISp and other protein targets are also available at the Cancermuts webserver, https://services.healthtech.dtu.dk/services/Cancermuts-1.0/.

In addition, annotations from variant lists from other studies, such as data on cohort-based or nationwide studies or other disease-related genomic initiatives can be introduced. For example, we recently analyzed variants found in germline samples of childhood cancer patients[57] using one of the tools applied in the STABILITY module[37]. We included in the MAVISp database a full assessment for the proteins investigated in the previous study and retained the cohort-based annotations on the source of variants.

Currently, MAVISp includes data on four million+ variants from 500 proteins (at the date of 01/01/2024). An overview of the currently available data and how to use them for different research questions, analyses and variant selection are described in detail in the next Sections. Additional source of inspiration and more details can also be found in the first target studies where MAVISp has been applied to understand variants impact in rare genetic diseases[58] or involved in cancer hallmarks[59,60].As explained below, we have defined a priority list of targets that are analyzed in batches by the biocurators. Additional targets of interest for the research community can be requested, as explained in the documentation on Gitbook.

COSMIC Tumor Suppressor Genes and Oncogenes

We prioritized MAVISp data collection of known driver genes in cancer, i.e., tumor suppressors and oncogenes. To this goal, we analyzed the entire list of COSMIC Tumor Suppressor Genes (COSMIC v96), while the collection of the COSMIC Oncogene and Dual Role targets is ongoing. Furthermore, we have been including genes reported as a candidate driver by the Network of Cancer Genes (NGC)[61].

The MAVISp datasets on cancer driver genes can assist the identification of molecular mechanisms of predicted or known pathogenic variants in these genes, as well as to aid the characterization of Variants of Uncertain Significance (VUS). A recent example is the study we performed on BRCA2[60].In this study, we analyzed mutations reported in ClinVar for BRCA2, comparing the predictions form the STABILITY and LOCAL_INTERACTIONS modules of MAVISp with results from AVENGERS, a multiplex assay for variant impact on cell viability. The study resulted in 84 variants classified as non-functional by AVENGERS for which we could pinpoints effects on protein stability or binding to SEM1 with MAVISp.

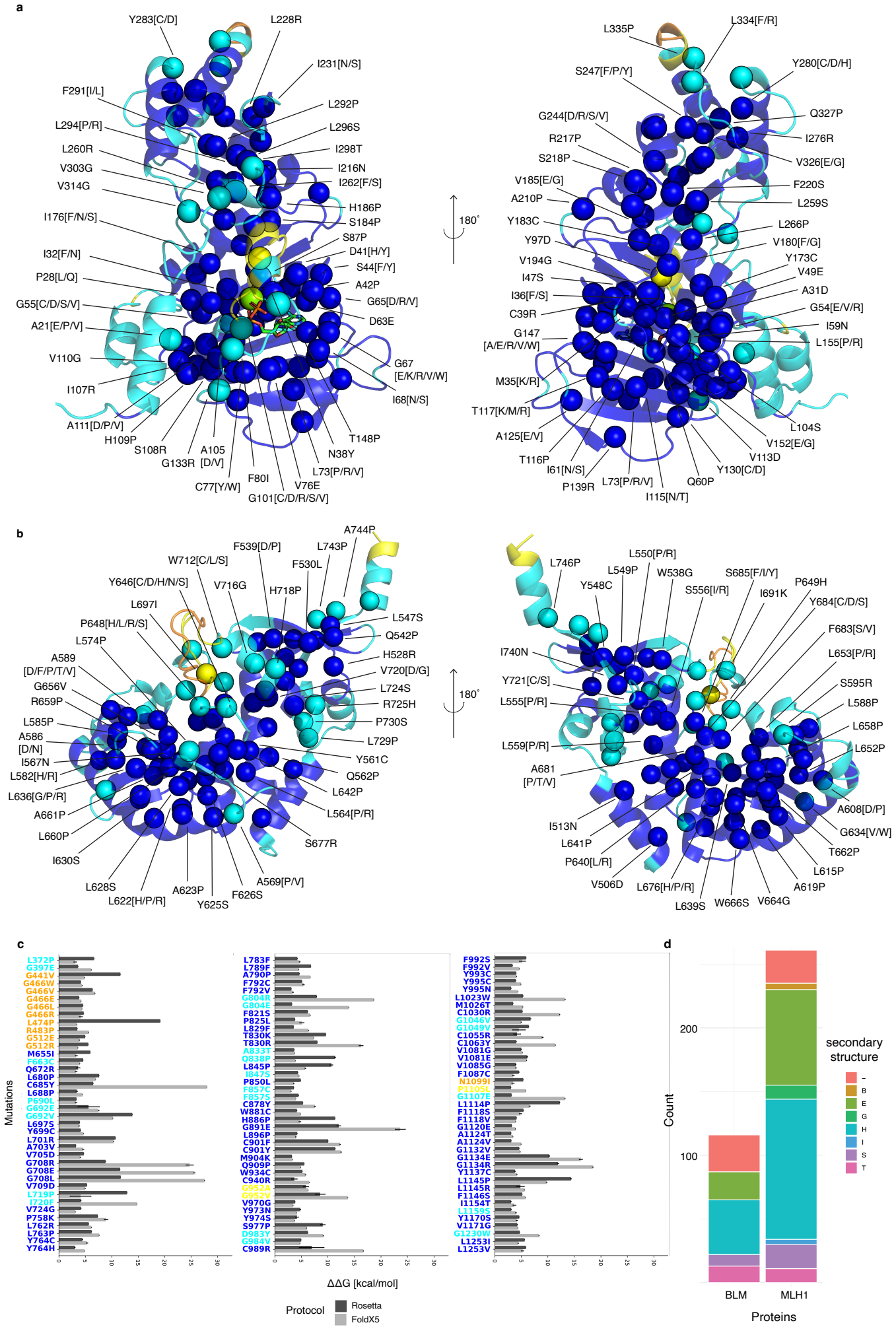
a
Y283[C/D]
L228R
I231[N/S]
F291[I/L]
L292P
L294[P/R]
L296S
L260R
I298T
V303G
I216N
V314G
I262[F/S]
H186P
I176[F/N/S]
S184P
S87P
I32[F/N]
D41[H/Y]
S44[F/Y]
P28[L/Q]
A42P
G55[C/D/S/V]
G65[D/R/V]
A21[E/P/V]
D63E
V110G
G67
[E/K/R/V/W]
I107R
I68[N/S]
A111[D/P/V]
H109P
T148P
S108R
A105
[D/V]
N38Y
G133R
L73[P/R/V]
F80I
C77[Y/W]
V76E
G101[C/D/R/S/V]
180˚
L335P
L334[F/R]
S247[F/P/Y]
Y280[C/D/H]
G244[D/R/S/V]
Q327P
R217P
I276R
S218P
V326[E/G]
V185[E/G]
F220S
A210P
L259S
Y183C
L266P
Y97D
V180[F/G]
V194G
Y173C
I47S
V49E
I36[F/S]
A31D
C39R
G54[E/V/R]
G147
[A/E/R/V/W]
I59N
L155[P/R]
M35[K/R]
T117[K/M/R]
L104S
A125[E/V]
V152[E/G]
T116P
V113D
I61[N/S]
Y130[C/D]
P139R
L73[P/R/V]
Q60P
I115[N/T]
b
A744P
F539[D/P]
L743P
W712[C/L/S]
F530L
Y646[C/D/H/N/S]
V716G
H718P
L697I
L547S
P648[H/L/R/S]
Q542P
L574P
H528R
A589
[D/F/P/T/V]
V720[D/G]
G656V
L724S
R659P
R725H
L585P
P730S
A586
[D/N]
L729P
I567N
Y561C
L582[H/R]
Q562P
L636[G/P/R]
L642P
A661P
L564[P/R]
L660P
S677R
I630S
L628S
A623P
A569[P/V]
L622[H/P/R]
Y625S
F626S
180˚
L746P
L550[P/R]
L549P
W538G
S685[F/I/Y]
Y548C
S556[I/R]
P649H
I691K
Y684[C/D/S]
F683[S/V]
I740N
L653[P/R]
Y721[C/S]
S595R
L555[P/R]
L588P
L658P
L559[P/R]
L652P
A681
[P/T/V]
A608[D/P]
I513N
G634[V/W]
L641P
T662P
P640[L/R]
L615P
V506D
L676[H/P/R]
A619P
L639S
W666S
V664G
c
Mutations
ΔΔG [kcal/mol]
Protocol
Rosetta
FoldX5
d
Count
secondary structure
BLM
MLH1
Proteins

Fig. 2. Variants with effects on structural stability in the tumour suppressors MLH1 and BLM. (a) Cartoon representation of $MLH1_{1-341}$ with spheres highlighting the C_α atom of the 84 positions where 145 mutations have been found annotated in the cBioPortal, COSMIC and ClinVar databases. The Mg^{2+} and ADP cofactors are shown as a green sphere and a stick, respectively. (b) Cartoon representation of $MLH1_{501-756}$ with spheres indicating the C_α atom of the 75 positions where 116 mutations have been found annotated in the cBioPortal, COSMIC and ClinVar databases. (c) The barplots illustrates the predicted ΔΔG values in kcal/mol of the 116 destabilizing mutations found in BLM. Both FoldX5 and Rosetta average ΔΔG values in kcal/mol are reported as in light grayblue and dark graylight blue bars, respectively, with standard deviation values shown as black winged lines. The ΔΔG values have been truncated at 25 kcal/mol for sake of clarity. The labels of the mutations are colored according to the AlphaFold pLDDT scores, i.e., blue - very high (pLDDT > 90), cyan - confident (70 < pLDDT <= 90), yellow - low (50 < pLDDT <= 70), and orange - very low (pLDDT <= 50). (d) Panel d illustrates the distribution stacked barplot showing the distribution of the MLH1 and BLM destabilizing mutations among secondary structure elements defined by the DSSP tool (i.e., H = α-helix, B = residue in isolated β-bridge, E = extended strand, participates in β ladder, G = 3-helix (3_{10} helix), I = 5-helix (π-helix), T = hydrogen bonded turn, S = bend, and "-" = no secondary structure identified). The results refers to the data available in the MAVISp database on 13th of July 2023.

In the case of tumor suppressors, it is especially interesting to identify variants that could cause loss of function, for example, by altering the structural stability of the protein or its function. As an example of loss-of-function due to changes in stability, we report the analysis of the MAVISp datasets of MLH1 and BLM. We identified 116 and 261 destabilizing variants for BLM and MLH1, respectively (**Fig. 2a-c**), of which more than half are reported in ClinVar. Of these, 90% of the destabilizing MLH1 mutations and 76% of BLM mutations occur in structured elements, while the remaining 10% and 24% respectively, are located in disordered regions (**Fig. 2d**). The results on MLH1 are consistent with previous analyses of disease variants performed using free energy calculations[40]. These results could be further investigated to select a subset of variants that have a predicted pathogenic impact from AlphaMissense and we predict damaging for stability for experimental validation. Depending on the size of the library to validate, one could decide if to perfom FACS sorting or cycloheximide chase assays[62,63] or use approaches based on multiplex technologies[64–67].

A useful feature of MAVISp is a module to curate and import experimental data with scores on the effects of the variants on biological readouts. These data can be compared and correlated with the structural properties we calculate, as a validation step to MAVISp predictions. For example, **Fig.3** illustrates the comparison with experimental data from a multiplex assay based on VampSEQ[64,68,69] that provides deep mutational scans (DMS) scores to study loss-of-function variants in PTEN (**Fig.3a**). We focused on 1690 variants for which we both have the MAVISp predictions on stability and the DMS scores[64] from the ProteinGym benchmarking dataset[70]. The consensus approach provided by the STABILITY module of MAVISp (accuracy 0.814) gives overall better performances in identifying variants that are found to be damaging in the DMS assay than the loss-of-function scores from GEMME or DeMasK (**Fig. 3b-c**). Nevertheless, this approach gives a lower sensitivity (i.e., 0.66) compared to GEMME. In the comparison, we applied the same threshold suggested for this dataset from the benchmarking dataset ProteinGym[70], which is based on the median value of the DMS scores. We thus wonder if the low sensitivity was due to cases with DMS scores too close to the median (**Fig. 3d**). However, we only found some of the cases of disagreement in this category. Additionally, in the original study for PTEN and as deposited in the MaveDB[71], a different classification for the variant scoring based on four abundance levels was proposed. We thus performed a comparison of the MAVISp results with the experimental dataset for the PTEN experiment from the MAVEDB using the abundance level classes as a threshold (**Fig. 3e**), resulting in increases sensitivity for the methods applied within MAVISp. The results on PTEN from MAVISp fits nicely with recent computational studies of PTEN variants using Rosetta calculations of protein stability and analyses of sequence conservation[13,14].

Overall, with the examples in this section, we illustrate different applications of the MAVISp results, spanning from benchmarking purposes, using the experimental data to link predicted functional effects with structural mechanisms or using experimental data to validate the predictions from the MAVISp modules.

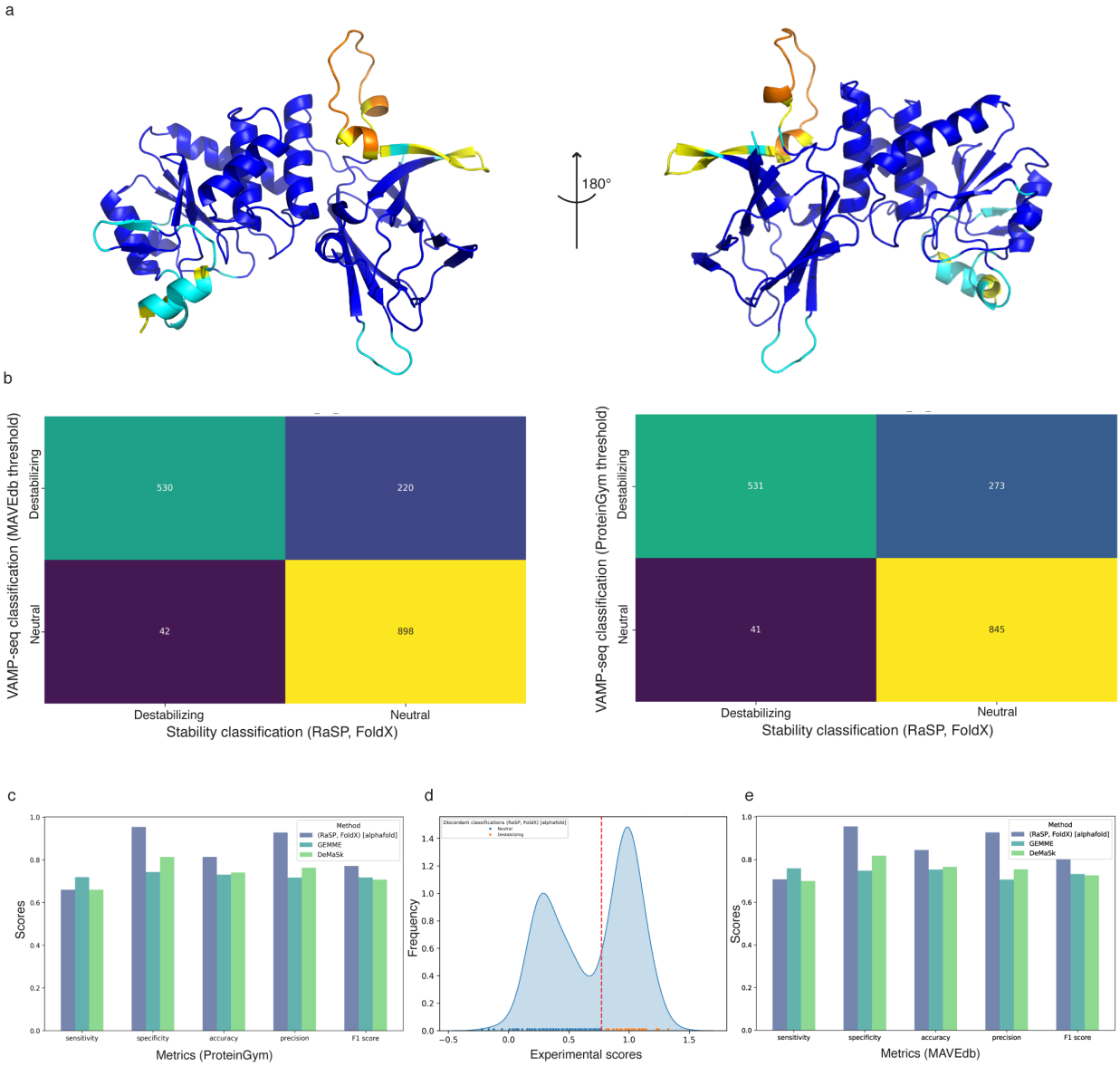

Fig. 3. Correlation of GEMME, DeMaSk, and MAVISp Stability Module predictions with protein abundance scores from VAMP-seq experiments.
(a) The trimmed AlphaFold structure (residues 1-351) of PTEN used for MAVISp stability module calculations is shown, colored according to pLDDT scores. (b) Confusion matrices comparing MAVISp stability module predictions (Destabilizing vs. Neutral) to protein abundance scores classified as Destabilizing or Neutral using two threshold approaches: ProteinGym and MAVE databases. (c and e) Performance metrics of MAVISp stability module, DeMaSk, and GEMME in predicting the effect of variants on protein function (GEMME, DeMaSk) or stability (MAVISp) using VAMP-seq scores with ProteinGym (c) and MAVE database (e) thresholds. (d) Distribution plot showing discordant classifications from the MAVISp stability module compared to experimental classifications, using the ProteinGym-defined threshold.

Transcription factors

Alterations in transcription factors are often linked to aberrant gene expression, including processes such as proliferation, cell death, and other cancer hallmarks[72]. Different mechanisms are at the base of alterations in the activity of transcription factors in cancer, including point mutations. Transcription factors are often multi-domain proteins and are involved in many protein-protein interactions. In previous works, we used part of the methods now integrated within MAVISp to study DNA-binding domains (ARID domains and p53[20,73]) or domains for protein-protein interaction (MZF1 and its SCAN domain[74]) from a mechanistic point of view. Additionally, the previous study on TP53 served as a platform to develop different modules currently available in MAVISp [20]. We aim to expand MAVISp data on transcription factor. To this goal, one of the high priority dataset under data collection is provided by the protein targets from the TRRUST2 database[75], which includes experimentally characterized transcription factors and their targets.

Proteins involved in cancer hallmarks

To expand the MAVISp contents, we are introducing also protein targets related to cancer hallmarks[76]. In this direction, we started to focus on proteins involved in cancer hallmarks related to protein clearance at the

cellular level, i.e. the ability to escape cell death (i.e., apoptosis and autophagy), and kinases or transcription factors involved in the regulation of cellular proliferation. Mitochondrial apoptosis is tightly regulated by a network of protein-protein interactions between pro-survival and pro-apoptotic proteins. We investigated the mutational landscape in cancer of this group of proteins in a previous study[77]. Here, we included structural analyses with different *simple-mode* MAVISp modules for both the pro-survival proteins BCL2, BCL2L1, BCL2L2, BCL2L10, MCL1 and BCL2A1, as well as the pro-apoptotic members of the family BOK, BAX and BAK1. In these analyses, the C-terminal transmembrane helix has been removed since the current version of our approach does not support transmembrane proteins or domains, illustrating an example on how we handle the STRUCTURE_SELECTION part.

Autophagy is a clearance mechanism with a dual role in cancer. The autophagy pathway relies on approximately 40 proteins, constituting the core autophagy machinery[78] (34 proteins currently included in the MAVISp database at the date of 01/11/2024 and the rest in production). As an example of the application of MAVISp to this group of proteins, we applied the *simple mode* to the markers of autophagosome formation MAP1LC3B and the central kinase ULK1, expanding on the knowledge provided by previous works[18,19].

In the case of ULK1, we included new structural coverage applying the AlphaFold models so that both the N-terminal (residues 7-279) and C-terminal domains (837-1046) have been used for the MAVISp assessment. ULK1 also serves as an example of how to customize the trimming of a model from AlphaFold to exclude disordered regions or linkers with low pLDDT scores in the *simple mode* of the assessment. In fact, their inclusion could lead to predictions of questionable quality. Disordered regions cannot be properly represented by one conformation and the *ensemble mode* would be necessary to derive more reliable conclusions.

ULK1 featured 215 variants reported in COSMIC, cBioPortal and/or ClinVar (**Fig. 4A**), illustrating an example of output provided by the downstream analysis tools of MAVISp. Using the *simple mode*, 59 among these had predicted long-range mixed effects. Furthermore, eight had a damaging effect on stability, one had a damaging PTM effect on regulation (S954N), one had a possible damaging PTM effect in function (S1042T), and four variants (L53P, G183V, E191G, and L215P) are characterized by both effects on stability and long-range communication (**Fig. 4B-D**). Most of the variants resulting in long-range and structural destabilizing effects are in the N-terminal kinase domain of the protein, suggesting that mutations on this protein could result in the inactivation of ULK1 by compromising its 3D architecture. We then performed a one-microsecond MD simulation of the ULK1 N-terminal kinase domain (residues 3-279, PDB ID: 5CI7) to generate a structural ensemble for the MAVISp *ensemble mode*. In this case, we applied an approach base on path analysis from a contact-based PSN (Methods), as provided by the LONG_RANGE module, which verified if structural long-range communication occurs between mutation and response sites predicted by the coarse grain model of the *simple mode*. The *ensemble mode* also offers further validation to the destabilizing variants predicted using a single input structure since it overcomes the limitations due to the lack or limited backbone flexibility with the free energy calculation scheme used by the framework. Overall, the application of the *ensemble* mode allowed to validate five variants with predicted long-range damaging effects (H72D, H72N, E73D, E73K, and R160L) and two variants with a damaging effect on stability (G183V and L215P). The predicted destabilizing (**Fig. 4E**) variant L215P has been also identified in samples from The Cancer Genome Atlas (TCGA) [18].

Structure-based assessment of variants with MAVISp

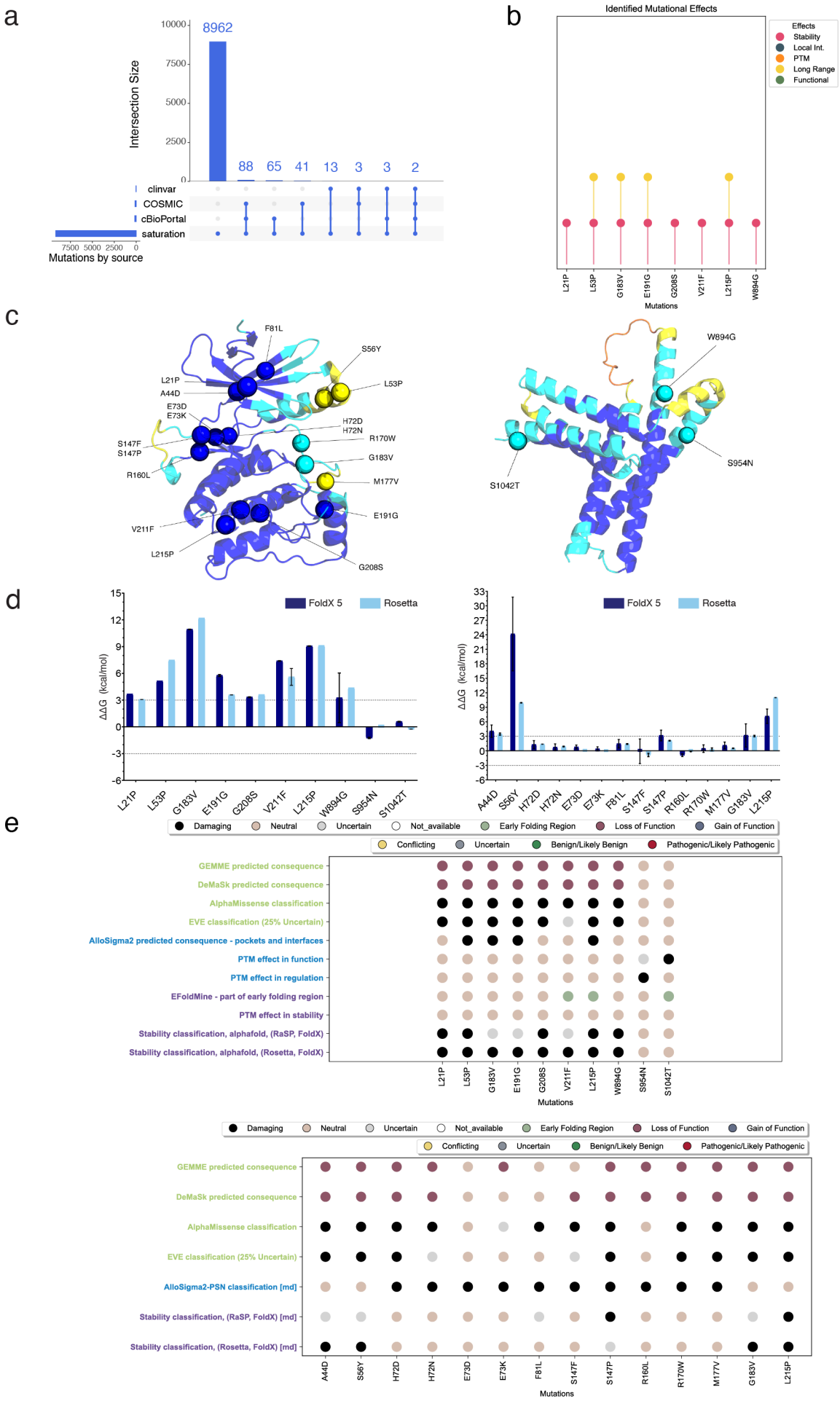

Fig. 4. MAVISp *ensemble mode* to identify damaging variants in the autophagy kinase ULK1. a) We examined the central autophagy kinase ULK1 using MAVISp, generating a saturation of all possible variants within the N-terminal (residues 7-279) and C-terminal domains (residues 837-1046), leading to a total of 8,962 variants. Of these, 215 variants have been identified in COSMIC, cBioPortal, and/or ClinVar databases. b) Among the ones reported in the previous databases, eight variants were reported as pathogenic by Alphamissense (L21P, L53P, G183V, E191G, G208S, V211F, L215P, and W894G) and among these, four variants are predicted to have a damaging effect on both protein stability and long-range communication (L53P, G183V, E191G, and L215P). c) Using MAVISp *simple* and *ensemble* modes, we identified 22 variants with destabilizing effects in terms of folding free energy, long-range effects, or PTM effects in regulation or in function. The mutation sites are highlighted with spheres on the AlphaFold models of the ULK1 N-terminal (left) and C-terminal (right) domains. d) We showed the predicted changes in folding free energy upon mutations for each of the 22 variants as calculated by the STABILITY module of MAVISp with MutateX and RosettaDDGPrediction with the *simple mode* (left) or with the *ensemble mode* (right). Interestingly, most of the variants that alter structural stability are located in the catalytic domain of the enzyme. This suggests potential mechanisms for ULK1 inactivation. e) Summary of the predicted effects on the 22 variants of ULK1 that have been found damaging with at least one MAVISp module with the *simple mode* (upper) or with the *ensemble mode* (lower).

The case of the autophagy marker MAP1LC3B provides an example on how the data for the LOCAL INTERACTION module can be obtained in a case of a protein that interacts with a functional motif embedded in intrinsically disordered proteins, i.e., a SLiM. In MAVISp, we report the results for the effects of variants in MAP1LC3B or in its binding partners using three prototypes for this interaction mode, i.e., the LC3-interacting region (LIR) of its binding partner SQSTM1 (**Fig.5a**), as well as ATG13 and optineurin. In this case, we first applied the protocols for (phospho)-SLiM identification developed within the MAVISp framework (Methods) and PDBMiner to identify possible starting structures. In the case of optineurin, we further model the flanking regions[59]. We identified ten variants reported in ClinVar as VUS (E102K, H86D, T29I, V91I, P2R, L44P, L44F, D56G, R11L, and E25Q, **Fig.5a**). In four of these, the MAVISp analyses have identified a mechanistic indicator altering the protein (**Fig.5b-d**). In details, T29I is predicted with changes due to lack of regulation by phosphorylation, L44P with effects on both structural stability and long-range effects to distal sites, L44F and R11L with long-range effects (**Fig 5b**). Additionally, a mutation found in cancer studies, P32Q, is predicted with a destabilizing effect on structural stability, confirming previous experimental results which showed propensity for aggregation[19]. Of note, this variant is identified with an uncertain prediction when a single structure is used for the analyses, whereas two different approaches for generating a conformational ensembles accounting for protein dynamics succeeded in predicting a destabilizing effect (**Fig 5d**). Additionally, all the variants with a mechanistic indicator from MAVISp are also predicted as pathogenic variants by AlphaMissense (**Fig.5c-d**) and are good candidates to further experimental studies for their effects on the autophagy flux or other functional readouts. E25Q and V91I are likely to be benign variants since all the predictors have identified neutral effects, both in terms of functionality and pathogenic potential (**Fig. 5c-d**).

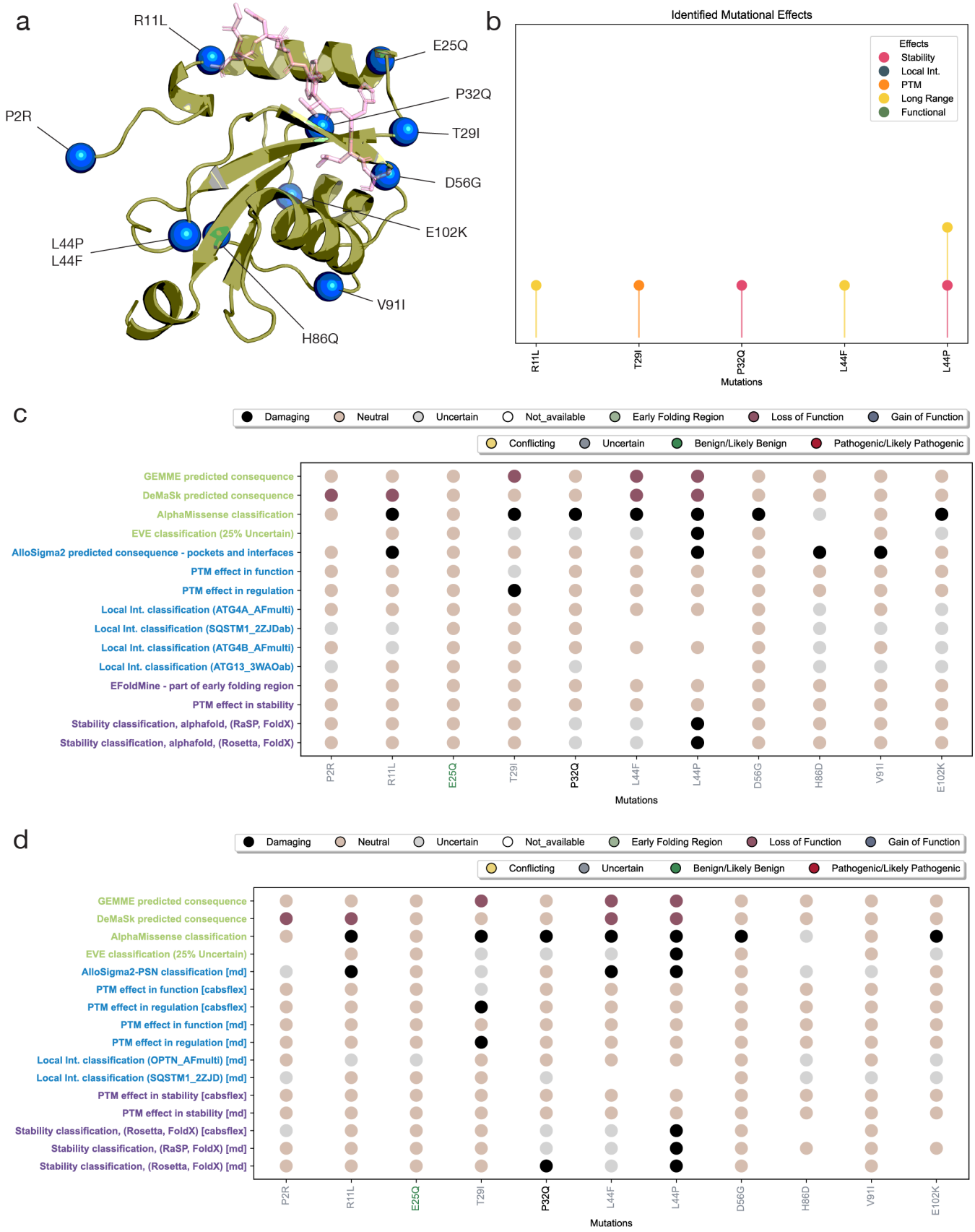

Fig. 5. Analysis of MAP1LC3B VUS Variants from ClinVar. a) A structural model (PDB ID: 2ZJD) of the MAP1LC3B (green) interaction with the LIR motif of SQSTM1(pink) highlights ten ClinVar-reported variants (E102K, H86D, T29I, V91I, P2R, L44P, L44F, D56G, R11L, and E25Q) along with the cancer-related variant P32Q. These variants are depicted as blue spheres on the structure. (b) Among these variants, five (R11L, T29I, P32Q, L44F and L44P) are reported as pathogenic by Alphamissense. Interestingly, L44P shows a predicted damaging effect on both long-range communication and stability.
(c-d) Summary of the predicted effects on the 11 variants of MAP1LC3B as reported by MAVISp dotplot with the simple mode (c) or with the ensemble mode (d).

Application of MAVISp to transmembrane proteins and to variants associated to other diseases

The STABILITY and LOCAL_INTERACTION modules do not support predictions for variants in transmembrane regions. A survey on methods to predict folding free energy changes induced by mutations on transmembrane proteins indicated the weakness of the existing protocols based on FoldX or Rosetta, which are suited soluble proteins[79]. For transmembrane protein, the current MAVISp protocol retains only the variants in the domains not in contact with the membrane. An example of this class is PILRA, where the transmembrane region is part of the AF model with low accuracy and has been trimmed, focusing on the analyses on the variants in the region 32-153. In addition, we included other transmembrane proteins in the database such as ATG9A and EGFR. PILRA is a protein target in connection with neurodegenerative diseases[80], along with KIF5A, CFAP410, and CYP2R1, illustrating the broad application of MAVISp to proteins related to different diseases. Proteins associated with other diseases have been also included, such as TTR and SOD1 and SMPD1. The latter has been recently investigated in a targeted study with the ensemble mode of MAVISp and other methodologies and compared with experimental data on the residual catalytic activity of the enzyme[58]. As stated above, MAVISp integrates curated experimental data for a target protein, which can be analyzed together with the results from the computational modules. To this goal, the dotplot representation provided by the downstream analyses toolkit of MAVISp and by the MAVISp database is handy to have a complete overview on the experimental and computational results (**Fig. 6a-c)** for SMPD1. Additionally, when a set of experimental data is available it is possible to correlate the predictions with the experimental data (**Fig. 6d-e**). For SMPD1, we have data on the residual catalytic activity of the enzyme for 135 variants[58]. Thanks to the MAVISp protocol, we collected data on the effects on changes in folding free energies and data for predicted loss-of-function/gain-of-function for these variants, which can be used to compare with the experimental data. Both the approaches supported by MAVISp for loss-of-function (i.e., GEMME and DeMasK) resulted in a correlation of 0.6 with the experimental data. Of note, most of the variants that have a predicted destabilizing effects on the stability are found at values of experimental residual activity lower than 20%, confirming what observed in our previous study[58] and suggesting that changes in stability for SMPD1 can be use as one of the proxies for damaging variants of this enzyme. In this case, the experimental readout cannot be explained by stability changes alone. Thus, variants found with low residual activity and a loss-of-function prediction (GEMME and DeMasK scores lower than -3 and -0.25, respectively) and that are neutral for stability according to MAVISp can be selected for further investigation. For example, they can be used to investigate the effects of these variants on the catalytic mechanism of the enzyme and its lipid transport using biomolecular simulations. Finally, variants, such as Y500H, which have a low residual activity, high loss-of-function scores and are uncertain for THE STABILITY module, can be analyzed for their propensity to fall in early folding regions (see entry in the MAVISp database) and could be investigated in the *ensemble mode* using enhanced sampling simulations to accurately estimate the folding free energy profiles.

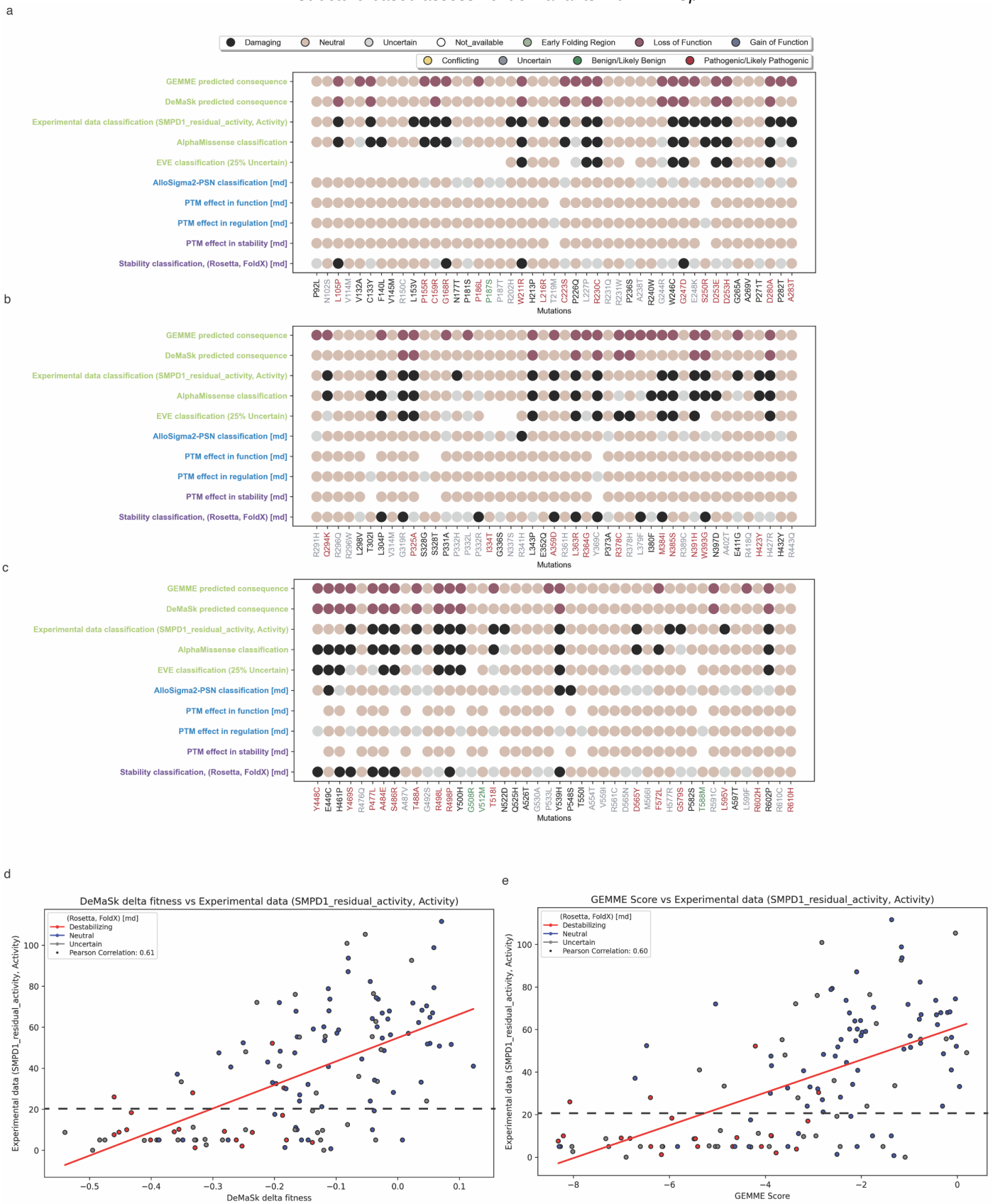

Fig. 6. MAVISp, GEMME, and DeMaSK predictions on the impact of SMPD1 variant subset. A subset of SMPD1 variants, for which experimental data on enzyme activity have been selected, is shown with predictions from MAVISp, GEMME, and DeMaSK. **a-c)** Panels a, b, c display dot plots of protein features analyzed by MAVISp, providing a classification into Damaging, Neutral, or Uncertain, alongside predictions from variant effect predictors (GEMME, DeMaSK) and pathogenicity classifiers (EVE, AlphaMissense). Variants are colored according to the classification in ClinVar. **d-e)** Panels d and e show scatter plots comparing DeMaSK (left) and GEMME (right) scores against experimental assay scores for enzymatic activity. The red line represents the regression, while the dotted line marks the threshold below which enzyme activity is considered inactive. Dots are colored based on the MAVISp STABILITY module classification (Rosetta/FoldX consensus): Destabilizing, Neutral, or Uncertain.

Conclusions and Future Perspective

MAVISp provides a multi-layered assessment of the effects of variants found in cancer studies or other diseases using structural methods. MAVISp results are especially useful for variant interpretation and prioritization. The results can be seen as a complementary resource to available pathogenic scores or high-throughput experiments. MAVISp can help to pinpoint the effects linked to a pathogenic mutation for further studies.

A significant advantage of MAVISp is its comprehensive coverage, expanding beyond clinically identified variants, by including novel variants yet to be characterized in other databases and tools. This makes MAVISp a valuable resource for researchers and clinicians, facilitating the exploration of novel variants and their underlying pathogenic mechanisms. Finally, we envision that MAVISp could become, in time, a community-driven effort and serve as a repository of data for the effects of disease-related variants more broadly. The results reported in MAVISp will provide an atlas of functional annotations for disease-related variants. MAVISp can help on one side to associated mechanistic indicators to variants that are known or predicted pathogenic, as well as to aid in the characterization of the effects of VUS or variants with conflict evidence at the molecular level.

As the database grows, it will provide high quality data on different structural properties that can also be used for benchmarking purposes or to be used as features in machine learning models. To this goal, the stringent data collection that we designed and present here is pivotal to produce a handful of data to build meaningful predictive models.

Additionally, we showed the versatility of the usage of MAVISp in a study where we characterized the effects induced by a redox post-translational modification (*S*-nitrosylation) with structural methods[81] . We focused on mutations found in cancer samples for their capability to alter the propensity of the cysteine to be *S*-nitrosylated or a population-shift mechanism induced by the PTM. The collection modules of MAVISp have been pivotal to aggregate variants for each target of interest in the study on *S*-nitrosylation. This study showcases the versatility of usage of MAVISp thanks to its modular design. Moreover, the pipelines developed in the study of *S*-nitrosylation provide a solid ground for future integration with MAVISp to go beyond phosphorylations with the PTM module.

Furthermore, MAVISp provides pre-calculated values of changes in folding or binding free energies and other metrics that can also be reanalyzed in the context of other research projects. With the examples on PTEN and SMPD1 provided here, we introduced the curation of experimental data in MAVISp, as a source of experimental validation. MAVISp supports either data from multiplex assays of variant effects (MAVE) or experimental data from literature mining of the biocurators. The purpose of collecting experimental data is to validate our findings, update protocols, and continuously improve the included methodologies. Trusting predictive data in this context depends on its alignment with experimental results, which helps refine the accuracy and reliability of predictions over time.

At this stage, it is possible to provide MAVISp annotations for variants transmembrane proteins in the domains not in contact with the membrane. Approaches as the one recently published [82] could broaden the possibility of applying the STABILITY module to transmembrane proteins. In addition, we will include support to intrinsically disordered regions in the ensemble mode, designing new modules to reflect the most important properties of these regions.

We foresee that MAVISp will provide a large amount of data on structure-based properties related to the changes that can exert at the protein level, which could be exploited for design of experimental biological readouts, also towards machine-learning applications for variant assessment and classification or to understand the importance of specific variants in connection with clinical variables, such as drug resistance, risk of relapse and more.

Methods

Initial structures for MAVISp

As a default, in the high-throughput data collection, we use models from the AlphaFold2 database[16] for most of the target proteins and trimmed them to remove regions with pLDDT scores < 70 at the N- or C-termini or very long disordered linkers between folded domains. For proteins coordinating cofactors, in the low-throughput targeted studies, we re-modeled the relevant cofactors upon analyses with AlphaFill[25] and where needed through MODELLER[83]. A summary of the initial structures used for each protein included in the database is reported in OSF (https://osf.io/y3p2x/). In selected cases, we have replaced long disordered loops with short

residue stretches using a custom pipeline based on MODELLER (https://github.com/ELELAB/MAVISp_loop_replacer). This was done to avoid potential bias in our structural calculations, due to the arbitrary conformation of such loops and their spurious contacts with the rest of the structure. In addition, for proteins with transmembrane regions, we used the PPM (Positioning of Proteins in Membrane) server 3.0 from OPM (Orientations of Proteins in Membrane)[84,85]. In cases the target proteins are larger than 2700 residues, they are not provided by AlphaFold2 database, and we model them through AlphaFold3.
The advantage of using AlphaFold-predicted structures in the default high-throughput data collection of MAVISp lies in their ability to achieve resolutions comparable to experimental data, as demonstrated in previous work[2] but circumventing limitations typically associated with experimental approaches, such as artifacts, missing atoms, and incomplete or absent residues.

Interactome

In the INTERACTOME module, we identify known interactors of the target protein by extracting data from the Mentha database[26] and matched it with available PDB structures. Mentha2PDB also examines experimentally validated dimeric complexes generated with AlphaFold2 from the HuRI and HuMAP databases by Burke et al.[27]. Mentha2PDB provides annotations of the interactors and generates input files for AlphaFold-Multimer. Mentha2PDB is freely available at https://github.com/ELELAB/PPI2PDB. A similar strategy based on STRING is under construction.

We also use other methods to identify four different classes of short linear motifs (BRCT, LIR, BH3 and UIM) in our target proteins. Depending on the type, we use a combination of simple regular expression matching, a method designed by us for structure-based identification of short linear motifs SLiMfast (available at https://github.com/ELELAB/SLiMfast) together with another method for predicting changes in secondary structure propensity that may be induced by phosphorylation in the core of putative LIR motifs, phosphor-iLIR (https://github.com/ELELAB/phospho-iLIR), or DeepLoc 2.0 for predicting the subcellular localization of the protein, especially useful for BRCT motifs[86].

Free energy calculations

We applied the BuildModel module of FoldX5 suite[87] averaging over five independent runs for the calculations of changes in free energy of folding upon mutation with MutateX and the FoldX5 method. We used the cartddg2020 protocol for folding free energy calculations with Rosetta suite and the ref2015 energy function. In this protocol, only one structure is generated at the relax step and then optimized in Cartesian space. Five rounds of Cartesian space optimization provide five pairs of wild-type and mutant structures for each variant. The change in folding free energy is then calculated on the pair characterized by the lower value of free energy for the mutant variant, as described in the original protocol[88].
We used MutateX to calculate changes in binding free energy for the LOCAL_INTERACTION module using the BuildModel and AnalyzeComplex functions of FoldX5 suite and averaging over five runs. With Rosetta, we used the flexddg protocol as implemented in RosettaDDGPrediction and the talaris2014 energy function. We used 35,000 backrub trials and a threshold for the absolute score for minimization convergence of 1 Rosetta Energy Unit (REU). The protocol then generates an ensemble of 35 structures for each mutant variant and calculates the average changes in binding free energy. We used Rosetta 2022.11 version for both stability and binding calculations. In the applications with RosettaDDGPrediction the Rosetta Energy Units (REUs) were converted to kcal/mol with available conversion factors[88]. We also applied RaSP using the same protocol provided in the original publication[38] and adjusting the code in a workflow according to MAVISp-compatible formats (https://github.com/ELELAB/RaSP_workflow). We have included data on 108 complexes at the date of 01/11/2024 (https://osf.io/y3p2x/).
For the calculations of allosteric free energy, we used the structure-based statistical mechanic model implemented in AlloSigMA2[46] . The model describes the mutated variants as 'UP' or 'DOWN' mutations depending on difference in steric hindrance upon the substitution. We did not include mutations for which the absolute changes in the volume of the side chain upon the amino acid substitution was lower than 5 $Å^3$, as recently applied to p53[20]. As a default, we retained only variants that were exposed at the solvent (>20% relative solvent accessibility of the side chain), with changes in allosteric free energy higher than 2 kcal/mol or lower than -2 kcal/mol to sites that were at distances higher than 15 Å and which belongs to pockets as identified by Fpocket[89] (**Supplementary Text S2** and workflow at https://github.com/ELELAB/MAVISp_allosigma2_workflow/)

Efoldmine

The EFOLDMINE module, integrated within the simple mode of MAVISp, predicts residues with early folding propensity using the EfoldMine tool[56]. Trained on residue-level hydrogen/deuterium exchange nuclear magnetic resonance (HDX NMR) folding data from the Start2Fold database[90], this tool uses secondary structure propensity and backbone/side-chain dynamics in a support-vector machine algorithm to predict early folding regions based on the target's primary structure.
In MAVISp, we incorporated EfoldMine to determine whether point mutations in variants fall within the predicted early folding regions, using a threshold of 0.169 to define residues involved in early folding events and considering only regions with a minimum length of three early folding residues to avoid isolated peaks, as suggested by the developers of the method[56].

Molecular dynamics simulations

We used either previously published[19,20,73,74,91–93] or newly collected within this study, all-atom molecular dynamics simulations of one microsecond with the CHARMM22star or CHARMM36m force fields[94]. All the simulations have been carried out in the canonical after equilibration steps and using explicit solvent and periodic boundary conditions. The templates files used for the simulations are provided in OSF (https://osf.io/y3p2x/). Ensembles generated using simulations are then subject to quality control, either using Mol_Analysis[95] or MetaD_Analysis (https://github.com/ELELAB/MetaD-Analysis) tools.
As a first example of how we intend to use metadynamics data for the FUNCTIONAL_DYNAMICS module we used the simulations from TP53 where the effects of mutations on an interface for protein-protein interaction (residues 207-213) was investigated. We used a collective variable based on distances between two residues (D208-R156) that were effective in capturing open (active) and closed (inactive) conformations of the loop. See repositories associated with the enhanced sampling simulations of TP53[20]. All the newly generated trajectories will be deposited as different entries in OSF, and the link is reported in the metadata on the MAVISp webserver. At the date of 01/11/2024, we have included 45 protein targets in the *ensemble mode* using as source of ensemble mostly unbiased MD simulations of 500 ns or one-μs, as detailed in the corresponding metadata on the MAVISp webserver. In some cases, we included ensembles generated by a coarse-grain model of flexibility or using the conformation provided by NMR structures from the PDB (see INPUT STRUCTURES tables in https://osf.io/y3p2x/).

Protein Structure Networks and path analysis

In the ensemble mode we apply a module building upon the simple mode LONG_RANGE module, LONG_RANGE: AlloSigma2-PSN where we constructed an atomic-contact PSN on the full trajectories using PyInteraph2[47]. Pairs of residues were retained only if their sequence distance exceeded Proxcut threshold of 1 and their edge calculations remained within less than 4.5Å, based on the thresholds described in PyInteraph2[47]. We retained edges with an occurrence greater than Pcrit threshold of 50% across the ensemble frames, weighted on the interaction strength Imin of 3.
Subsequently, we used the path_analysis function of PyInteraph2 to identify the shortest paths of communication between each pair of AlloSigma2[46] predicted mutation and respective response sites, using a minimum distance threshold of 15Å and retained paths that were four residues or longer (**Supplementary Text S2**).

CABS-flex ensembles

We used the coarse-grained CABS-flex 2.0 method and software[32] as a part of a Snakemake[96] pipeline, available at https://github.com/ELELAB/MAVISp_CABSflex_pipeline. The pipeline includes the possibility to tune the calculations by different restraints, secondary structure definition, ligand binding and more. It also contains a quality control step to evaluate the secondary structure content of the generated structures with respect to the starting one, using DSSP[97] and the SOV-refine score[98].

Mutation impact predictors

We used DeMaSk[54], GEMME[11], EVE[55], REVEL[53] and AlphaMissense[21] as predictors for the effect of mutations of any possible mutation to natural amino acids, on the full protein sequence of the main UniProt[99] isoform of each protein. We used available default parameters for each method unless noted otherwise. We used the standalone version of DeMaSk as available on its public GitHub (commit ID 10fa198), with BLAST+

2.13.0. We followed the protocol available on GitHub: we first generated the aligned homologs sequence file by using the demask.homologs module, and then calculated fitness impact predictions. Finally, we classified as damaging those mutations having a DeMaSk delta fitness score in absolute value larger or equal than 0.25, and neutral otherwise (**Supplementary Text S4**). We used the available online webserver to obtain mutation effect predictions with GEMME, upon setting the number of JET iterations to 5, to obtain more precise results. We have rank-normalized the GEMME scores per protein[100] and classified mutations having a normalized score of 0.5 or higher as damaging, and neutral otherwise. To obtain EVE scores, we have used the scripts, protocol and parameters available on the EVE GitHub (commit iD 740b0a7) as part of a custom-built Snakemake[96]-based pipeline, available at https://github.com/ELELAB/MAVISp_EVE_pipeline Using EVE first requires building a protein-specific Bayesian variational autoencoder model, which learns evolutionary constraints between residues from a multiple sequence alignment. In the current MAVISp release, we generated such alignments using EVcouplings[101], using the Uniref100[102] sequence database released on 01/03/2023, by keeping sequences with at least 50% of coverage with the target protein sequence, alignment positions with a minimum of 70% residue occupancy, and using a bit score threshold for inclusion of 0.5 bits with no further hyperparameter exploration. We then used our pipeline to perform model training, calculation of the evolutionary index, and used a global-local mixture of Gaussian Mixture Models to obtain a pathogenicity score and classification. We have used pre-computed REVEL scores for our mutations as available in dbSNFP[103,104], accessed through myvariants.info[105,106], as implemented in Cancermuts. We have classified as damaging those mutation have a REVEL score larger or equal to 0.5[107]. We included AlphaMissense pathogenicity prediction scores and classification as available by the dataset of prediction for all possible amino acid substitutions in UniProt canonical isoforms, release version 2[108].

Annotations from experimental data

We developed Python scripts to identify the overlap in coverage between MAVEdb[71] and MAVISp, and to systematically and automatically retrieve the score sets associated with the shared entries from the MAVE database[71] database through their API (https://api.mavedb.org/docs). Where available, we also extracted information on methods and classification thresholds. For entries where this information was incomplete, the corresponding publications were manually reviewed to extract thresholds for variant classification.
The ProteinGym[70] repository was locally downloaded from github, and a custom Python script was used to process the datasets based on the reference files provided in the repository. The datasets used for the analysis contained the experimental scores and the classification provided by the authors either based on the median of the score distributions or via manual annotation. The scores and their classifications were then integrated into the final CSV file generated by MAVISp. The aggregated scores, along with their classifications, were compiled into the final CSV file produced by MAVISp through a module dedicated to the experimental data.

Workflows for data collection

We provide and maintain two SnakeMake workflows for the data collection of the default modules of MAVISp. The first is a Snakemake pipeline to automate MutateX runs as much as possible. It is designed to automatically download and trim chosen structure(s) from the AlphaFold structural database, or a custom structure input file, when necessary, trim them as requested, and generate desired MutateX folding free energy scans with a predictable directory structure. It only requires as input a csv file with metadata on the desired scan and a configuration file with details on the run to be performed. It is available at https://github.com/ELELAB/mutatex_pipelines/tree/main/custom_collect_scan.
Once such a scan is available, it is possible to use a second Snakemake pipeline, called MAVISp_automatization, which performs most of the steps that are necessary to annotate a protein for a MAVISp simple mode entry. Similarly to the previous pipeline, it only requires metadata on the target protein to be analyzed, as well as a MutateX mutational scan. It generates a dataset that can then be imported into the MAVISp database, except for predictions performed using Rosetta-based methods, since these are much more computationally expensive and need to be performed separately using the RosettaDDGPrediction pipeline[37]. Using a Snakemake pipeline allows to improve efficiency and scalability, allowing to take advantage of multi-core system to process several proteins or perform different analyses in parallel. It is available at https://github.com/ELELAB/MAVISp_automatization.

Data Availability

The data can either be consulted through our web server (https://services.healthtech.dtu.dk/services/MAVISp-1.0/) or as individual CSV files in the OSF repository https://osf.io/ufpzm/. Other raw data and utilities can be found at the MAVISp extended data OSF repository (https://osf.io/y3p2x/) Reports for each protein are available at https://elelab.gitbook.io/mavisp/.

References

1. Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Lett* **340**, 284–295 (2013).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics 2018 137:9* **137**, 665–678 (2018).
4. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol* **24**, 147 (2023).
5. Burke, W., Parens, E., Chung, W. K., Berger, S. M. & Appelbaum, P. S. The Challenge of Genetic Variants of Uncertain Clinical Significance. *Ann Intern Med* **175**, 994–1000 (2022).
6. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).
7. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
8. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* (2012) doi:10.1158/2159-8290.CD-12-0095.
9. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835–D844 (2020).
10. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15**, 816–822 (2018).
11. Laine, E., Karami, Y. & Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol Biol Evol* **36**, 2604–2619 (2019).
12. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep* **38**, 110207 (2022).
13. Cagiada, M. *et al.* Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol Biol Evol* **38**, 3235–3246 (2021).
14. Jepsen, M. M., Fowler, D. M., Hartmann-Petersen, R., Stein, A. & Lindorff-Larsen, K. Classifying disease-associated variants using measures of protein activity and stability. *Protein Homeostasis Diseases* 91–107 (2020) doi:10.1016/B978-0-12-819132-3.00005-1.
15. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv* (2022) doi:10.1101/2021.10.04.463034.
16. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439–D444 (2022).
17. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
18. Kumar, M. & Papaleo, E. A pan-cancer assessment of alterations of the kinase domain of ULK1, an upstream regulator of autophagy. *Sci Rep* **10**, 14874 (2020).
19. Fas, B. A. *et al.* The conformational and mutational landscape of the ubiquitin-like marker for autophagosome formation in cancer. *Autophagy* 1–24 (2020).
20. Degn, K. *et al.* Cancer-related Mutations with Local or Long-range Effects on an Allosteric Loop of p53. *J Mol Biol* **434**, 167663 (2022).
21. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
22. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
23. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
24. Degn, K., Beltrame, L., Tiberti, M. & Papaleo, E. PDBminer to Find and Annotate Protein Structures for Computational Analysis. *J Chem Inf Model* **63**, 7274–7281 (2023).
25. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**, 205–213 (2022).
26. Calderone, A., Castagnoli, L. & Cesareni, G. Mentha: A resource for browsing integrated protein-interaction networks. *Nat Methods* **10**, 690–691 (2013).

27. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat Struct Mol Biol* **30**, (2023).
28. Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* **39**, (2023).
29. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun* **13**, (2022).
30. Di Rita, A. *et al.* HUWE1 E3 ligase promotes PINK1/PARKIN-independent mitophagy by regulating AMBRA1 activation via IKKα. *Nat Commun* **9**, 3755 (2018).
31. Holdgaard, S. G. *et al.* Selective autophagy maintains centrosome integrity and accurate mitosis by turnover of centriolar satellites. *Nat Commun* **10**, 1–19 (2019).
32. Kuriata, A. *et al.* CABS-flex 2.0: A web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res* **46**, W338–W343 (2018).
33. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2**, 19–25 (2015).
34. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **185**, 604–613 (2014).
35. Bonomi, M. *et al.* Promoting transparency and reproducibility in enhanced molecular simulations. *Nat Methods* **16**, 670–673 (2019).
36. Tiberti, M. *et al.* MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* **23**, bbac074 (2022).
37. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. *Protein Science* **32**, e4527 (2023).
38. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *Elife* **12**, e82593 (2023).
39. Nielsen, S. v. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* **13**, e1006739 (2017).
40. Abildgaard, A. B. *et al.* Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. (2019) doi:10.7554/eLife.49138.001.
41. Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J. & Moal, I. H. SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **35**, 462–469 (2019).
42. Sampson, J. M. *et al.* Robust Prediction of Relative Binding Energies for Protein–Protein Complex Mutations Using Free Energy Perturbation Calculations. *J Mol Biol* **436**, (2024).
43. Sargsyan, K. & Lim, C. Using protein language models for protein interaction hot spot prediction with limited data. *BMC Bioinformatics* **25**, (2024).
44. Sapozhnikov, Y., Patel, J. S., Ytreberg, F. M. & Miller, C. R. Statistical modeling to quantify the uncertainty of FoldX-predicted protein folding and binding stability. *BMC Bioinformatics* **24**, 426 (2023).
45. Peccati, F., Alunno-Rufini, S. & Jiménez-Osés, G. Accurate Prediction of Enzyme Thermostabilization with Rosetta Using AlphaFold Ensembles. *J Chem Inf Model* **63**, 898–909 (2023).
46. Tan, Z. W., Guarnera, E., Tee, W. V. & Berezovsky, I. N. AlloSigMA 2: Paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. *Nucleic Acids Res* **48**, W116–W124 (2020).
47. Sora, V. *et al.* PyInteraph2 and PyInKnife2 to Analyze Networks in Protein Structural Ensembles. *J Chem Inf Model* **63**, 4237–4245 (2023).
48. Tiberti, M. *et al.* The Cancermuts software package for the prioritization of missense cancer variants: a case study of AMBRA1 in melanoma. *Cell Death Dis* **13**, 872 (2022).
49. Orioli, S., Henning Hansen, C. G. & Lindorff-Larsen, K. Transient exposure of a buried phosphorylation site in an autoinhibited protein. *Biophys J* **121**, 91–101 (2022).
50. Henriques, J. & Lindorff-Larsen, K. Protein Dynamics Enables Phosphorylation of Buried Residues in Cdk2/Cyclin-A-Bound p27. *Biophys J* **119**, 2010–2018 (2020).
51. Potel, C. M. *et al.* Impact of phosphorylation on thermal stability of proteins. *Nat Methods* **18**, 757–759 (2021).
52. Huang, H. *et al.* iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res* **46**, D542–D550 (2018).
53. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877–885 (2016).
54. Munro, D. & Singh, M. DeMaSk: A deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36**, 5322–5329 (2020).
55. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

56. Raimondi, D., Orlando, G., Pancsa, R., Khan, T. & Vranken, W. F. Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. *Sci Rep* **7**, 8826 (2017).
57. Byrjalsen, A. *et al.* Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes. *PLoS Genet* **16**, (2020).
58. Scrima, S., Lambrughi, M., Tiberti, M., Fadda, E. & Papaleo, E. ASM variants in the spotlight: A structure-based atlas for unraveling pathogenic mechanisms in lysosomal acid sphingomyelinase. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1870**, 167260 (2024).
59. Antonescu, O. N. *et al.* Decoding Phospho-Regulation and Flanking Regions in Autophagy-Associated Short Linear Motifs: A Case Study of Optineurin-LC3B Interaction. *BioRxiv* 1–43 (2023) doi:10.1101/2023.09.30.560296.
60. Sahu, S. *et al.* AVENGERS: Analysis of Variant Effects using Next Generation sequencing to Enhance BRCA2 Stratification. *BioRxiv* 1–31 (2023) doi:10.1101/2023.12.14.571713.
61. Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**, 1 (2019).
62. Miao, Y. *et al.* Cycloheximide (CHX) Chase Assay to Examine Protein Half-life. *Bio Protoc* **13**, (2023).
63. McKinnon, K. M. Flow Cytometry: An Overview. *Curr Protoc Immunol* **120**, 5.1.1-5.1.11 (2018).
64. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874–882 (2018).
65. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-Resolution Mapping of Protein Concentration Reveals Principles of Proteome Architecture and Adaptation. *Cell Rep* **7**, 1333–1340 (2014).
66. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
67. Yen, H.-C. S., Xu, Q., Chou, D. M., Zhao, Z. & Elledge, S. J. Global Protein Stability Profiling in Mammalian Cells. *Science (1979)* **322**, 918–923 (2008).
68. Post, K. L. *et al.* Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nat Commun* **11**, (2020).
69. Mighell, T. L., Thacker, S., Fombonne, E., Eng, C. & O'Roak, B. J. An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. *Am J Hum Genet* **106**, 818–829 (2020).
70. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Adv Neural Inf Process Syst* **36**, 64331–64379 (2023).
71. Esposito, D. *et al.* MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**, (2019).
72. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
73. Invernizzi, G., Tiberti, M., Lambrughi, M., Lindorff-Larsen, K. & Papaleo, E. Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. *PLoS Comput Biol* **10**, e1003744 (2014).
74. Nygaard, M. *et al.* The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci* **3**, (2016).
75. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46**, D380–D386 (2018).
76. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
77. Kønig, S. M., Rissler, V., Terkelsen, T., Lambrughi, M. & Papaleo, E. Alterations of the interactome of Bcl-2 proteins in breast cancer at the transcriptional, mutational and structural level. *PLoS Comput Biol* **15**, e1007485 (2019).
78. Suzuki, H., Osawa, T., Fujioka, Y. & Noda, N. N. Structural biology of the core autophagy machinery. *Curr Opin Struct Biol* **43**, 10–17 (2017).
79. Geng, C., Xue, L. C., Roel-Touris, J. & Bonvin, A. M. J. J. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Computational Molecular Science* **9**, (2019).
80. Li, Y. *et al.* Genomics of Alzheimer's disease implicates the innate and adaptive immune systems. *Cellular and Molecular Life Sciences* **78**, 7397–7426 (2021).
81. Papaleo, E. *et al.* TRAP1 S-nitrosylation as a model of population-shift mechanism to study the effects of nitric oxide on redox-sensitive oncoproteins. *Cell Death Dis* **14**, (2023).
82. Katarina Sooe Tiemann, J., Zschach, H., Lindorff-Larsen, K. & Stein, A. Interpreting the molecular mechanisms of disease variants in 2 human membrane proteins. doi:10.1101/2022.07.12.499731.
83. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).

84. Lomize, A. L., Todd, S. C. & Pogozheva, I. D. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. *Protein Science* **31**, 209–220 (2022).
85. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* **40**, D370–D376 (2012).
86. Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res* **50**, W228–W234 (2022).
87. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. *Bioinformatics* 1–2 (2019) doi:10.1093/bioinformatics/btz184.
88. Frenz, B. *et al.* Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front Bioeng Biotechnol* **8**, (2020).
89. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11 (2009).
90. Pancsa, R., Varadi, M., Tompa, P. & Vranken, W. F. Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res* **44**, D429–D434 (2016).
91. Salamanca Viloria, J., Allega, M. F., Lambrughi, M. & Papaleo, E. An optimal distance cutoff for contact-based protein structure networks using side chain center of masses. *Sci Rep* **7**, 2838 (2016).
92. Papaleo, E., Sutto, L., Gervasio, F. L. & Lindorff-Larsen, K. Conformational changes and free energies in a proline isomerase. *J Chem Theory Comput* **10**, 4169–4174 (2014).
93. Srinivasan, S. *et al.* The conformational plasticity of structurally unrelated lipid transport proteins correlates with their mode of action. *PLoS Biol* **22**, e3002737 (2024).
94. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100**, (2011).
95. Lambrughi, M. *et al. Analyzing Biomolecular Ensembles*. *Methods in Molecular Biology* vol. 2022 (2019).
96. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33 (2021).
97. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* (1983) doi:10.1002/bip.360221211.
98. Liu, T. & Wang, Z. SOV-refine: A further refined definition of segment overlap score and its significance for protein structure similarity. *Source Code Biol Med* **13**, 1–10 (2018).
99. Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
100. Okada, T., Miyagi, H., Sako, Y., Hiroshima, M. & Mochizuki, A. Origin of diverse phosphorylation patterns in the ERBB system. *Biophys J* **121**, 470–480 (2022).
101. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).
102. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
103. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894–899 (2011).
104. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* **12**, 1–8 (2020).
105. Xin, J. *et al.* High-performance web services for querying gene and variant annotation. *Genome Biol* (2016) doi:10.1186/s13059-016-0953-9.
106. Lelong, S. *et al.* BioThings SDK: a toolkit for building high-performance data APIs in biomedical research. *Bioinformatics* **38**, 2077 (2022).
107. Hopkins, J. J., Wakeling, M. N., Johnson, M. B., Flanagan, S. E. & Laver, T. W. REVEL Is Better at Predicting Pathogenicity of Loss-of-Function than Gain-of-Function Variants. *Hum Mutat* **2023**, 1–6 (2023).
108. Cheng, J. *et al.* Predictions for AlphaMissense. (2023) doi:10.5281/ZENODO.8360242.

Author Contributions (CRediT Classification)

Conceptualization: EP, *Data Curation*: MA, MU, KD, LBa, SS, KK, PSB, KM, LBb, TD, AM, LF, EK, AO, AHE, JB, JS, FM, BAF, GT, PB, JV, ML, MT, EP *Formal Analysis*: MA, LBa, MU, TD, ML, EP. *Funding Acquistion:* EP, MA, CC, SI, MN, AR. *Methodology*: MA, MU, LBa, KD, SS, KK, KM, LBb, EK, JB, AP, JV, ML, MT, EP. *Project administration:* EP. *Resources*: PWD, EP, SI, CC. *Software*: MA, KD, KK, SS, LBb, AP, ML, MT, PWS. *Supervision*: EP, ML, MT, SS, MU, MA, LBa. *Validation*: All the coauthors. *Visualization*: MA, MU, KD, LBa, MU, MT. *Writing – Original Draft:* EP, MT. *Writing – Review and Editing*: All the coauthors.

Acknowledgements

Our research has been supported by Carlsberg Foundation Distinguished Fellowship (CF18-0314), Danmarks Grundforskningsfond (DNRF125), Hartmanns Fond (R241-A33877), LEO Foundation (LF17006), NovoNordisk Fonden Bioscience and Basic Biomedicine (NNF20OC0065262) to E.P. group. Part of the calculations have been supported by a EuroHPC Benchmark Access Grant (EHPC-BEN-2023B02-010) and a EuroHPC Regular Grant (EHPC-REG-2023R01-051) on Discoverer. The work is also supported by A PhD Fellowship from the Danish Data Science Academy (DDSA) to M.A. SI and JS would like to thank the Merkin Institute of Transformative Technologies in Healthcare. AR is supported by The Assar Gabrielsson's foundation and The Healthcare Board, Region Västra Götaland. PSB is supported by a KBVU Pre-Graduate Fellowship (R361-A21156).

This project is also in part based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology) which supported a short-term visit for GT in EP group.

KD, MT and EP would like to thank Kresten Lindorff-Larsen, Lasse M. Blaabjerg and Matteo Cagiada for their inputs and feedback to port the RaSP code within our workflow. We also would like to thank the group of Stefano Vanni for providing published MD trajectories for analysis within the ensemble mode of MAVISp. We would like to thank Jeppe Samuelsen, Kledi Salla, Amalie Drud Nielsen, Julie Bruun Brockhoff, Laura Mattioli, Zeming Hou, Subayan Akhuli and Ona Saulianskaite for preliminary work on targets reported in the database.