MAVISp: A Modular Structure-Based Framework for Genomic Variant Interpretation

Matteo Arnaudi[1,2†], Ludovica Beltrame[1†], Kristine Degn[1†], Mattia Utichi[1,2†], Simone Scrima[2], Pablo Sánchez-Izquierdo Besora[1,2], Karolina Krzesińska[1,2], Alberte Heering Estad[1,2], Francesca Maselli[2,3], Terézia Dorčaková[2], Jordan Safer[4], Katrine Meldgård[1,2], Philipp Becker[1], Valentina Sora[1,2], Alberto Pettenella[2], Julie Bruun Brockhoff[1,2], Amalie Drud Nielsen[1,2], Jérémy Vinhas[2], Peter Wad Sackett[1], Claudia Cava[5], Anna Rohlin[6,7], Mef Nilbert[8,9], 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]Department of Science, Technology and Society, Scuola Universitaria IUSS, Istituto Universitario di Studio Superiori, Piazza della Vittoria 15, 27100 Pavia, Italy
[6] Department of Clinical Genetics and Genomics, Sahlgrenska university hospital
[7] Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[8] Institute of Clinical Medicine, Dept. Oncology and Pathology, Lund University, Sweden
[9] 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, including cancer, continues to expand thanks to the advent of advanced sequencing techniques integrated into clinical practice. The rapid growth in the identification of genomic variants has led to the classification of many variants as Variants of Uncertain Significance (VUS) or with conflicting evidence, posing challenges in their interpretation and application. Here we introduce MAVISp (Multi-layered Assessment of Varlants by Structure for proteins), a modular structural framework for variant interpretation. We also provide a web server (https://services.healthtech.dtu.dk/services/MAVISp-1.0/), to enhance data accessibility, consultation, and re-usability. Currently, MAVISp offers analyses for more than 200 different proteins, encompassing approximately 85000 variants. A dedicated team of biocurators and reviewers continuously analyze and update protein targets using standardized workflows, incorporating high-throughput free energy calculations or biomolecular simulations. Here, we illustrate the potential of the MAVISp approach through a selection of case studies. Our framework aids in the interpretation of genomic variants, particularly those categorized as VUS, and holds great potential for advancing the understanding and application of genomics in disease research.

Keywords: variants of uncertain significance, cancer genomics, protein structures, free energy calculations, protein stability, protein function, long-range structural communication

Introduction

Cancer is caused by the accumulation of mutations in the genome, especially in genes with tumor suppressor or oncogenic functions[1]. Cancer alterations are heterogenous even when considering tumors from the same tissue. This heterogeneity affects predisposition to cancer, treatments response, and risk of relapse[2]. In parallel, it is also challenging to understand the molecular mechanisms and effects related to the alterations that can be found by sequencing approaches. We are witnessing unprecedented advances in cancer genomics, sequencing[3], structural biology[4], and high-throughput multiplex-based assays[5,6]. Consequently, Variants of Uncertain Significance (VUS) or conflicting classifications are continuously identified. Many variants have also been reported in publicly available repositories, including polymorphisms observed in healthy individuals[7] or compendia of cancer-related variants[8–10]. In other studies, the focus has been on specific genes[11] or hereditary cancer types[12] or, for variant classification from a clinical point of view[13]. Similar challenges apply to other diseases[14].

The bioinformatics community has been providing different approaches to predict variant impact in agreement with experiments and cellular readouts[15–19] thanks to advances in machine learning and coevolution-based methods. Furthermore, the advent of AlphaFold[4,20,21] provides the possibility to obtain accurate three-dimensional (3D) structures of proteins and their complexes with quality often close to experimental techniques. In addition, the possibility to perform long and accurate biomolecular simulations with GPU-based infrastructures and robust physical models allows to investigate conformational changes and dynamics in protein structures over different timescales[22].

In three previous pilot projects[23–25], we provided the foundations for a structure-based framework to analyze and classify the impact of germline and somatic variants related to cancer. This applies to variants that alter the coding region of a gene and have consequences on the protein product. For example, the framework can be used to prioritize variants for experimental and functional studies[24]. The framework could also be applied, more broadly, to study disease-related variants. We foresee that such a framework, if properly formalized and standardized, could complement available pathogenicity scores. The overall goal is to link the effects of the variants to the specific underlying molecular mechanisms[25]. In this context, pathogenicity scores only estimate the likelihood of a variant being pathogenic. They lack a direct connection to the effects of the variants in relation to the function that the altered protein carries out at the cellular level. A mechanistic understanding of the effect of variants is fundamental knowledge to design strategies for disease prevention, genetic counseling, clinical care, and drug treatments. From a fundamental research viewpoint, the mechanistic knowledge is also essential in designing and prioritizing experiments to understand the underlying molecular details.

In light of the above observations, we developed MAVISp (Multi-layered Assessment of VarIants by Structure for proteins) to handle high-throughput variant analysis within standardized workflows and integrating pathogenicity scores. The data are also accessible through a Streamlit-based website for consultation and download for further studies (https://services.healthtech.dtu.dk/services/MAVISp-1.0/).

We release the results for over 200 proteins for illustrative purposes of how to operate within MAVISp. New data and updates of existing entries will be continuously released in connection with the related publications where the assessment will be applied.

Results

Overview of MAVISp and its database

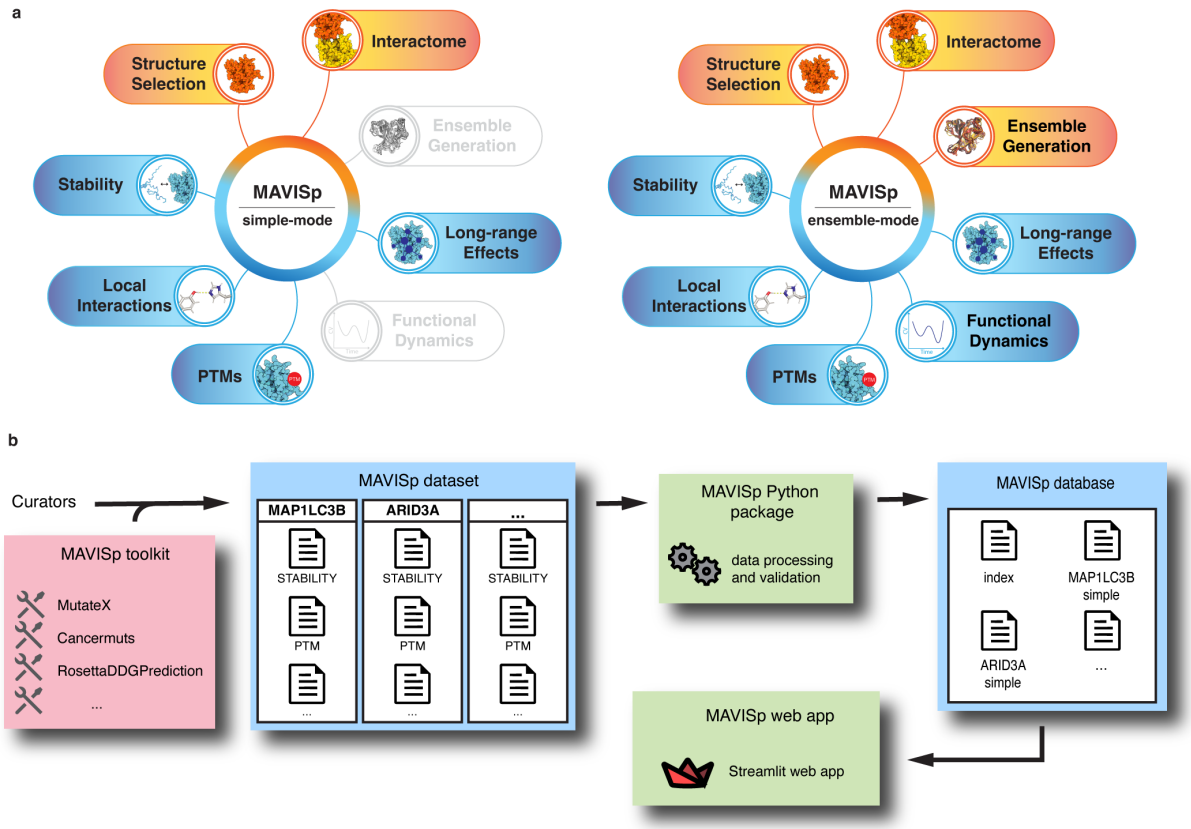

Fig. 1. Overview of MAVISp components. (a) MAVISp includes different modules, each handled by workflow engines or dedicated tools. The modules in orange are used for selection and collection of the protein structures. The modules in blue are devoted to structural analyses of the effects of the variants. The procedure is applied to start from the gene name and a list of variants of interest. All the steps, for a gene, can be carried out on a standard server with 32-64 CPUs, apart from the ENSEMBLE GENERATION module, which includes all-atom MD simulations. Depending on the simulation length and system size, it might require access to HPC facilities. The left hand side illustrates the *simple mode* for the assessment, which uses single experimental structures or models from AlphaFold2. The right hand side illustrates the *ensemble mode,* in which a conformational ensemble for the target protein or its complexes is applied. (b) Scheme of the current workflow for the MAVISp database and app. Curators use specific workflows and protocols within each MAVISp module to derive structure-based predictions of the changes connected to the variants identified for a protein target. The results are gathered into a text-file based database. The generated data are then processed by the MAVISp Python package, which performs consistency checks, further processes data and outputs human-readable CSV table files, i.e., the MAVISp database. The CSV files are imported by the Streamlit web app, which powers the MAVISp webserver (https://services.healthtech.dtu.dk/services/MAVISp-1.0/) and makes the data available both for interactive visualization and download. In addition, we provide reports for each protein to facilate the interpretation of the results through Gitbook: https://elelab.gitbook.io/mavisp/.

MAVISp works by performing a set of independent predictions, each one predicting the effect of a specific set of amino acid type change on a different aspect of protein function or structural stability, starting from one or more protein structures. These independent predictions are performed by the so-called MAVISp modules (**Fig. 1a**). MAVISp can be applied on individual 3D structures of proteins and their complexes (*simple mode*) or on an ensemble of structures (*ensemble mode*). The framework is completely modular, so that, depending on the case study, all the modules or only a selection of them can be applied. Each module relies on Snakemake, Dask workflows, or Python scripts, often supported by specific virtual environments. The modules can be divided into two main categories: i) modules to retrieve and select the structures for the analyses (in orange in **Fig. 1a**) and modules for the individual layers of the assessment (in blue in **Fig. 1a**).
The overall workflow of MAVISp (**Fig. 1b**) is designed so that several biocurators can work concomitantly and independently on different proteins. The curators generate and validate data for the different MAVISp modules, which are stored as text files. A purposefully designed Python package parses such data, further processes them, performs sanity checks and writes the results as human-readable csv files, which constitute the MAVISp database. Finally, the MAVISp web app reads the database and displays it as a user-friendly

Streamlit-based website (https://services.healthtech.dtu.dk/services/MAVISp-1.0/) where all the results of our analyses are made available.
The MAVISp Python package and web app source code is also available on GitHub (https://github.com/ELE-LAB/MAVISp**)**, while the full 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 will add newly available data in batches on a regular basis. We provide the information on the list of entries to be processed, modified, or released in a CHANGELOG file in our GitHub repository. If data on specific variants or proteins have been published as a part of a scientific publication applying the MAVISp framework, the PMID or DOI of the associated publication will be annotated in the MAVISp database and reported in the Gitbook page.
We welcome requests on targets or variants that are neither available in MAVISp or in the pipeline for curation. We also welcome contributors as curators, after training and adherence to our guidelines (see Gitbook page for information, https://elelab.gitbook.io/mavisp/). A full update and revision of entries will be carried out once per year and will guarantee that we maintain up-to-date resources for free energy calculations, identification of biological interactions, post-translational modifications. Moreover, we aim to keep expanding the toolkit and develop new modules, so that even more comprehensive assessments can be achieved.

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**). Our framework has been designed with flexibility in mind so that new modules can be added to cover essential missing properties for variant assessment, and existing ones can be easily modified.
The STRUCTURE_SELECTION module includes structure discovery in the Protein Data Bank (PDB)[26], the AlphaFold Protein Structure Database[21], or to run AlphaFold and AlphaFold-multimer[20] locally. In addition, it streamlines the selection of the structures in terms of experimental resolution, missing residues, amino acidic substitutions with respect to the UniProt reference sequence, and the AlphaFold per-residue confidence score (pLDDT), thanks to workflows such as PDBminer [27]. Moreover, this module also takes advantage of the programmatic access to AlphaFill [28] to identify cofactors to be included in the model structure or to identify mutation sites that should be flagged if in the proximity of a cofactor that has not been included in the starting model. In some cases, we reconstruct missing residues or design linker regions to replace large, disordered linkers between protein domains using Modeller[29]. As an example of the latter case, we reported in the database AMBRA1 and ALPK1.
The INTERACTOME module facilitates the identification of protein interactors for the target protein and their complex structures by querying the Mentha database [30], the PDB, and experimentally validated proteome-wide AlphaFold models [31]. It also generates the input files for AlphaFold-Multimer and retains only the models with reasonable accuracy in terms of pLDDT and Predicted Aligned Error (PAE) scores. In addition, for complexes in which the interaction happens between a protein and a peptide harboring a Short Linear Motif (SLiMs), we support structure collection using AlphaFold-based protocols [20,32] or comparative modeling [33,34].
The ENSEMBLE_GENERATION module allows collecting coarse-grained models for protein flexibility (with CABS-flex [35]) or all-atom Molecular Dynamics (MD) simulations (with GROMACS [36] and PLUMED[37,38]) of the structure of the protein (or its complexes), depending on the accuracy in the ensemble that the user wants to achieve and available computational resources. Once structures or structural ensembles for the protein candidate are selected – either alone or with interactors - the analysis modules can be used.
The minimal requirement to include a protein and its variants in the MAVISp database is the application of the STABILITY, PTM modules and the pathogenicity scores. In the STABILITY step, the effect of the variants on the structural stability of the proteins is investigated using folding free energy calculations with the FoldX5 suite[39], the cartddg2020 protocol, and the ref2015 energy function of Rosetta[40,41]. This module takes advantage of recently developed workflows for high-throughput in-silico mutagenesis scans, i.e., MutateX [42] and RosettaDDGPrediction[43] , or RaSP[44]. 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 [45,46]. 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. As a default, we rely on the classification based on the consensus between Rosetta and FoldX. We included RaSP since we plan to compare the influence of using a consensus among

FoldX and RaSP on larger datasets and evaluate if we can fully replace the CPU-intensive scans based on Rosetta with RaSP predictions.
The LOCAL_INTERACTION module can be applied if at least one structure of a complex between the target protein and another biomolecule is identified by the STRUCTURE_SELECTION and INTERACTOME modules. This structure can be a complex deposited in the PDB or predicted with reasonable accuracy with AlphaFold-multimer. 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 the FoldX5 and the flexddg protocol and talaris2014 energy function of Rosetta[47]. In this step, we apply MutateX and RosettaDDGPrediction and apply a cut-off-based consensus strategy for classification similar as what done in the STABILITY module. We classify a mutation as stabilizing (both FoldX and Rosetta predict $\Delta\Delta G <= -1$ kcal/mol), neutral (-1 kcal/mol $< \Delta\Delta G < 1$ kcal/mol) or destabilizing ($\Delta\Delta G >= 1$ kcal/mol). Cases in which the two methods are in disagreement 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 in relation to protein and DNA interactions and homodimers. Of note, a strength of our approach is that we provide annotations for the effects of protein variants on different biological interfaces for the same target protein.
In the *ensemble mode*, the calculations for the STABILITY and LOCAL_INTERACTION modules can be applied to a set of 50 or 25 frames from the simulations (FoldX-based calculations) [42] or on the main representative structures of the ensemble after structural clustering carried out with GROMACS (Rosetta-or RasP-based calculations). In addition, other metrics based on Protein Structure Network as the changes in interaction strength upon mutations and changes in hub behavior are currently under evaluation for inclusion in the framework, using PyInteraph2 [48] to expand the portfolio of tools in the *ensemble mode*.
The LONG_RANGE module applies coarse-grained models to estimate allosteric free energy changes upon mutation based on Allosigma2[49]. For each mutation that causes a significant change in residue size (at least 5 $Å^3$), we use the AlloSigma2 allosteric signaling map to single out other residues in the protein that show a significant ($|\Delta g| > 2$ kcal/mol) positive or negative allosteric free energy change (see Materials and Methods). We classify each mutation as destabilizing if it exclusively causes positive changes of allosteric free energy. stabilizing if it exclusively causes negative changes of allosteric free energy. Mixed effects if both conditions occur, or Neutral if the mutation doesn't cause any significant change. Finally, we classify as those mutations that do not cause a significant change in residue side-chain volume as uncertain. In the *ensemble mode*, other graph theory metrics based on changes in the shortest communication paths or centrality metrics could be applied using PyInteraph2 [48] and are currently under investigation.
The FUNCTIONAL_DYNAMICS module is only available in the *ensemble mode*. For selected variants, enhanced sampling simulations are designed to allow further verification of the long-range effects of a mutation to the corresponding functional site, similar to what we recently applied to p53 [25]. In the database, we exemplified this application using the published mutations on p53.
The PTM module currently supports phosphorylations, and it classifies the effect of variants at phosphorylatable sites. The PTM module considers the possible effect of the loss or change of phosphorylation sites for consequences on regulation, stability, or function changes. The module uses information such as solvent accessibility of the mutation site, being part of a phosphorylatable linear motif or predicted changes of free energy of stability or binding upon mutation or upon modification to classify the mutation for each category. We use custom decision logic (see **Fig. S1-S3** for more details) to classify each mutation as neutral, damaging, unknown effect, potentially damaging or uncertain. The PTM module takes advantage of the annotations provided by the Cancermuts package[50] for searching SLiM motifs and identifying the phospho-sites. In addition, we rely on a manually curated selection of phospho-modulated SLiMs (see https://github.com/ELE-LAB/MAVISp/blob/main/mavisp/data/phosphoSLiMs_09062023.csv). Finally, it should be noted that in the case of solvent-inaccessible phosphorylatable residues, the effects are classified as uncertain in the *simple mode* for all the PTM-related classifications. In these cases, the *ensemble mode* is required to investigate the possibility that a cryptic phosphorylated site could become accessible through conformational changes[51,52].
Additionally, MAVISp provides annotations for the variant interpretation reported in ClinVar[13], or calculated with REVEL[53], DeMask[54] , GEMME[17], EVE (Evolutionary model of variant effect)[55] , and AlphaMissense[56]. The pathogenicity scores are used together with the results from the MAVISp layers for variant interpretation and prioritization. MAVISp also included the pLDDT score if an AlphaFold model is used, the secondary structure, and the relative solvent accessibility of the wild-type side chain, which can assist in the interpretation of the results. For each entry, we provide a report with the main results through Gitbook

(https://elelab.gitbook.io/mavisp/) that can be exported as a PDF file. Furthermore, we provide a review status for each entry to provide guidance on the quality and integrity of the curated data. To this goal, we have defined four levels (i.e., stars) associated with the review status of each protein.

Sources of variants supported by MAVISp

By default, all variants reported for the target protein in COSMIC, cBioPortal, and ClinVar are annotated within MAVISp. We will update and maintain the entries in the MAVISp database to include new variants or correct previously misreported ones. The variants are retrieved from ClinVar and using Cancermuts[50].
In addition, we support variant lists from other studies, such as data on cohort-based or nationwide studies or other disease-related genomic initiatives. For example, we recently analyzed variants found with low frequency in the healthy population on germline samples of childhood cancer patients[57] using one of the tools applied in the STABILITY module[43]. We now included in the database a full assessment for the proteins investigated in the previous study, also including variants from COSMIC, cBioPortal or ClinVar.
Currently, MAVISp includes data on 84427 variants and 205 proteins (at 16/11/2023). Some examples of the proteins and the currently available data and analyses from MAVISp are described in detail in the next sections and in **Supplementary Text S1**.

Transcription factors

Alterations in transcription factors are often linked to aberrant gene expression, including processes such as proliferation, cell death, and other cancer hallmarks[58]. Different mechanisms are at the base of altering transcription factor activity 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 some of the methods now integrated within MAVISp and MD simulations to study DNA-binding domains (ARID3A and p53[25,59]) or domains for protein-protein interaction (MZF1 and its SCAN domain[60]) from a mechanistic point of view. Here we also included other transcription factors such as ESRRB and GATA3. As a more detailed example, we applied the *simple* and the *ensemble mode* of MAVISp to ARID3A for a comprehensive understanding of mutations found in cancer samples. Overall, we characterized 45 mutations (**Fig. 2a**) in the ARID3A DNA-binding domain selected for the study. The usage of MAVISp ensemble mode allowed to shed further light on six variants classified as uncertain in the *simple mode* for effects on structural stability (i.e., neutral effects predicted for G237W, M274T, I289M, S333N, and R350W, along with a destabilizing alteration in the case of P263H)**.** All the variants that MAVISp identified with a damaging effect on stability, long-range interactions, or post-translational regulation are also characterized by REVEL scores [53] higher than 0.5, suggesting that the assessment here provided can help to link the pathogenic effect with specific mechanisms. We also identified mutations that might have allosteric (i.e., distal) effects both stabilizing other regions of the protein or destabilizing them (**Fig. 2a**).

Pro-survival and pro-apoptotic proteins belonging of the BCL2 family proteins:

One of the cancer hallmarks is the ability to escape cell death (i.e., apoptosis)[58]. Among the different cell-death pathways, 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[61]. In the previous work, we were missing a comprehensive characterization of their missense mutations at the structural level. 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 the analyses, the C-terminal transmembrane helix has been removed since the current version of our approach does not support transmembrane proteins or domains. We focused our attention on pathogenic variants, as predicted by REVEL and EVE, to evaluate if the structure-based assessment can help to shed light on their effects. The results are reported in **Table S1** for BCL2, as an example. Similar analyses for the other pro-apoptotic or pro-survival proteins are available by accessing the corresponding entries of the MAVISp database.

Autophagy proteins

Autophagy is a clearance mechanism with a dual role in cancer. The autophagy pathway relies on approximately 40 proteins, constituting the core autophagy machinery[62]. 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 our previous works[23,24]. Moreover, we collected data on other autophagic proteins (see webserver or OSF repository).
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 204 variants reported in COSMIC, cBioPortal and/or ClinVar. Among these, 15 had predicted long-range destabilizing effects, eight had a damaging effect on stability, one had a damaging PTM effect on regulation (S954N), and four variants (L53P, G183V, E191G, and L215P) are characterized by both effects on stability and long-range communication (**Fig. 2B-C**). Most of the variants resulting in long-range and structural destabilizing effects are located 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. In addition, the K162N variant (REVEL score 0.741 and predicted as pathogenic by AlphaMissense) has a detrimental distal effect, and according to Cancermuts annotation might impair a known acetylation for ULK1 activation and, consequently, autophagy regulation[63]. The predicted destabilizing (**Fig. 3D**) variants V211F and L215P have been identified in samples from The Cancer Genome Atlas (TCGA) in a previous study[23].
We then analyzed MAP1LC3B regarding the effect of the mutations on the interaction with the LC3-interacting region (LIR) of its binding partner p62, as well as with other interactors. This provides an example of the application of structural models of SLiM-target interactions to understand changes in binding affinity upon mutations. We identified 44 mutations in MAP1LC3B reported in COSMIC and/or cBioPortal. The MAVISp stability module produced results in line with our previous study. [24].

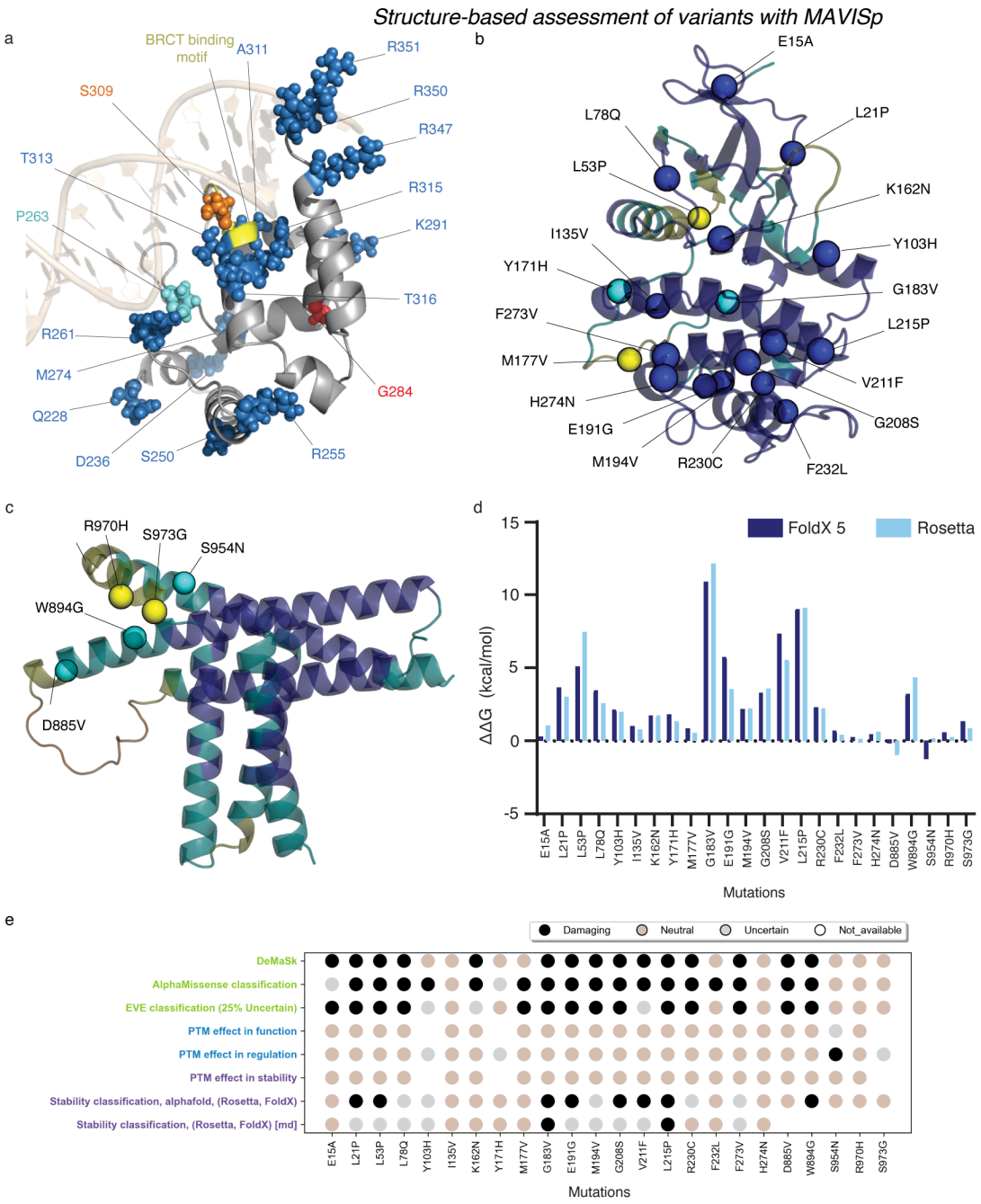

Fig. 2. MAVISp *ensemble mode* to study variants in ARID3A and the destabilizing variants in the autophagy kinase ULK1. a) We characterized 45 variants in the ARID3A DNA-binding domain (residues 228-351). We highlighted the sites where MAVISp identified variants with effects on: i) structural stability (red spheres), ii) PTMs (orange spheres), iii) long-range perturbation (cyan spheres). We mapped the sites on the NMR structure of ARID3A used in the analysis (PDB entry 2KK0) and showed as a reference the location of the DNA from the NMR structure of an ARID3A paralog (PDB entry 1KQQ) as brown transparent cartoon**.** b-c) We identified 24 variants with destabilizing effects in terms of folding free energy, long-range effects, or PTM effects in regulation. The mutation sites are highlighted with spheres on the AlphaFold models of the ULK1 N-terminal (Panel B) and C-terminal (Panel C) domains. d) We showed the predicted changes in folding free energy upon mutations for each of the 38 variants as calculated by the STABILITY module of MAVISp with MutateX and RosettaDDGPrediction. 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 STABILITY, LONG_RANGE, and PTM modules predicted effects on the 24 variants of ULK1 that have been found damaging with at least one MAVISp module.

Only R11A resulted in a destabilizing effect on the binding with p62 using the simple mode, whereas when we applied the ensemble mode this variant turned out with a neutral effect on the binding. Moreover, in the ensemble mode, K51A resulted damaging to the binding. This suggests that the current LOCAL_INTERACTION module for the assessment is stricter than what was originally proposed in the pilot study on MAP1LC3B and in better agreement with the published experimental binding assays[24]. Our assessment also confirms the possibly damaging effect on protein regulation due to the abolishment of a phosphorylation at position 29 (T29A variant), which is not predicted to alter the structural stability of the protein in our assessment. T29 is in a solvent-accessible region on the side opposite to the region where the p62 LIR motif binds and it is also far from the interface for interaction with ATG4B. T29A is predicted as an allosteric mutation that can have

stabilizing long-range effects on the residue N76, which is located at the interface for interaction with ATG4B. Our results disclose a possible long-range mechanism exerted by the phosphorylation of T29 or mutations of this residue.

Chaperones

Chaperones are important cellular players since they assist the folding and unfolding of other proteins. We analyzed using the MAVISp framework two chaperones, i.e., PARK7 and TRAP1. The tumor necrosis factor receptor-associated protein 1 (TRAP1) is a mitochondrial chaperone involved in different biological processes of relevance in cancer[64]. TRAP1 acts a as 'moonlight tumor gene', i.e., can have a dual role as both oncogene and tumor suppressor in a context-dependent manner[65]. We applied the STABILITY, LONG-RANGE, and PTM modules of MAVISp to understand the impact of cancer alterations in the coding region of the gene. We identified 26 variants with destabilizing effects on the protein which could cause loss of function when the protein acts as a tumor suppressor (G160D, G126R, V217A, S219F, W231R, G250R, I255T, R271P, F280L, W299C, I339N, L380P, V391A, F440S, Y444N, S427F, L475P, Y513C, L525P, L539F, R540P, S549Y, L574P, W585R, T600P, and M615R) using a consensus approach between FoldX and Rosetta. When substituting Rosetta with RaSP in the consensus, V217A, S291F, V391A, S472F, S549Y were classified as uncertain. Moreover, 45 variants have the potential to trigger long-range effects and deserve further studies to verify which of them could relate to changes in the conformational changes required for its catalytic cycle[66]. Among these variants, 14 variants are expected to decrease the flexibility at the distal site. The remaining 35 variants have the opposite effect and is predicted to cause enhanced flexibility at the distal sites (see OSF repository or webserver) or are predicted with mixed effects (M141T, R368H, and R495C/H).

Peptidyl-prolyl cys-trans isomerases

Prolyl cis-trans isomerases, such as cyclophilins and FK506-binding proteins, play a central role in integrating protein conformers and signaling events [67]. In the last decades, these enzymes have become interesting targets for the design isoenzyme-specific inhibitors of their activity. Therefore, it is interesting to characterize the landscape of their alterations in cancer samples. We identified 50 mutations for one representative of this enzymes class (PPIA, also known as CypA). The methods used in the STABILITY module have been evaluated for their accuracy against experimentally determined unfolding free energy changes in FKBP1A, a protein with a similar function, in one of our previous studies[43].
Interestingly, we found only three destabilizing variants in terms of change of folding free energy for PPIA (S77F, G75V, and G12V). On the other side, the analysis of FKBP1A revealed 32 mutations at 30 positions reported in COSMIC, ClinVar and/or cBioPortal. Among these, seven are predicted also with long-range destabilizing effects (H26R, S39C/Y, R41W, R58Q, W60C and Q66H), nine with long-range stabilizing effects (P10S, K35N, F47L, K53E, M67V, P94A/S, T97A, G63V and V99I) and two with mixed effects (G29E and F100L. Interestingly, G63V is predicted to be destabilizing to the global protein but stabilizing long-range on position R14 (**Fig. S4**). The damaging effect on stability could suggest a destabilization, causing loss of function, and, eventually, leading to a decreased effect of FKBP1A as an immunoregulator. This regulation was studied in the local interaction between FKBP1A and the TGF-β1 kinase receptor, using the PDB entry 1B6C (chains A and B)[68] and the mutations within 10 Å of the binding interface (21 mutations). This study revealed no stabilizing mutations of interaction which would have suggested increased binding, leading to a reduced immune response.

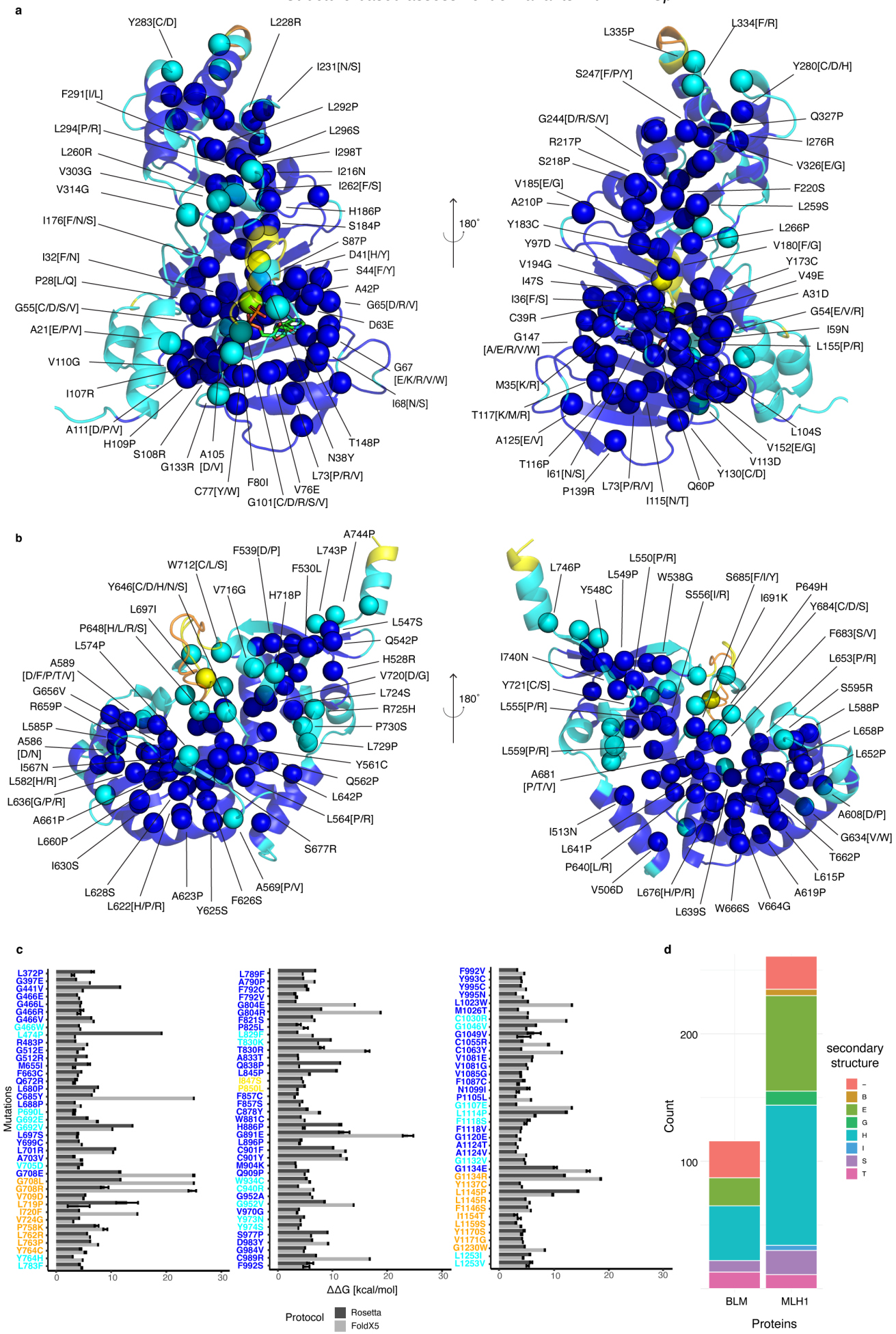
a
Y283[C/D]
L228R
I231[N/S]
F291[I/L]
L292P
L294[P/R]
L296S
L260R
I298T
V303G
I216N
V314G
I262[F/S]
H186P
I176[F/N/S]
S184P
S87P
I32[F/N]
D41[H/Y]
S44[F/Y]
P28[L/Q]
A42P
G55[C/D/S/V]
G65[D/R/V]
A21[E/P/V]
D63E
V110G
G67
[E/K/R/V/W]
I107R
I68[N/S]
A111[D/P/V]
H109P
T148P
S108R
A105
G133R
[D/V]
N38Y
L73[P/R/V]
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
A569[P/V]
L622[H/P/R]
A623P
F626S
Y625S
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
L372P
G397E
G441V
G466E
G466L
G466R
G466V
G469V
L474P
R483P
G512E
G512R
M655I
F663C
Q672R
L680P
C685Y
L688P
P690L
G692E
G692V
L697S
Y699C
L701R
A703V
V705D
G708E
G708L
G708R
V709D
L719P
I720F
V724G
P758K
L762R
L763P
Y764C
Y764H
L783F
L789F
A790P
F792C
F792V
G804E
G804R
F821S
P825L
L829F
T830K
T830R
A833T
Q838P
L845P
I847S
P850L
F857C
F857S
C878Y
W881C
H886P
G891E
L896P
C901F
C901Y
M904K
Q909P
W934C
C940R
G952A
G952V
V970G
Y973N
Y974S
S977P
D983Y
G984V
C989R
F992S
F992V
Y993C
Y995C
Y995N
L1023W
M1026T
C1030R
G1046V
G1049V
C1055R
C1063Y
V1081E
V1081G
V1085G
F1087C
N1099I
P1105L
G1107E
L1114P
F1118S
F1118V
G1120E
A1124T
A1124V
G1132V
G1134E
G1134R
Y1137C
L1145P
L1145R
F1146S
I1154T
L1159S
Y1170S
Y1171G
G1230W
L1253I
L1253V
0
10
20
30
ΔΔG [kcal/mol]
Protocol
Rosetta
FoldX5
d
Count
100
200
secondary structure
–
B
E
G
H
I
S
T
BLM
MLH1
Proteins

Figure 3. 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).

COSMIC Tumor Suppressor Genes and Oncogenes

MAVISp can be applied to both known or newly discovered driver genes in cancer, i.e., tumor suppressors and oncogenes. This application can assist the identification of potential driver mutations in these genes, or to understand the role of Variants of Uncertain Significance (VUS) in cancer predisposition. Here, we applied the *simple mode* of MAVISp to examples of tumor suppressors (such as ACVR1B, ACVR2A, ARID1A, ATR, B2M, BARD1, BAX, BCL10, BRCA1, BLM, BTG1, BTG2, CASP3, CBFB, CTNNA1, CUL3, FANCA, FANCE, FANCG, FANCI, FANCL, ERCC4, GRIN2A, KEAP1, MLH1, MSH2, NBN, PPARG, POLE, POLD1, PTEN, RFWD3, RAD51C and VHL) and oncogenes (such as BRAF, BRD3, BRD4, IDH1, IDH2, CDK4, CALR, EGFR, HRAS, KRAS, PIK3CA, PIM1, RAF1,SRC, and STAT3). As examples of dual role genes in cancer we included TP53, GATA3, RUNX1 and CBL. Furthermore, we included NQO1, which is an important regulator of many tumor suppressor genes and NXF1 which is reported as a candidate driver by the Network of Cancer Genes (NGC)[69]. We also included genes of interest previously identified by the analyses of germline variants in childhood cancer (for example RAD51C)[43]. Additionally, we included in the database the BRCT domain of TP53BP1 since it is central for the regulation of different proteins involved in DNA repair and known to interact with TP53.

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, we report the results of MLH1 and BLM. We identified 116 and 261 destabilizing variants for BLM and MLH1, respectively (**Fig. 3a-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 more disordered regions (**Fig. 3d**). The results on MLH1 are consistent with previous analyses of disease variants performed using free energy calculations[46].

Moreover, in the study of PTEN with MAVISp, we assessed 1238 variants and identified 238 variants with changes in stability from 3 to 18.5 kcal/mol. We compared our results to those reported in other studies on PTEN and identified that at least 176 out of the 238 destabilizing variants altered PTEN abundance levels in cells[70–72]. The results from MAVISp also fits nicely with recent computational studies of PTEN variants using Rosetta calculations of protein stability and analyses of sequence conservation[19,73]. In addition, we applied the LOCAL_INTERACTION module to study the effects of the variants on the interaction between PTEN and SPOP. SPOP is involved in PTEN ubiquitination by binding to the SBC motif (residues 359-363) in the disordered C-terminal tail[74]. The existing X-ray structure of this interaction (PDB entry 401V[75]) only models 15 amino acids long PTEN peptide (residues 354-368). To acquire more information about the flanking regions, we first attempted to use AlphaFold-multimer with the full-length sequences of both interactors. We also experimented with the AlphaFold-Linker protocol[32]. However, we did not manage to reproduce the known binding mode with the obtained models. Therefore, we studied this interaction using the X-ray structure (i.e., PDB entry 4O1V). This allowed us to study 18 variants of PTEN (P357A/L/Q/S/T, E358D/K/Q, A359S/V, S360C/G/T, S361I/N/R, T363A/I), and none was determined to be destabilizing for interaction with SPOP. Interestingly, two variants (A359S, A359V) were annotated as uncertain. In details, FoldX predicts them A359S as a stabilizing variant and A359V as destabilizing, while RossettaDDGPrediction suggests no impact on binding free energies. Bearing in mind the limitations due to local sampling and/or backbone rigidity of these approaches, one of the future steps could be the application of the *ensemble mode* of MAVISp to this complex. On the other hand, the results also make sense from the viewpoint of the tumor suppressor functions of PTEN, where mutations that would disfavor the binding to a regulator promoting its cellular degradation are likely not to provide a selective advantage for cancer evolution.

In the case of the oncogene KRAS, many mutations with long-range effects have been found compared to other case studies analyzed with MAVISp. This finding might suggest that some of these alterations could promote activation of the protein through distal effects and deserve further investigation (**Fig. 4**). For example,

V103A induces stabilization in the residues of the Switch I region, which forms part of the binding interface for effector proteins (PMID: 31988705, PMID: 24889603).

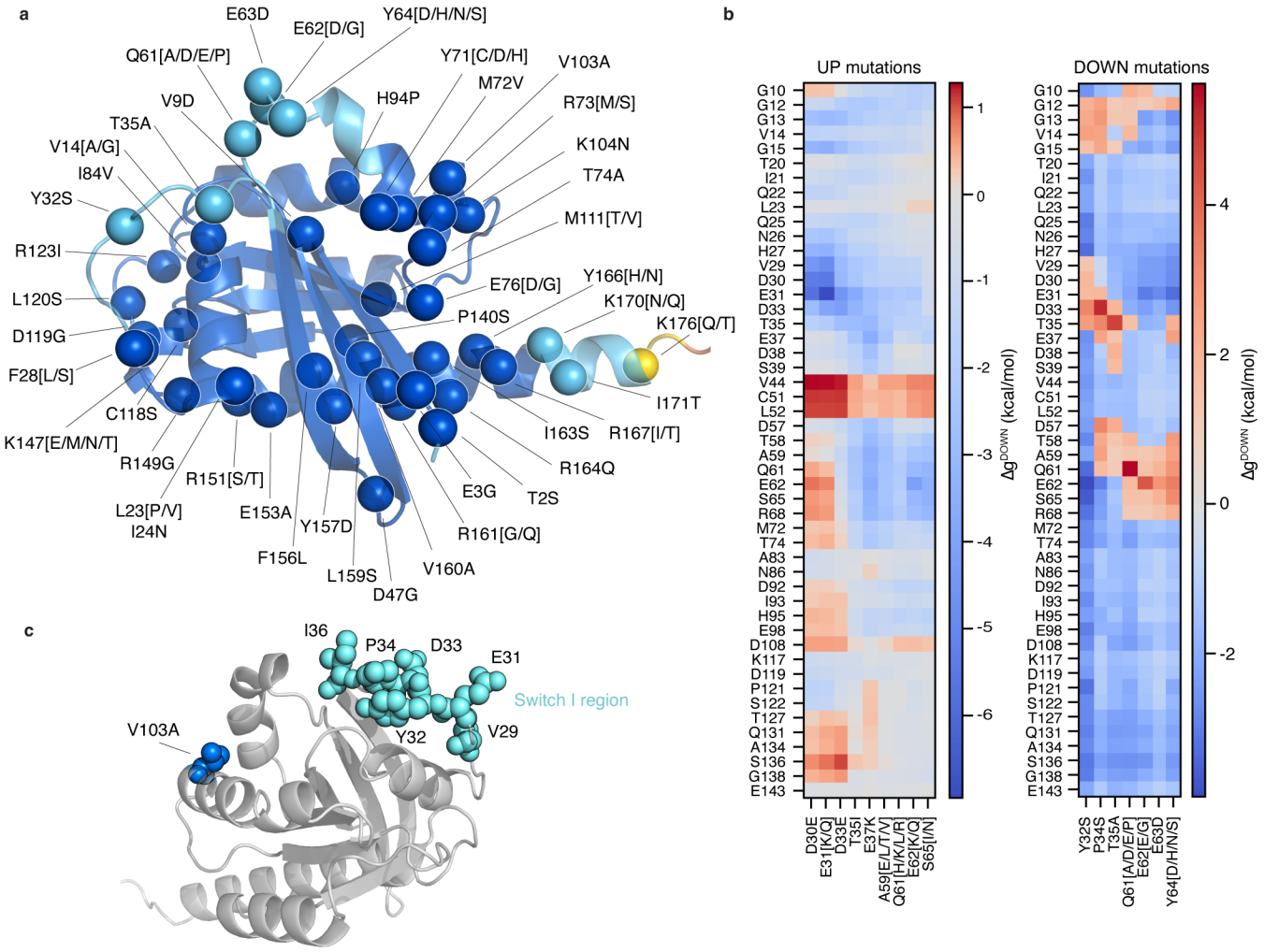

Figure 4. KRAS and allosteric mutations. (a) KRAS AlphaFold model with each residue colored according to the pLDDT score, i.e., blue – very high (pLDDT > 90), cyan - confident (70 < pLDDT <= 90), yellow - low (50 < pLDDT <= 70), and orange - very low (pLDDT <= 50). The C_α of the residues for which DOWN mutations are involved in stabilizing allosteric effects are shown as spheres. **(b)** Heatmaps of the allosteric effects promoted by a subset of UP and DOWN mutations. The colormap indicates the changes in allosteric free energy induced upon mutation; blue means stabilization due to decreased dynamics, while red indicates increased motions (i.e., destabilizing effect). On the x-axis there are the mutations that can have an impact on the allosteric signalling path to a distal site (i.e., at least 8 Å away from the target mutation). The distal sites are shown on the y-axis. **(c)** MAVISp identifies variants with stabilizing effects at distal sites. For example, the V103A mutation induces stabilization in the residues of the Switch I region (i.e., V29,31-34, I36), which forms part of the binding interface for effector proteins.

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

In the current release, MAVISp does not support predictions for the transmembrane regions of proteins within the STABILITY module. A survey on methods to predict folding free energy changes induced by mutations on transmembrane proteins pointed out the weakness of the protocols based on FoldX or Rosetta that can be applied to soluble proteins[76]. Once a transmembrane protein is identified, we thus retain only the variants in the domain that are not in contact with the membrane. As an example, we studied PILRA, which belongs to the immunoglobulin superfamily and play a role in immune response. The transmembrane region is part of the AF model with low accuracy and, thus, we only analyzed variants located in the region 32-153. The application of MAVISp to a protein target such as PILRA also points out the relevance of the structure-based framework to study the effects of variants in other diseases. In fact, PILRA is an important protein target in connection with neurodegenerative diseases[77], along with KIF5A, CFAP410, and CYP2R1. Interestingly, we found 55 variants for PILRA with a predicted destabilizing effect that could suggest a potential pathogenic role for these variants. In addition, we included other transmembrane proteins in the database such as ATG9A and EGFR. Proteins associated with other diseases than cancer have been also included, such as TTR and SOD1.

Application of MAVISp for classification of variants in tumor suppressor and for VUS characterization

Furthermore, we aimed at exploring the potential of MAVISp beyond variant interpretation, as a tool for classification. To this goal, we retained a subset of proteins that are known tumor suppressor genes and variants found in ClinVar with a review status of 2 or higher for the proteins in the MAVISp database. This property has been showed before to capture pathogenic mutations in tumor suppressors [43,45,46], where mutations are mostly linked to loss-of-function [78], making tumor suppressors a good dataset to study. We analyzed the 35 tumor suppressor proteins available in the MAVISp database and a subset of 167 pathogenic and 75 benign variants in ClinVar. It should be noted that, in the current release, not all the proteins have been necessarily analyzed for all their possible biological interactors or PTMs. Thus, we attempted a first classification only based on changes in structural stability, as a proof of principle (**Fig. 5a**). Thus, if a known pathogenic variant is predicted as neutral, it might mean that additional layers of evaluation are needed to achieve a full picture. On the contrary, we would expect that most (or virtually all) variants currently classified as benign in ClinVar would be predicted to have neutral effects on stability, if our framework is effective. Interestingly, all the variants predicted as damaging with these MAVISp modules are pathogenic according to ClinVar except for the variant E198G in MSH2. Mutations to glycines encounter the risk to be overestimated by the local conformational sampling provided by FoldX. Rosetta predicts a change very close to the 3 kcal/mol threshold for E198G, suggesting that this variant would be an interesting case to analyze, in future studies, using the *ensemble mode* of MAVISp to account for conformational changes during the assessment. Indeed, E198G has been showed as a benign mutation also in experimental studies and predictions with EVE in a previous study[55]. In addition, we noticed that the same analysis using a consensus based on FoldX and RasP yielded a higher proportion of uncertain variants (**Fig. 5b**).

Moreover, we then retained the 92 variants for which all the methods could provide a classification without uncertainty, and we compared the performances of the different methods in classifying pathogenic and benign variants (**Table 1, Fig 5c-f**). As observed in the analysis above, the classification according to MAVISp has very high precision (i.e., correctly identify the true positive variants) and it is higher or comparable to the performances of the other pathogenicity scores. On the other side, the sensitivity is very low when compared to the new generation pathogenicity scores, supporting the notion that the additional modules of MAVISp can provide further mechanistic indicators for the underlying pathogenic effect. It should be considered that these results serve as an example for illustrative purposes of the potential of MAVISp even beyond the original aim of variant interpretation. More in-depth investigation and a much larger dataset of predictions performed with MAVISp will be needed to further verify these results and the potential to use the data for classification.

At last, we evaluated how MAVISp can shed light on VUS. To this aim, we identified a group of 114 variants reported as 'Uncertain significance' in familial cancer predisposing syndromes in ClinVar querying the entire MAVISp database. The variants have been selected so that the corresponding protein entry was associated with a MAVISp review status of 2 to guarantee the quality of the data used in the analysis (**Table S2**). The variants were found mostly in BRCA1 (36 variants) and RAD51C (73 variants). Moreover, we found two variants in PPARG (D144G, D19A), and one in RAF1 (N553H), GRIN2A (K160R) or ALPK3 (T1190M). Of these, we identified T1684P, V1810G, and V1842E for BRCA1 as destabilizing in terms of structural stability (**Figure 5g**). Notably, the destabilizing mutations were found in the BRCT tandem domains region, suggesting a potential crucial impact on BRCA1 functionality. This region is involved in numerous phospho-dependent interactions, regulated by DNA damage response kinase, which results in signal transduction for the assembly of protein-protein complexes at discrete nuclear foci, initiating the DNA damage response (DDR). Interestingly the MAVISP stability prediction for V1810G agrees with the findings of a study where the destabilizing effect of the mutation was validated in a cellular model[79]. Furthermore, T1684P and V1842E were classified as non-functional based on a Homology Directed Repair assay combined with Saturation Genome Editing[80]. This result aligns with the MAVISP readout, which provides as a possible underlying mechanism a change in stability of the protein product in the T1684P and V1842E variants.

On the other hand, A1708V, F1734L, and E1754Q were predicted to exert damaging allosteric effects (**Figure 5g-h**). Specifically, A1708V was predicted to have a destabilizing effect associated with an increase in flexibility (**Figure 5h**), while F1734L and E1754Q were predicted to have mixed effects (**Figure 5h**). These mutations are located in the BRCT tandem domain region, particularly in the first domain, except for E1754Q, which occurs in the linker between the two domains. However, these mutations, along with the affected long-range residues, do not directly impact the phosphopeptide binding site, involving residues S1655, G1656,K1702 as the first anchor; R1699, L1701, F1704, N1774, M1775,L1839 as the second anchor; V1696, C1697, and V1740 as hydrophobic patch 1; L1657 and P1659 as hydrophobic patch 2 and V1654, F1662, M1633 and Y1666 as hydrophobic patch 3[81]. Additionally, it has been shown that BRCT domains of BRCA1

can establish independent interactions, as demonstrated for DNA-PKcs[82]. Therefore, the residues affected by the mutations could be involved in phosphorylation-independent interactions with the BRCA1 partner. However, the studies conducted on these mutations were unable to fully clarify their impact. For example, A1708V was classified with an intermediate effect through HDR, transcription activation, and Centrosome amplification assays[83] while no literature regarding the impact of E1754Q on BRCA1 functionality has been found.

Regarding RAD51C, C104R, G153D, V156G, and I240T have been predicted to be destabilizing in terms of structural stability (**Figure 5g**). Intriguingly, these mutations occur in the region binding the RAD51 paralog, members of the BCDX2 and CX3 complexes involved in genome stability and DNA damage response. Thus, these destabilizing mutations of the RAD51C structure could impact complexes assembly and consequently their function. Except for G153D, which resulted non-functional according to assays in yeast[84], the other variants are poorly characterized and functionally validated. Therefore, C104R, V156G, and I240T appear to be promising candidates for further functional validation.

Table 1. Performances of the pathogenicity scores EVE, DeMask and Alphamissense, as well of a classification based on changes in stability or in stability upon the loss of a phosphorylation as predicted by MAVISp using tumour suppressor proteins.

	MAVISp stability	AlphaMissense	EVE – 25% Uncertain	DeMaSk
Sensitivity	0.554	0.954	0.908	0.908
Specificity	1.000	0.913	0.696	0.913
Accuracy	0.670	0.943	0.852	0.909
Precision	1.000	0.969	0.894	0.967
F1 score	0.713	0.961	0.901	0.937

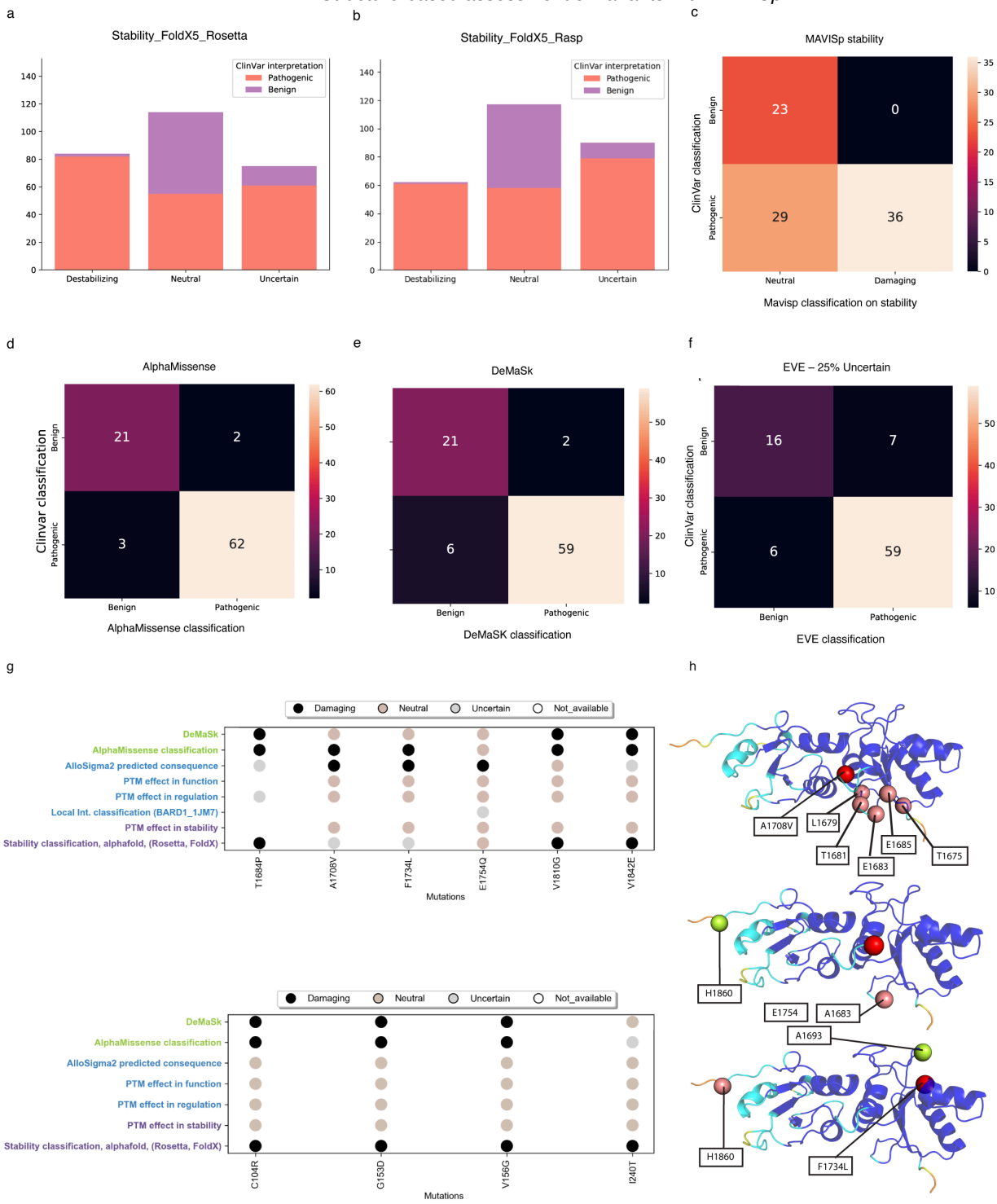

Figure 5. Example of usage of MAVISp data for classification of variants based on protein stability and to characterize Variant of Uncertain Significance in tumor suppressor genes.

(a-b) The stacked bar plots illustrates the results of a classification of variants reported as benign or pathogenic in ClinVar with a review status of 2 or higher using the consensus prediction between Rosetta and FoldX (a) or RasP and FoldX (b) from MAVISp. (c-f) Confusion matrices from variant classification using the MAVISp Stability Module (c), AlphaMissense (d), DeMaSk (e), and EVE (f) in comparison with the classification reported in ClinVar for variants with review status of 2 or higher. The corresponding performances are reported in Table 1. (g) The dotplots illustrated the results of the aggregated analyses from MAVISp for VUS in BRCA1 and RAD51C for which we could associate a mechanistic indicator with MAVISp at the base of a possible pathogenic effect. (h) Illustration of the variants with long-range effects in BRCA1 and the response residues. The mutation site is shown in red, the response sites with a destabilizing effect upon mutation in light red, whereas the response sites with a stabilizing effect are colored in green.

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. 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 a comprehensive atlas of functional annotations for cancer-related variants or variants related to other diseases. MAVISp can help to understand the effects of VUS or variants with conflict evidence at the molecular level.

Longer term, the assessment provided by MAVISp might have the potential to be used for the classification of variants, as exemplified here in a case study on tumor suppressor genes. In this context, future studies are needed to benchmark against variants with known effects to estimate the accuracy, sensitivity, and specificity of the different layers and their collective contributions. To this goal, the data collection that we designed and present here is pivotal to produce a handful of data to build meaningful predictive models.

Recently, 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[85] . 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. Indeed, to focus on functional variants with marginal effect on stability, we used the STABILITY module and retained for further analyses only those variants with neutral or unknown effects as estimated by changes in folding free energy. The analyses were carried out for TP53, CALR, EGRF, GRIN2A, and CBFB, which are now included in the MAVISp database. Other MAVISp modules were used as well to obtain a more comprehensive overall picture. 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.

MAVISp can also be extended to other disease-related variants and does not necessarily need to focus only on variants reported in cancer studies. 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. In the next releases, we also aim to include, as source of experimental validation, experimentally derived scores in the MAVE database[86], which includes data from multiplex assays of variant effects (MAVE).

At this stage, it is possible to provide MAVISp annotations for variants in regions of transmembrane proteins that are external to the membrane, as exemplified with PILRA, EGFR, TMEM175 and ATG9A. Approaches as the one provided by Tiemann et al.[87] 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 mutations 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

We used models from the AlphaFold2 database[21] for most of the target proteins and trimmed them to remove regions with pLDDT scores < 70 at the N- or C-termini or long disordered linkers between folded domains. For proteins coordinating cofactors we re-modeled the relevant cofactors upon analyses with AlphaFill[28] and where needed through a step with MODELLER[29]. We used an X-ray (PDB entry 3K0N[88]) and an NMR structure (PDB entry 2KK0[89]) for PPIA and ARID3A, respectively. For AMBRA1 we used a recently solved CryoEM structure in complex with DDB1[90], isolating the chain for AMBRA1 and reconstructing missing coordinates with MODELLER. Additionally, in the case of ALPK1 AlphaFold2 model, we designed a 20-residues linker to connect the residues 495 and 956, including 10 residues from the N- and the C-terminal regions to avoid modelling a long unstructured region with low pLDDT score. 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)[91,92].

Interactome

In the INTERACTOME module, we used the Mentha2PDB tool to identify known interactors of the target protein by extracting data from the Mentha database[30] 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.[31]. Mentha2PDB provides annotations of the interactors and generates input files for AlphaFold-Multimer. Mentha2PDB can be found freely available at https://github.com/ELELAB/Mentha2PDB.

Free energy calculations

We applied the BuildModel module of FoldX5 suite[39] 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[41].
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[41]. We also applied RaSP using the same protocol provided in the original publication[44] and adjusting the code in a workflow according to MAVISp-compatible formats (https://github.com/ELELAB/RaSP_workflow).
For the calculations of allosteric free energy, we used the structure-based statistical mechanic model implemented in AlloSigMA2[49] . 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[25]. 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 8 Å and which belongs to pockets as identified by Fpocket[93].

Molecular dynamics simulations

For ARID3A, MEFV, MAP1LC3B, NQO1, PPIA, PARK7, $ULK1_{3-279}$, $MZF1_{37-128,}$ and the DNA binding domain we used either previously published[24,25,59,60,94,95] or collected within this study all-atom molecular dynamics simulations of one microsecond with the CHARMM22star or CHARMM36m force fields[96]. 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 will be provided in OSF. As an 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[25]. The trajectories will be deposited as different entries in OSF.

CABS-flex ensembles

We used the coarse-grained CABS-flex 2.0 method and software[35] as a part of a Snakemake[97] 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[98] and the SOV-refine score[99].

Mutation impact predictors

We used DeMaSk[54], GEMME[17], EVE[55], REVEL[53] and AlphaMissense[56] as predictors for the effect of mutations of any possible mutation to natural amino acids, on the full protein sequence of the main UniProt[100] 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. 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. 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[97]-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 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 (10.5281/zenodo.8360242).

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/. Reports for each protein are available at https://elelab.gitbook.io/mavisp/.

References

1. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science (1979)* **339**, 1546–1558 (2013).
2. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: A looking glass for cancer? *Nat Rev Cancer* **12**, 323–334 (2012).
3. 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).
4. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. 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).
6. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biology* vol. 24 Preprint at https://doi.org/10.1186/s13059-023-02986-x (2023).
7. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation* vol. 43 1012–1030 Preprint at https://doi.org/10.1002/humu.24309 (2022).
8. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).
9. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
10. 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.
11. Bouaoun, L. *et al.* TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Hum Mutat* **37**, 865–876 (2016).
12. Plazzer, J. P. *et al.* The InSiGHT database: Utilizing 100 years of insights into Lynch Syndrome. *Fam Cancer* **12**, 175–180 (2013).

13. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835–D844 (2020).
14. Burke, W., Parens, E., Chung, W. K., Berger, S. M. & Appelbaum, P. S. The Challenge of Genetic Variants of Uncertain Clinical Significance: A Narrative Review. *Annals of Internal Medicine* vol. 175 Preprint at https://doi.org/10.7326/M21-4109 (2022).
15. 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. in *Protein Homeostasis Diseases* 91–107 (Elsevier, 2020). doi:10.1016/b978-0-12-819132-3.00005-1.
16. 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).
17. Laine, E., Karami, Y. & Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Molecular Biology and Evolution* vol. 36 2604–2619 Preprint at https://doi.org/10.1093/molbev/msz179 (2019).
18. 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).
19. 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).
20. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv* (2022) doi:10.1101/2021.10.04.463034.
21. 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).
22. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
23. 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).
24. Fas, B. A. *et al.* The conformational and mutational landscape of the ubiquitin-like marker for autophagosome formation in cancer. *Autophagy* 1–24 (2020).
25. 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).
26. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
27. Degn, K., Beltrame, L., Tiberti, M. & Papaleo, E. PDBminer to Find and Annotate Protein Structures for Computational Analysis. *BioRXiv* (2023).
28. 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).
29. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).
30. Calderone, A., Castagnoli, L. & Cesareni, G. Mentha: A resource for browsing integrated protein-interaction networks. *Nat Methods* **10**, 690–691 (2013).
31. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat Struct Mol Biol* **30**, (2023).
32. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun* **13**, (2022).
33. Di Rita, A. *et al.* HUWE1 E3 ligase promotes PINK1/PARKIN-independent mitophagy by regulating AMBRA1 activation via IKKα. *Nat Commun* **9**, 3755 (2018).
34. Holdgaard, S. G. *et al.* Selective autophagy maintains centrosome integrity and accurate mitosis by turnover of centriolar satellites. *Nat Commun* **10**, 1–19 (2019).
35. 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).
36. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2**, 19–25 (2015).
37. 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).

38. Bonomi, M. *et al.* Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* Preprint at https://doi.org/10.1038/s41592-019-0506-8 (2019).
39. 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.
40. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J Chem Theory Comput* **12**, 6201–6212 (2016).
41. Frenz, B. *et al.* Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front Bioeng Biotechnol* **8**, (2020).
42. Tiberti, M. *et al.* MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* **23**, bbac074 (2022).
43. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. *Protein Science* **32**, e4527 (2023).
44. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *bioRxiv* (2022) doi:10.1101/2022.07.14.500157.
45. Nielsen, S. V. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* **13**, e1006739 (2017).
46. 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.
47. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J Phys Chem B* **122**, 5389–5399 (2018).
48. Sora, V., Tiberti, M., Mahdi Robbani, S., Rubin, J. & Papaleo, E. PyInteraph2 and PyInKnife2 to analyze networks in protein structural ensembles. *bioRxiv* (2020) doi:https://doi.org/10.1101/2020.11.22.381616.
49. 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).
50. 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).
51. 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).
52. Henriques, J. & Lindorff-Larsen, K. Protein Dynamics Enables Phosphorylation of Buried Residues in Cdk2/Cyclin-A-Bound p27. *Biophys J* **119**, 2010–2018 (2020).
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. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (1979)* **381**, eadg7492 (2023).
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. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* vol. 12 31–46 Preprint at https://doi.org/10.1158/2159-8290.CD-21-1059 (2022).
59. 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).
60. Nygaard, M. *et al.* The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci* **3**, (2016).
61. 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).
62. Suzuki, H., Osawa, T., Fujioka, Y. & Noda, N. N. Structural biology of the core autophagy machinery. *Curr Opin Struct Biol* **43**, 10–17 (2017).

63. Shu-Yong, L. *et al.* GSK3-TIP60-ULK1 signaling pathway links growth factor deprivation to autophagy. *Science (1979)* **336**, 477–481 (2012).
64. Masgras, I., Sanchez-Martin, C., Colombo, G. & Rasola, A. The Chaperone TRAP1 As a Modulator of the Mitochondrial Adaptations in Cancer Cells. *Front Oncol* **7**, 1–10 (2017).
65. Matassa, D. S., Agliarulo, I., Avolio, R., Landriscina, M. & Esposito, F. Trap1 regulation of cancer metabolism: Dual role as oncogene or tumor suppressor. *Genes (Basel)* **9**, 1–12 (2018).
66. Elnatan, D. *et al.* Symmetry broken and rebroken during the ATP hydrolysis cycle of the mitochondrial Hsp90 TRAP1. 1–25 (2017) doi:10.1101/107094.
67. Theuerkorn, M., Fischer, G. & Schiene-Fischer, C. Prolyl cis/trans isomerase signalling pathways in cancer. *Current Opinion in Pharmacology* vol. 11 281–287 Preprint at https://doi.org/10.1016/j.coph.2011.03.007 (2011).
68. Huse, M., Chen, Y.-G., Massagué, J. & Kuriyan, J. Crystal Structure of the Cytoplasmic Domain of the Type I TGF β Receptor in Complex with FKBP12. *Cell* **96**, 425–436 (1999).
69. 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).
70. 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).
71. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874–882 (2018).
72. 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).
73. 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.
74. Guo, Z. Q. *et al.* Small-Molecule Targeting of E3 Ligase Adaptor SPOP in Kidney Cancer. *Cancer Cell* **30**, 474–484 (2016).
75. Li, G. *et al.* SPOP promotes tumorigenesis by acting as a key regulatory hub in kidney cancer. *Cancer Cell* **25**, 455–468 (2014).
76. 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? *Wiley Interdisciplinary Reviews: Computational Molecular Science* vol. 9 Preprint at https://doi.org/10.1002/wcms.1410 (2019).
77. Li, Y. *et al.* Genomics of Alzheimer's disease implicates the innate and adaptive immune systems. *Cellular and Molecular Life Sciences* vol. 78 7397–7426 Preprint at https://doi.org/10.1007/s00018-021-03986-5 (2021).
78. Wang, L. H., Wu, C. F., Rajasekaran, N. & Shin, Y. K. Loss of Tumor Suppressor Gene Function in Human Cancer: An Overview. *Cellular Physiology and Biochemistry* **51**, 2647–2693 (2018).
79. Adamovich, A. I. *et al.* The functional impact of BRCA1 BRCT domain variants using multiplexed DNA double-strand break repair assays. *Am J Hum Genet* **109**, 618–630 (2022).
80. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
81. Wu, Q., Jubb, H. & Blundell, T. L. Phosphopeptide interactions with BRCA1 BRCT domains: More than just a motif. *Progress in Biophysics and Molecular Biology* vol. 117 143–148 Preprint at https://doi.org/10.1016/j.pbiomolbio.2015.02.003 (2015).
82. Davis, A. J. *et al.* BRCA1 modulates the autophosphorylation status of DNA-PKcs in S phase of the cell cycle. *Nucleic Acids Res* **42**, 11487–11501 (2014).
83. Lovelock, P. K. *et al.* Identification of BRCA1 missense substitutions that confer partial functional activity: Potential moderate risk variants? *Breast Cancer Research* **9**, (2007).
84. Clague, J. *et al.* RAD51C germline mutations in breast and ovarian cancer cases from high-risk families. *PLoS One* **6**, (2011).
85. 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. *bioRxiv* (2022) doi:10.1101/2022.12.11.519943.

86. Esposito, D. *et al.* MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**, (2019).
87. 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.
88. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).
89. Liu, G. *et al.* Solution NMR structure of the ARID domain of human AT-rich interactive domaincontaining protein 3A: A human cancer protein interaction network target. *Proteins: Structure, Function and Bioinformatics* **78**, 2170–2175 (2010).
90. Liu, M. *et al.* Structural basis for substrate recruitment by AMBRA1 E3 ligase receptor 1 2. *bioRxiv* (2022) doi:10.1101/2022.12.04.519012.
91. 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).
92. 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**, (2012).
93. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11 (2009).
94. 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).
95. 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).
96. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100**, (2011).
97. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33 (2021).
98. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* (1983) doi:10.1002/bip.360221211.
99. 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).
100. Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
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).

Author Contributions (CRediT Classification)

Conceptualization: EP; *Data Curation*: MA, LB, KD, MU, SS, TD, FM, PSIB, KK, PB, AEH, KM, JBB, ADN, JV, JS, ML, MT, EP. *Formal Analysis*: MA, LB, KD, MU, TD, SS, FM, ML, EP. *Funding Acquistion:* EP, MA, CC, SI, MN, AR. *Investigation*: MA, LB, KD, JS, TD, FM, MU, AR, MN, SI, EP. *Methodology*: MA, LB, KD, AP, VS, ML, MT, EP. *Project administration:* EP. *Resources*: PWD, EP, SI, CC. *Software*: MA, LB, KD, JV, MT, PWS. *Supervision*: EP, ML, MT, SS, MU, MA, LB. *Validation*: MA, LB, KD, MU, PSIB, KK, SS, PB, ML, MT, EP. *Visualization*: MA, LB, KD, MU, SS, TD, ML, 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) 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. 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.