MAVISp: Multi-layered Assessment of Varlants by Structure for proteins

Matteo Arnaudi[1,2†], Ludovica Beltrame[1†], Kristine Degn[1†], Mattia Utichi[1,2], Pablo Sánchez-Izquierdo[1,2], Simone Scrima[2], Francesca Maselli[2,3], Karolina Krzesińska[1,2], Terézia Dorčaková[2], Jordan Safer[4], Katrine Meldgård[1,2], Julie Bruun Brockhoff[1,2], Amalie Drud Nielsen[1,2], Alberto Pettenella[2], Jérémy Vinhas[2], Peter Wad Sackett[1], Claudia Cava[5], 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 Society Research Center, 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

† These three authors equally contributed to the work

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

Cancer is a complex group of diseases due to the accumulation of mutations in tumor suppressors or oncogenes in the genome. Cancer alterations can be very heterogeneous, even in tumors from the same tissue, affecting cancer predisposition, response to treatment, and risks of relapse in different patients. The role of genomics variants in this context continues to be realized. Thanks to advances in sequencing techniques and their introduction in the clinics, the number of genomic variants discovered is growing exponentially. Many of these variants are classified as Variants of Uncertain Significance (VUS), while other variants have been reported with conflicting evidence or with a 'likely' effect. Applications of bioinformatic-based approaches to characterize protein variants demonstrated their full potential thanks to advances in machine learning, comparisons between predicted effects and cellular readouts, and progresses in the field of structural biology and biomolecular simulations. We here introduce a modular structure-based framework for the annotations and classification of the impact of variants affecting proteins or their interactions and impacting on the corresponding protein product (MAVISp, Multi-layered Assessment of Varlants by Structure for proteins) together with a Streamlit-based web application (https://github.com/ELELAB/MAVISp) where the variants and the data generated by the assessment are made available to the community for consultation or further studies. Currently, MAVISp includes information for 127 different proteins and approximately 42000 variants. New protein targets are routinely analyzed in batches by biocurators through standardized Python-based workflows and high-throughput free energy and biomolecular simulations. We also illustrate the potential of the approach for each protein included in the database. New variants will be deposited on a regular base or in connection with future publications where the approach will be applied. We also welcome new contributors who are interested in participating to the collection in relation to their 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 suppressors or oncogenic functions[1,2]. Cancer alterations are heterogenous even when considering tumors from the same tissue types. This heterogeneity affects predisposition to cancer, treatments response, and risk of relapse[3]. On the other side, it also makes challenging to understand the molecular mechanisms and effects related to the alterations that can be found by sequencing approaches applied to tumor and healthy samples. We are witnessing unprecedented times with advances in cancer genomics, sequencing[4], structural biology[5,6,] and high-throughput multiplex-based assays.[7–10] Consequently, many Variants of Uncertain Significance (VUS) or conflicting classifications are continuously identified and hard to rationalize. Many variants have also been reported in different publicly available repositories, including polymorphisms observed in healthy individuals[11,12], compendia of cancer-related variants[13–15] on focusing on specific genes[16] or hereditary cancer types[17], and classifying the variants from a clinical point of view[18,19]. Similar challenges apply to other diseases. The bioinformatics community has been providing new approaches to predict the pathogenic impact of variants with often good agreement with experiments and cellular readouts [20–25] thanks to advances in machine learning approaches and a large amount of available data.

The advent of Alpha Fold[5,26,27] provides the possibility to get 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[28].

In light of the context above, in three recent pilot projects[29–31], we provide 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 thus have consequences on the protein product. For example, the framework can be used to prioritize variants for experimental and functional studies [30]. It can also complement other pathogenicity scores since the overall goal is to link the effects of the variants to the specific underlying molecular mechanisms[31]. However, pathogenicity scores only estimate the likelihood of a variant being pathogenic. They lack a clear 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 cancer prevention, genetic counseling, clinical care, and drug treatments. From a fundamental research viewpoint, this mechanistic knowledge is essential in designing and prioritizing experiments to understand the underlying molecular details.

Here, we consolidated and generalized our structure-based framework to develop MAVISp (Multi-layered Assessment of VarIants by Structure for proteins) to handle high-throughput variant analysis within common standardized workflows. The data are also released as a Streamlit-based web application for consultation and download for further studies.

We release the results for more than 120 proteins for illustrative purposes of how to work within MAVISp. New data will be introduced in connection with the related publications where the assessment will be applied.

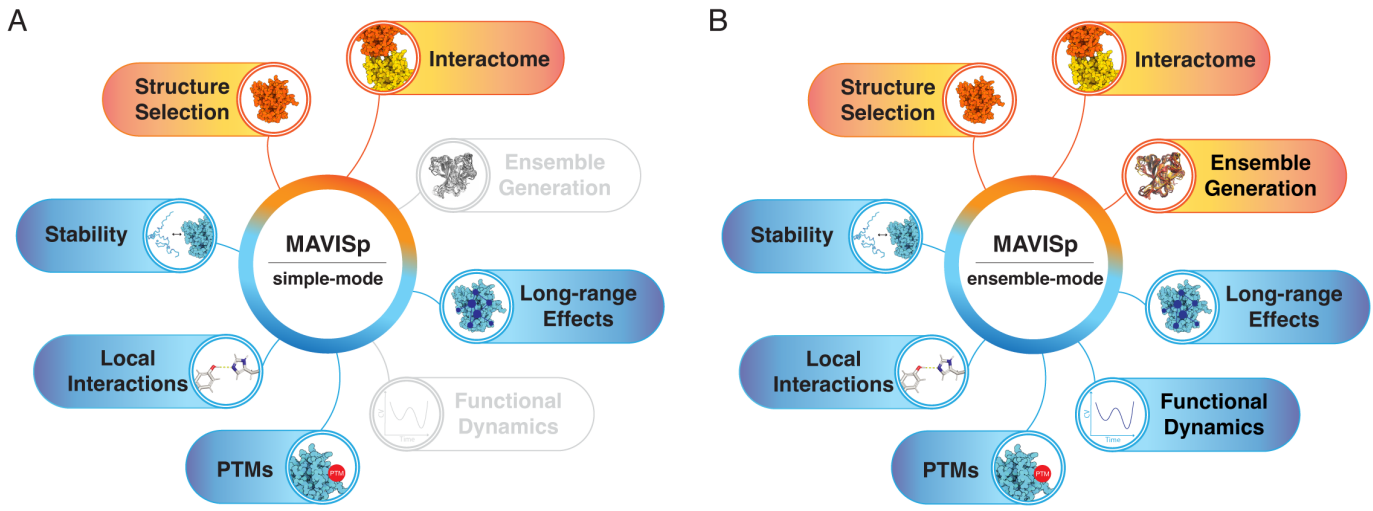


Figure 1. Overview of MAVISp components. MAVISp inlcludes different modules, each handled by workflow engines or dedicated tools. The modules in orange are the ones related to the selection/collection of the protein structures. The modules highlighted in light blue are devoted to the analyses. The procedure is applied to start from the gene name and a list of variants of interest. All the steps can be carried out on a standard server with approximately 24/32 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. Panel A illustrates the *simple mode* for the assessment, which uses single experimental structures or models from AlphaFold2. Panel B illustrates the *ensemble mode,* in which a conformational ensemble for the target protein or its complexes is applied.

Results

Overview of MAVISp and its database

MAVISp can be applied on individual three-dimensional (3D) structures of proteins and their complexes (*simple mode*) or an ensemble of structures (*ensemble mode*), as illustrated in **Figure 1**. We designed the framework in a modular way 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 categories: i) modules to retrieve and select the structures for the analyses (in orange in **Figure 1**) and modules for the different layers of the assessment (in blue in **Figure 1**).

The overall workflow of MAVISp (**Figure 2A**) 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 process them, performs sanity checks and write the results as human-readable csv files, which constitute the MAVISp database. The MAVISp web app reads the database and displays it as a user-accessible website, from which the data tables can be downloaded.

As mentioned above, we will release the results of the analyses in a Streamlit-based web server (**Figures 2B**), which is currently under development and available as a user-runnable MAVISp web app. We plan to release the web app as a fully hosted website that will be available for the general public as the MAVISp web server. The MAVISp database, MAVISp Python package and MAVISp web app are available on our GitHub repository (https://github.com/ELELAB/MAVISp) and released under an open-source license.

We will add newly available data batches on a regular basis. We provide updates 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 with the MAVISp framework, the PMID or DOI of the associated publication will be annotated in the MAVISp database.

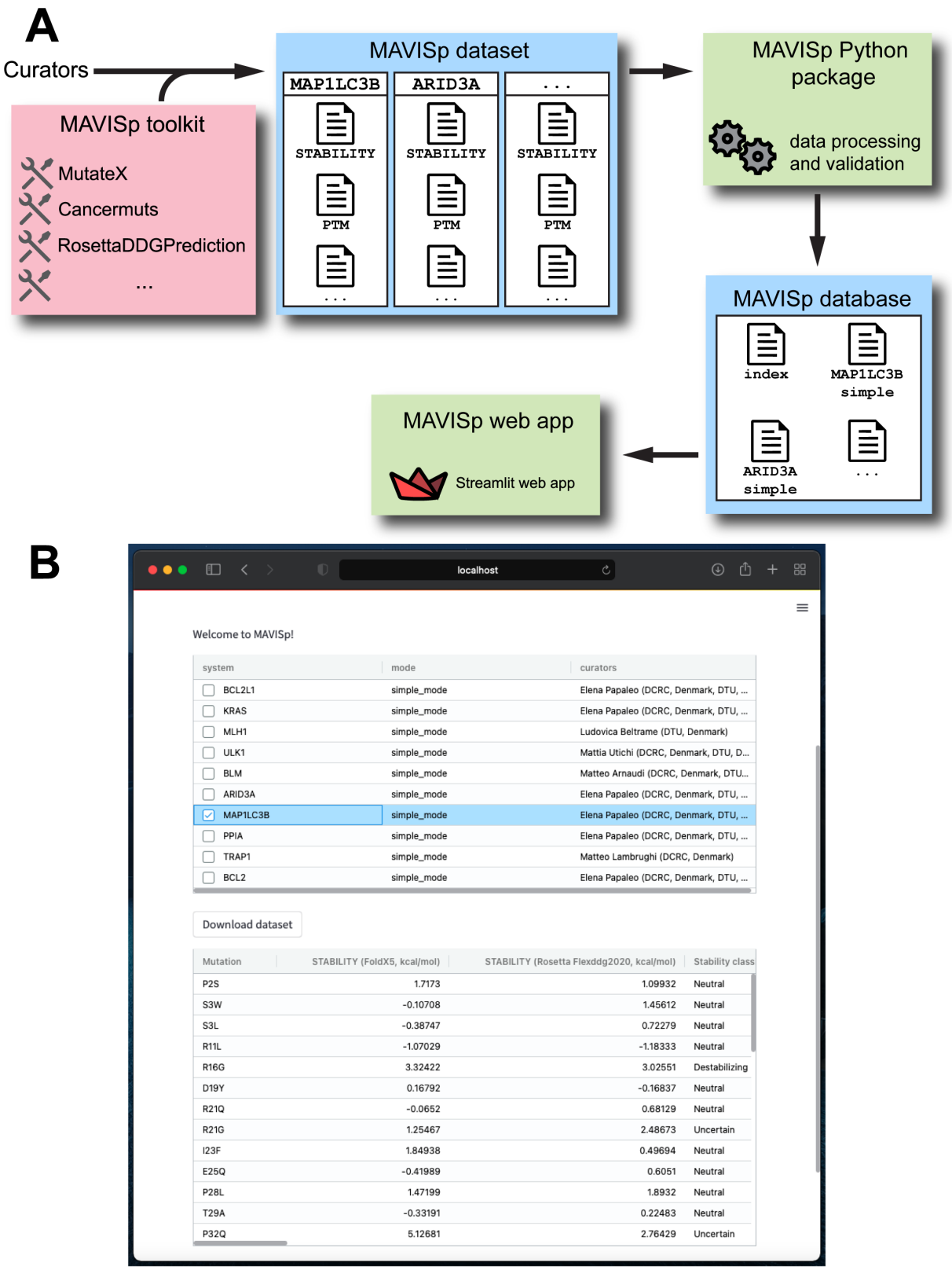


Figure 2. The MAVISp web app. (A) Scheme of the current workflow for the MAVISp database and app. Curators use specific workflows and protocols within each MAVISp module to derive structure-based prediction of the changes connected to the variants identified for a protein target. The results are gathered 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 a Streamlit web app, which further processes it and makes the data available both for interactive visualization and download. (B) Screenshot of the current main page of MAVISp web app.

We welcome requests on specific targets or variant lists that are not already available in MAVISp or in the pipeline for curation. We also welcome external contributors as curators, after training on the protocols to follow and adherence to our guidelines, and we will provide the needed resources where needed. Currently, we only provide the data on a single selected complex for each protein with interactors in the database. The support to annotations for the effects of protein variants on different biological interface for the same target protein is under development and will part of the next release. We also support as an independent MAVISp module the effects of variants on the interaction with DNA. Updates and revisions of entries will be carried out every six months 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 expand the tools to apply so that even more comprehensive assessments can be achieve.

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* (**Figure 1**). 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 allows the interaction with the Protein Data Bank (PDB) [32], the AlphaFold Protein Structure Database [27], or to run AlphaFold and AlphaFold-multimer [26]. 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).

The INTERACTOME module streamlines the identification of protein interactors for the target protein querying the Mentha database [33] and PDB. It also generates input files to predict the complexes with AlphaFold-multimer and retains only the models with reasonable accuracy in terms of pLDDT and Predicted Aligned Error (PAE) scores. Models that do not feature a low PAE score but provide suitable interaction interfaces should be first validated experimentally before using them in the assessment.

The ENSEMBLE_GENERATION module allows collecting coarse-grained models for protein flexibility (with CABS-flex [34,35]) or all-atom Molecular Dynamics (MD) simulations (with GROMACS [36]) 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 - alone or with interactors - the analysis modules can be applied.

The minimal requirement to include a protein and its variants in MAVISp is the application of the STABILITY and PTM modules. 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[37], the cartddg2020 protocol, and the ref2015 energy function of Rosetta[38,39]. This part takes advantage of our recently developed workflows for high-throughput in-silico mutagenesis scans, i.e., MutateX [40] and RosettaDDGPrediction[41]. Moreover, we added the estimates of changes in folding free energy provided by RaSP[42]. MAVISp applies a consensus approach between two methods. Variants with changes in stability below 3 kcal/mol are not featuring marked decreased stability at the cellular level [43,44]. Thus, MAVISp defines the following classes for changes in stability: stabilizing ($\Delta\Delta G$ =< - 3kcal/mol with both methods), destabilizing ($\Delta\Delta G$ >= 3 kcal/mol), neutral (-2 =<$\Delta\Delta G$ <=2 kcal/mol), and uncertain. The uncertain class result from marked disagreement between the two methods or if the prediction is in between 2-3 kcal/mol or -3/-2 kcal/mol with both methods. We mostly rely on the classification based on the consensus between Rosetta and FoldX. We included RaSP since we plan to compare on a larger dataset the influence of using a consensus among FoldX and RaSP to evaluate if we can 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 energy function and the flexddg protocol and talaris2014 energy function of Rosetta [45]. In this step, we also apply Mutatex and RosettaDDGPrediction.

In the *ensemble mode*, the calculations for the STABILITY and LOCAL_INTERACTION modules can be applied to a collection of 50 or 25 frames from the simulations (FoldX-based calculations)[40] or on the three main representative structures of the ensemble after structural clustering carried out with GROMACS (Rosetta-based calculations). In addition, the *ensemble mode* allows applying other metrics based on Protein Structure Network as the changes in interaction strength upon mutations and changes in hub behavior thanks to the usage of PyInteraph2 [46,47] for the LOCAL_INTERACTION and the STABILITY modules, respectively.

The LONG_RANGE module applies coarse-grain models to estimate allosteric free energy changes upon mutation based on Allosigma2 [48,49]. In the ensemble mode, other metrics from graph theory based on changes in shortest communication paths can also be applied using PyInteraph2 [46,47].

The FUNCTIONAL_DYNAMICS module applies metrics based on dimensionality reduction to recapitulate the effects of changes in conformations of functional sites. This module is only available in the *ensemble mode*. In addition, for selected variants, an approach based on parallel bias metadynamics will allow us to further verify the long-range effects of a mutation to the corresponding functional site, similarly to what we recently applied to p53 with other metadynamics techniques [31]. The interface for the FUNCTIONAL_DYNAMICS results in the MAVISp webapp is currently under development.

The PTM modules currently supports phosphorylations, as a fist proof of principle, and it classifies the effect of variants at phosphorylatable sites reported by PhosphoSitePlus [50]. In this step, the variant can be classified for consequences on regulation, stability, or function changes. In more detail, in a case in which the phosphorylatable site is not accessible to the solvent, the effect on regulation is classified as unknown in the *simple mode*. The *ensemble mode* is required to understand the possibility that a cryptic phosphorylated site is made accessible by conformational changes [51,52]. If the site has solvent accessibility of its side chain higher than 20%, the variant is expected to be damaging for the regulatory effect associated with the PTM, unless it is a substitution of serine to threonine or vice versa (neutral) or substitution of tyrosine to serine/threonine or vice versa (unknown). The effect on the stability is evaluated by comparing the effect that the phosphorylation would have on the wild-type variant of the protein target using folding free energy calculations with FoldX with the effect that the mutation under investigation has according to the STABILITY module. With regard to the functional impact of the substitution, we first verify if the variant is occurring in a Short Linear Motif (SLiM) that requires a phosphorylated residue. In the presence of this evidence, the variant is marked at first as 'potentially_damaging'. Furthermore, we evaluate if we can model the effect of the phosphorylation and the mutation on a structure of the complex between the protein SLiM and its interactor. This can be assessed with the LOCAL_INTERACTION module to have structure-based evidence of the damaging effect. If the lack of phosphorylation and substitution with a different residue impair the protein interaction, the annotation is changed from 'potentially_damaging' to 'damaging'. On the contrary, if after the structure-based calculations the substitution is expected to retain the properties of the phospho-site, the annotation is changed to 'neutral'. The PTM module takes advantage of the annotations provided by our recently developed Cancermuts package [53] for searching SLiM motifs and identifying the phospho-sites.

Additionally, MAVISp provides annotations for the variant interpretation reported in ClinVar [18,19], REVEL [54], DeMask [55] and GEMME[23] , the pLDDT score in the Alphafold model, relative solvent accessibility of the wild-type side chain, which can assist the interpretation of the results.

Table 1. Overview of the proteins and variants included in MAVISp

Protein	Residues	ClinVar	TOT. VARIANTS	MAVISp analysis modules
ADCK1	1-530	N.A.	28	STABILITY; PTM; LONG_RANGE
ADCK2	88-626	3	188	STABILITY; PTM; LONG_RANGE
ADCK5	47-580	N.A.	202	STABILITY; PTM; LONG_RANGE
ADH1B	1-375	N.A.	246	STABILITY; PTM; LONG_RANGE
ADH1C	1-375	N.A.	179	STABILITY; PTM; LONG_RANGE
AHR	35-591	120	304	STABILITY; PTM; LOCAL_INT.; LONG_RANGE
ALDH1A	1-501	10	212	STABILITY;PTM;LONG_RANGE
ALPK2	1781-2134	165	305	STABILITY;PTM
AMBRA1	6-204 & 856-1038	5	121	STABILITY; PTM;LOCAL_INT.;LONG_RANGE
AP2A1	1-977	7	352	STABILITY;PTM;LONG_RANGE
APOC1	1-83	1	41	STABILITY; PTM; LONG_RANGE
ARID3A	228-351	N.A.	44	STABILITY; PTM; LONG_RANGE
ATG2A	1-1938	22	783	STABILITY; PTM
ATG3	1-314	3	112	STABILITY; PTM; LOCAL_INT.; LONG_RANGE
ATG4A	25-398	5	145	STABILITY; PTM; LONG_RANGE
ATG4B	1-393	3	138	STABILITY; PTM; LONG_RANGE
ATG9A	36-725	7	220	STABILITY; PTM; LONG_RANGE
ATG12	44-140	N.A.	25	STABILITY;PTM;LONG_RANGE
ATR	1-2644	627	2309	STABILITY:PTM;LOCAL_INT
ATXN3	1-361	5	126	STABILITY; PTM; LOCAL_INT.; LONG_RANGE
B2M	1-119	20	185	STABILITY; PTM; LOCAL_INT ; LONG_RANGE
BARD1	33-160 and 422-777	1035	1190	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
BAX	12-192	1	72	STABILITY;PTM;LONG_RANGE
BCKDK	32-412	34	182	STABILITY;PTM
BCL2	1-211	1	345	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
BCL2A1	1-175	1	82	STABILITY;PTM;LONG_RANGE
BCL2L1	1-206	N.A.	138	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
BCL2L2	9-177	2	71	STABILITY;PTM;LONG_RANGE
BCL2L10	13-172	N.A.	53	STABILITY;PTM;LONG_RANGE
BCL10	10-117	10	81	STABILITY;PTM; LOCAL_INT.;LONG_RANGE
BLM	368-1290	830	1237	STABILITY;PTM;LONG_RANGE
BOK	1-212	9	59	STABILITY;PTM;LONG_RANGE
BRAF	41-722	379	968	STABILITY;PTM;LONG_RANGE
BRD4	43-173;350-459;608-678	15	242	STABILITY;PTM;LONG_RANGE
BTG1	1-131	1	234	STABILITY;PTM;LONG_RANGE
BTG2	7-130	2	257	STABILITY;PTM;LONG_RANGE
CALR	17-404	8	280	STABILITY; PTM; LONG_RANGE
CAMK1G	12-325	5	146	STABILITY;PTM
CAMKK1	105-456	N.A.	99	STABILITY;PTM
CAMKV	17-327	N.A.	149	STABILITY;PTM
CBFB	1-182	N.A.	165	STABILITY;PTM;LOCAL_INT;LONG_RANGE
CBL	47-439 & 840-906	195	564	STABILITY; PTM; LONG_RANGE
CDC34	6-193	2	78	STABILITY; PTM; LONG_RANGE
CDK4	1-303	143	246	STABILITY; PTM; LOCAL_INT ;LONG_RANGE
CDK10	10-340	16	145	STABILITY; PTM; LONG_RANGE
CDK14	131-449	1	184	STABILITY; PTM;LONG_RANGE
CDK15	94-416	4	155	STABILITY; PTM
CDK19	1-361	1	304	STABILITY; PTM
CDK20	1-304	3	140	STABILITY; PTM
CDKL1	1-302	N.A.	102	STABILITY; PTM
CDKL3	1-313	1	113	STABILITY; PTM
CDKL4	1-311	1	133	STABILITY; PTM
CFAP410	1-147	21	45	STABILITY;PTM;LONG_RANGE
CLK3	283-630	N.A.	118	STABILITY; PTM
CLK4	138-481	N.A.	135	STABILITY; PTM
COQ8A	173-647	156	302	STABILITY; PTM
COQ8B	80-525	47	171	STABILITY; PTM
CSNK2A3	1-338	1	148	STABILITY;PTM;LONG_RANGE
CSNK1A1	9-314	N.A.	136	STABILITY; PTM
CSKN1A1L	10-317	N.A.	176	STABILITY; PTM
CSNK1G1	37-337	4	117	STABILITY; PTM
CSNK1G2	39-338	N.A.	119	STABILITY; PTM
CSNK1G3	35-334	2	120	STABILITY; PTM
CSTB	1-98	N.A.	28	STABILITY;PTM;LONG_RANGE

CYB5A	1-134	4	48	STABILITY;PTM;LONG_RANGE
CYB5R3	1-301	62	105	STABILITY;PTM
CYP2R1	1-501	33	168	STABILITY;PTM;LONG_RANGE
DYRK1B	39-433	N.A.	182	STABILITY;PTM;LONG_RANGE
EGFR	1-1072	818	1955	STABILITY;PTM;LONG_RANGE
ERCC4	19-914	257	762	STABILITY;PTM;LONG_RANGE
ESRRB	100-431	30	204	STABILITY;PTM;LONG_RANGE
FANCA	37-1441	996	1696	STABILITY;PTM;LONG_RANGE
FANCE	12-534	202	356	STABILITY;PTM;LONG_RANGE
FANCF	2-369	161	320	STABILITY;PTM;LONG_RANGE
FANCG	12-616	162	360	STABILITY;PTM;LONG_RANGE
FANCI	1-1279	305	840	STABILITY;PTM;LONG_RANGE
FANCL	1-375	160	278	STABILITY;PTM;LONG_RANGE
FKB1PA	2-108	N.A.	28	STABILITY;PTM;LOCAL_INT;LONG_RANGE
GABARAP	1-116	N.A.	36	STABILITY;PTM;LONG_RANGE
GATA3	258-367	9	170	STABILITY;PTM;LONG_RANGE
GRIN2A	32-387	147	532	STABILITY;PTM;LONG_RANGE
HRAS	1-173	90	202	STABILITY;PTM;LONG_RANGE
IDH1	1-414	163	477	STABILITY;PTM
IDH2	41-452	69	355	STABILITY;PTM
KEAP1	52-624	12	934	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
KIF5A	6-380	128	266	STABILITY;PTM;LONG_RANGE
KRAS	1-176	108	332	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
LDHA	1-332	52	101	STABILITY;PTM
MAP1LC3B	1-120	3	44	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
MCL1	160-350	1	139	STABILITY;PTM;LONG_RANGE
MEFV	1-98	29	69	STABILITY;PTM
MLH1	1-341 & 501-756	1244	1373	STABILITY;PTM;LONG_RANGE
MZF1	38-126 & 355-734	3	41	STABILITY;PTM;LONG_RANGE
NBN	1-749	1289	1577	STABILITY;PTM;LONG_RANGE
NPLOC4	1-608	3	216	STABILITY;PTM;LONG_RANGE
NQO1	1-274	77	150	STABILITY;PTM;LONG_RANGE
NXF1	117-619	3	229	STABILITY;PTM;LOCAL_INT.
PARK7	1-189	40	89	STABILITY;PTM;LONG_RANGE
PELI1	11-418	6	158	STABILITY;PTM;LONG_RANGE
PELI2	13-420	N.A.	174	STABILITY;PTM;LONG_RANGE
PELI3	40-465	N.A.	157	STABILITY;PTM;LONG_RANGE
PIK3CA	1-1068	451	1489	STABILITY;PTM;LONG_RANGE
PILRA	32-153	1	43	STABILITY;PTM;LONG_RANGE
PIM1	30-305	N.A.	462	STABILITY;PTM;LONG_RANGE
PPIA	2-164	N.A.	50	STABILITY;PTM;LOCAL_INT;LONG_RANGE
PPARG	1-505	8	346	STABILITY;PTM;LONG_RANGE
PRKCA	27-672	3	281	STABILITY;PTM
PTEN	1-351	704	1172	STABILITY;PTM;LOCAL_INT;LONG_RANGE
PTK2	32-1052	4	427	STABILITY;PTM;LONG_RANGE
RAD51C	13-350	660	750	STABILITY;PTM;LONG_RANGE
RAF1	1-648	142	608	STABILITY;PTM;LONG_RANGE
RIPK2	1-523	5	209	STABILITY;PTM;LONG_RANGE
RFWD3	284-744	25	164	STABILITY;PTM;LONG_RANGE
RUNX1	49-181	156	181	STABILITY;PTM;LONG_RANGE
SMPD1	84-611	251	432	STABILITY;PTM;LONG_RANGE
SOD1	1-154	120	173	STABILITY;PTM
SRC	87-536	9	358	STABILITY;PTM;LONG_RANGE
STAT3	1-716	N.A.	650	STABILITY;PTM;LONG_RANGE
TP53	91-289	743	1138	STABILITY;PTM;LONG_RANGE
TP53BP1	1712-1972	N.A.	17	STABILITY;PTM;LOCAL_INT.;LONG_RANGE
TRAP1	94-704	44	256	STABILITY;PTM;LONG_RANGE
TTR	28-147	150	183	STABILITY;PTM
TXN	1-105	N.A.	30	STABILITY;PTM;LONG_RANGE
UBE2G2	1-165	N.A.	63	STABILITY;PTM;LONG_RANGE
ULK1	7-279 & 837-1046	1	193	STABILITY;PTM;LONG_RANGE
VCP	22-766	178	416	STABILITY;PTM
VHL	58-213	N.A.	572	STABILITY;PTM;LONG_RANGE

Sources of variants supported by MAVISp

As a default, all the variants reported for the target protein in COSMIC, cBioPortal, and ClinVar are annotated within MAVISp. We plan regular updates of the entries in the MAVISp database to add new variants or correct previously misreported ones. The variants are retrieved from ClinVar and using Cancermuts [53].

In addition, we support variant lists from other studies, such as data on cohort-based or nationwide studies or other cancer genomic initiatives. For example, we recently analyzed variants found with low

frequency in the healthy population on germline samples of childhood cancer patients [56] using one of the tools applied in the STABILITY module [41]. We here included a full assessment for some of the proteins investigated in the previous study, also including variants from COSMIC, cBioPortal or ClinVar.

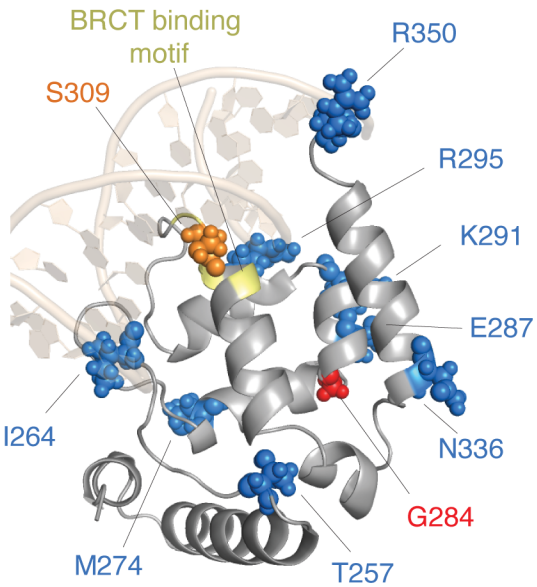
Currently, MAVISp includes data on 41880 variants and 127 proteins (**Table 1**). Some examples of the proteins and the currently available data from MAVISp are described in detail in the next sections, along with the analyses that can provided by MAVISp data.

Transcription factors

Alterations in transcription factors are often linked to aberrant gene expression, including processes such as proliferation, cell death, and other cancer hallmarks[57].

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 applied by MAVISp and MD simulations to study DNA-binding domains (ARID3A and p53[31,58]) or domains for protein-protein interaction (MZF1 and its SCAN domain[59]) from a mechanistic point of view. Here we included also other transcription factors 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 44 mutations overall (**Figure 3**) 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* (**Figure 3A**) 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 [54] 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 (**Figure 3B**).

Figure 3. MAVISp *ensemble mode* to study variants in ARID3A. We characterized 44 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 pertubation (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 a ARID3A paralog (PDB entry 1KQQ) as brown transparent cartoon.

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) [60]. 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 [61,62].

We investigated the mutational landscape in cancer of this group of proteins in a previous study [63]. In that work, we were missing a comprehensive characterization of their missense mutations at the structural level. Thus, we included, in this work, 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 and BAX. As a detailed example, we here reported the results for BCL2. 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 the possibly pathogenic variants, as predicted by REVEL, to evaluate if the structure-based assessment can help to shed light on their effects. The results are reported in **Table 2** for BCL2, as an example. Similar analyses for the other pro-apoptotic or pro-survival proteins are available by accessing the corresponding MAVISp database entries.

Table 2. MAVISp results for BCL2 variants with high pathogenic scores. We show the results of the assessment of BCL2 to illustrates the usage of MAVISp to rationalize variants with high pathogenic scores and link to possible molecular mechanisms. Here, we used REVEL first to identify the mutations with higher pathogenic score. We can also notice that DeMask agrees with a loss of function effect predicted for these mutations (i.e., negative scores). Then, we can use MAVISp results to identfy the possible mechanism behind the pathogenic effect. In this example, the predicted pathogenic variants G145E and G145R features a decrease binding with one of the BH3-containing interactor for BCL2, whereas for the variants L181P and Y180D the pathogenic potential seems linked to changes in structural stability, which could cause enhanced cellular degradation or aggregation.

Variant	REVEL	DeMask	GEMME	STABILITY	LOCAL INT. or PTM	LONG-RANGE
G145E	0.766	-0.418	-6.927	Neutral	Destabilizing	Neutral
G145R	0.764	-0.435	-7.198	Neutral	Destabilizing	Neutral
L181P	0.754	-0.417	-6.667	Destabilizing	Uncertain	Neutral
Y180D	0.722	-0.418	-6.496	Destabilizing	Uncertain	Neutral

Autophagy proteins

Autophagy is a clearance mechanism with a dual role in cancer [64–66]. The autophagy pathway relies on approximately 40 proteins that belong to the core autophagy machinery [67,68]. 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 works that served as a prototype for MAVISp [29,30]. Moreover, we studied ATG2A, ATG9A, ATG12, ATG3, ATG4A, ATG4B, AMBRA1, and GABARAP (Table 1 and OSF repository).

In the case of ULK1, we also included new structural coverage applying the AlphaFold models so that both the N-terminal (residues 7-279) and C-terminal domains (837-1046) can be used for the assessment with MAVISp. ULK1 also serves as an example on 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 captured properly by one conformation and an ensemble mode would be necessary to derive more reliable conclusions.

ULK1 featured 193 variants reported in COSMIC and/or cBioPortal. Among these, 29 had predicted long-range destabilizing effects, eight had a damaging effect on stability, and only one variant (L21P) is characterized by both effects on stability and long-range communication (**Figure 4**). 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 ULK1by compromising its 3D architecture. In addition, the K162N variant (REVEL score 0.741) has not only a detrimental distal effect but also impairs a known acetylation site which is important for ULK1 activation and, consequently, autophagy regulation[69]. Among the variants found destabilizing with the STABILITY module, V211F and L215P were also found in samples from The Cancer Genome Atlas (TCGA) as we previously reported[29].

We then analyzed MAP1LC3B regarding the effect of the mutations on the interaction with the LC3-interacting region (LIR) of its binding partner p62. 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, with results on the stability comparable to what was previously published [30]. In the assessment with the simple mode, we could not classify the P32Q variant. P32Q was predicted as destabilizing using structural ensemble from one-microsecond MD simulations in our previous study [30]. We also validated the prediction with cellular experiments that showed a decrease of protein levels for P32Q and a propensity to form cellular aggregates [30].

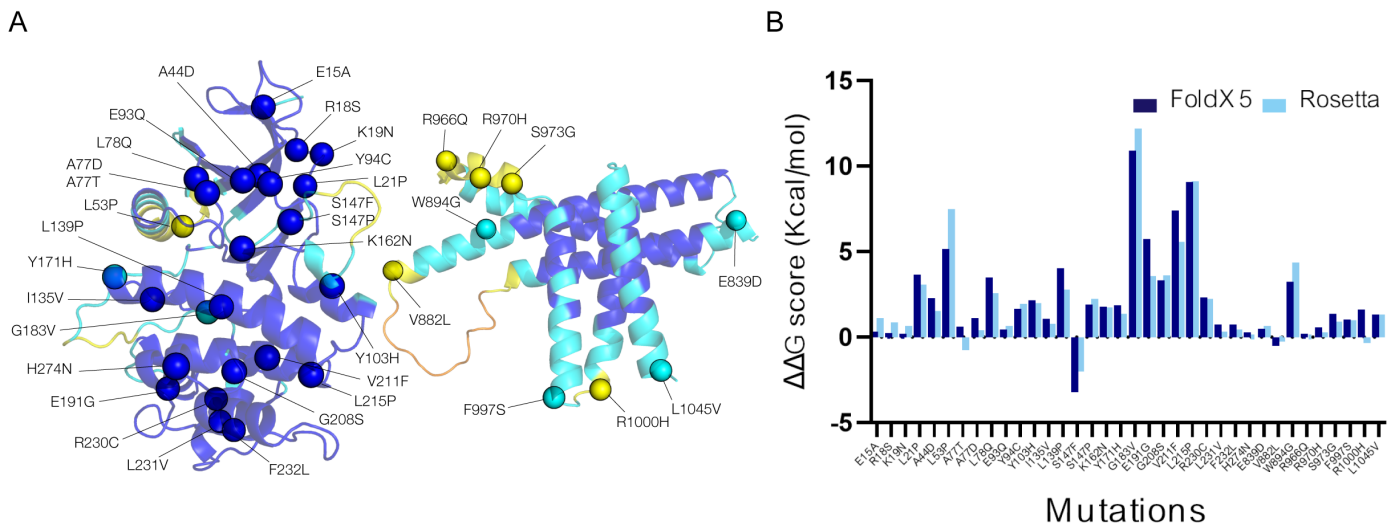


Figure 4. Destabilizing variants in the autophagy kinase ULK1. A-B) We identified 36 variants with destabilizing effects in terms of folding free energy or long-range effects.These mutations are highlighted with spheres on the AlphaFold models of the ULK1 N-terminal (Panel A, left) and C-terminal (Panel A, right) domain. B) We showed the predicted changes in folding free energy upon mutations for each of the 36 variants as calculated by the STABILITY module of MAVISp with MutateX and RosettaDDGPrediction. Interestingly, the majority of the variants that alter structural stability are located in the catalytic domain of the enzyme. This suggests potential mechanisms for ULK1 inactivation.

We did not identify variants with marked effect on the binding with p62. This suggests that the current LOCAL_INTERACTION module for the assessment is stricter than what was proposed in the first pilot study on MAP1LC3B and in better agreement with the published cellular binding assays [30]. Our assessment also confirms the possibly damaging effect of 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 (see OSF repository) and TRAP1. The tumor necrosis factor receptor-associated protein 1 (TRAP1) is a mitochondrial chaperone involved in different biological processes of relevance in cancer [70,71]. 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 [72]. The results for TRAP1 are discussed in detail.

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 only accounted for the monomeric structure of TRAP1 and we did not include the cofactors for regulating the catalytic cite for simplicity. In the future, we will investigate in more details the modification in the context of the different open and closed conformation of the TRAP1 dimer and including ADP or ATP in the modeling.

We identified 17 variants with destabilizing effects on the protein which could cause loss-of-function when the protein acts as a tumor suppressor (G160D, G126R, S219F, W231R, G250R, F280L, W299C, A319E, I339N, V391A, F440S, Y444N, S427F, Y513C, L539F, T600P, and M615R) if we use the consensus approach with FoldX and Rosetta calculations. When we replace Rosetta with RaSP S291F, V391A, S472F results in an uncertain classification. If future studies will confirm that the majority of disagreement between the two consensus approach results in an uncertain classification when RaSP is applied, one potential application to save computational time would be to use the consensus from FoldX and RaSP and apply Rosetta only on the cases that cannot be properly classified.

Moreover, 44 variants have the potential to trigger long-range effects and could deserve further study to verify which of them could relate to changes in the activity of TRAP1 or in the conformational changes required for its catalytic cycle [73]. Among these variants, 14 variants are expected to decrease the flexibility at the distal site. The remaining 30 variants have the opposite effect and is predicted to cause enhanced flexibility at the distal sites (see OSF repository).

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 [74]. In the last decades, these enzymes have become interesting targets for the design isoenzyme-specific inhibitors of their activity. It is thus also interesting to characterize the landscape of their alterations in cancer samples. We identified 50 mutations for one representative of this class of enzymes (PPIA, also known as CypA). The methods used in the STABILITY module of MAVISp have been evaluated for their accuracy against experimentally determined unfolding free energy changes in FKB1A in one of our previous studies [75].

Interestingly, we found only three destabilizing variants in terms of change of folding free energy for PPIA (S77F, G75V, and G12V) and a group of possibly allosteric mutations, of which P4L, S21F, V29I, and H92Y cause a long-range decrease of flexibility in the area where N129 is, which is in proximity of the catalytic site.

The peptidylprolyl isomerase (FKBP1A) is a small, 108-amino acid, cytoplasmatic protein. FKBP1A inactivates the cytokine receptor TGF-β 1 kinase receptor in the absence of its ligand by preventing phosphorylation and stabilizing the inactive conformation[76], thus making FKBP1A a player in a healthy immune response. Indeed, upregulation of FKBP1A has been associated with head and neck squamous cell carcinoma (HNSCC) [77]. Acting in a similar inhibitory way, FKBP1A also regulates the activin receptor complex by preventing the association of SMAD2 and SMAD3 [78]. The present analysis of FKBP1A revealed 28 mutations at 27 positions reported in COSMIC and/or cBioportal. Among these two are predicted to have long-range destabilizing effects (S39C, S39Y affecting the histidine at position 25), 10 are predicted by both MutateX and RosettaDDGPrediction to have damaging effects on stability (G20V, G29E, S39V, S39Y, W60C, G63V, G70C, G84C, P94S, F100L) (**Figure S1A**), 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-β 1 kinase receptor, using the PDB entry 1B6C (chains A and B)[79]and the mutations within 10 Å of the binding interface (16 mutations). This study revealed no stabilizing mutations of interaction which would have suggested increased binding, leading to a reduced immune response (**Figure S1B**). On the other side, we identified three variants (G29R, F47L and E61K) that are predicted destabilizing by MutateX, whereas they do not pass the defined threshold with RosettaDDGPrediction. These variants might lead to decreased binding, which could cause a more pronounced immune response. However, they would require further investigation due to the lack of consensus between the two approaches. Of note, in the case of G29E, the changes in the interaction may stem from the destabilization of the protein architecture itself (**Figure S1A**).

Using an atomic contact Protein Structure Network (acPSN) we identified the most important contacts between F47 of FKBP1A and multiple side chains on the TGF-β receptor (i.e., M184, P194, L196, V197, R199, T200, and I201, **Figure S1B**). Moreover, we identified a salt bridge between E61 and R199 of the TGF-β receptor (**Figure S1B**). These interactions are likely to be important for the molecular recognition or the stability of the complex and they would be dramatically altered by mutations as E61K and F47L.

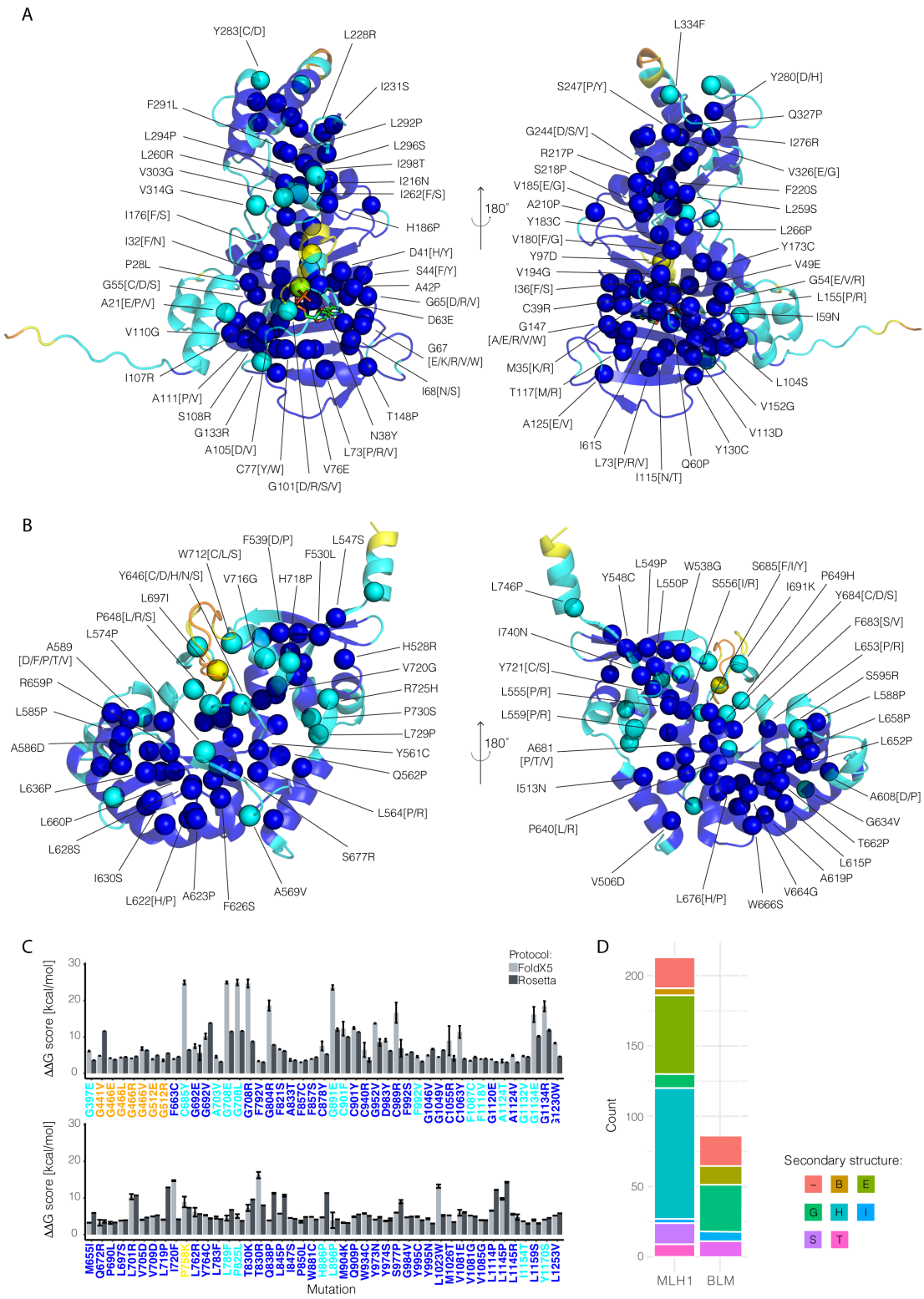


Figure 5. Variants with effects on structural stability in the tumour suppressors MLH1 and BLM. (A) Cartoon representation of $MLH1_{1-341}$ with spheres highlighting the Cα atom of the 75 positions where 120 mutations have been found annotated in the cBioPortal, COSMIC and ClinVar databases. The Mg^{2+} and ADP cofactors are shown as a green sphere and a stick, respectively. (B) Cartoon representation of $MLH1_{501-756}$ with spheres indicating the Cα atom of the 63 positions where 93 mutations have been found annotated in the cBioPortal, COSMIC and ClinVar databases.The 213 and 90 destabilizing mutations found in MLH1 and BLM, respectively, annotated in cBioPortal, COSMIC and ClinVar databases are shown in panels A and B, respectively. (C) Barplot illustrating the predicted ΔΔG values in kcal/mol of the 90 destabilizing mutations found in BLM. Both FoldX5 and Rosetta average predictionsΔΔG values in kcal/mol are reported asin light grayblue and dark graylight blue bars, respectively, with standard deviation values shown as black winged lines. Values above 25 kcal/mol have been set to 25 kcal/mol. 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 C 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 known or newly discovered driver genes in cancer, i.e., tumor suppressors and oncogenes. This application can help the identification of potential driver mutations in these genes, or to understand the role of Variants of Uncertain Significance (VUS) in cancer predisposition. We here applied the *simple mode* of MAVISp to examples of tumor suppressors (ATR, B2M, BARD1, BAX, BCL10, BLM, BTG1, BTG2, CBFB, ERCC4, FANC proteins, GRIN2A, KEAP1, MLH1, NBN, PPARG, PTEN, RFWD3 and VHL) and oncogenes (BRAF, BRD4, IDH1, IDH2, CDK4, CALR, EGFR, HRAS, KRAS, PIK3CA, PIM1, RAF1,SRC, and STAT3) reported in COSMIC. 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)[80]. We also included genes of interest previously identified by the analyses of germline variants in childhood cancer (for example RAD51C)[41]. Additionally, we investigated the BRCT domain of TP53BP1 since it is central for the regulation of different proteins involved in DNA repair and also 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 90 and 213 destabilizing variants for BLM and MLH1, respectively (**Figure 5**), of which more than half are reported in ClinVar and, thus, of medical relevance. 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 (**Figure 5**).

Moreover, in the study of PTEN with MAVISp, we assessed 1162 variants and identified 283 of them as destabilizing. Of these, 236 were found as detrimental for structural stability with changes in stability ranging from 3 to 18.5 kcal/mol. 57 variants featured destabilizing effects triggered long-range to distal sites. Of those, only ten variants (Y16S, H93P, W111C, W11G, W11S, W111R, R142P, Y188S, Y188D, and R335P) have combined effects causing both changes to the folding free energy and allosteric perturbations. We compared our results to what reported in other studies on PTEN variants, and we identified that at least 176 out of the 236 destabilizing variants have been observed to alter PTEN abundance levels in cells [81_83] . The results from MAVISp also fits nicely with recent computational studies of PTEN variants[25]. 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 [84]. The existing X-ray structure of this interaction (PDB entry 401V [85]) only models a 15 amino acids long PTEN peptide, which corresponds to 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 [86]. 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 corresponding to the PDB entry 4O1V. This allowed us to study nine variants of PTEN (A359S, A359V, S360C, S360G, S360T, S361N, S361R, T363A, T363I), and none was determined to be destabilizing for interaction with SPOP. Interestingly, two variants (A359S, A359V) were annotated as uncertain, because FoldX predicts them to be stabilizing and destabilizing respectively, while RossettaDDGPrediction suggests no impact on binding free energies. Bearing in mind the limitations in conformational sampling of these approaches, the next step in the analysis 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 an advantage to the cancer evolution.

Interestingly, 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 (**Figure 6**). For example, V103A induces stabilization in the residues of the Switch I region, which forms part of the binding interface for effector proteins.

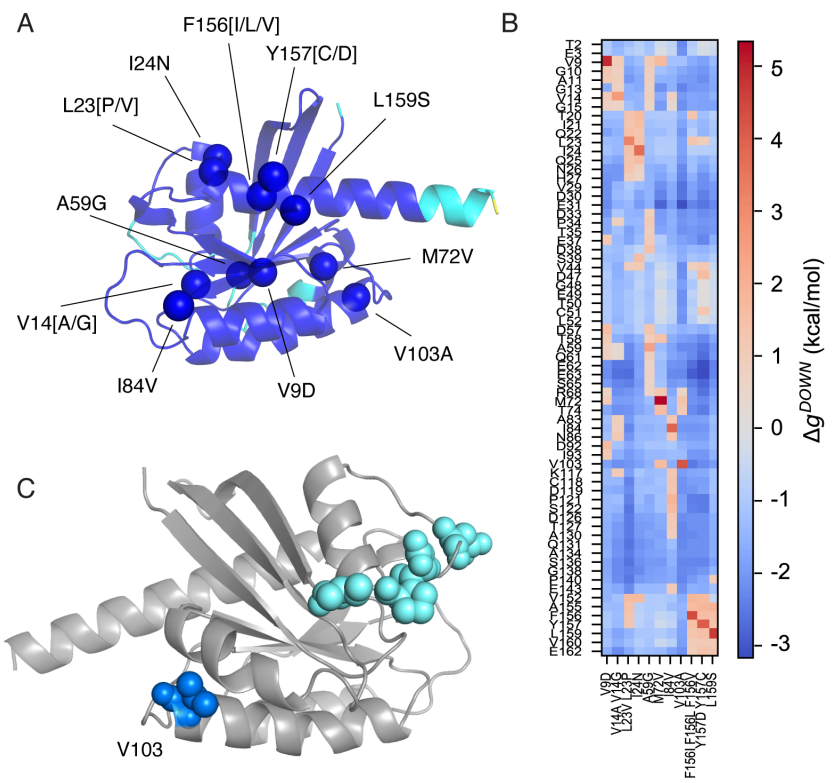


Figure 6. KRAS and allosteric mutations. A) KRAS AlphaFold model with each residue colored according to the pLDDT score (i.e., blue - very high (pLDDT > 90), cyan - confident (70 < pLDDT <90), yellow - low (50 < pLDDT < 70), and orange - very low (pLDDT <50)). The Cα of the residues involved in stabilizing allosteric effects are shown as spheres. **B)** Heatmaps of the DOWN mutations obtained with the Allosigma2 tool. 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 8Å 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, 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 prediction on the transmembrane regions of proteins with the *simple mode*. 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 [87]. Once a transmembrane protein is identified, we will retain only the variants in the domain that are external to the membrane. As an example, we studied PILRA, which belongs to the immunoglobulin superfamily and play a role in immune response [88]. 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 point 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[89], along with KIF5A[90], CFAP410[91], and CYP2R1[92]. We used the stability module of the framework to classify them as destabilizing, neutral, uncertain, or stabilizing by predicting the ΔΔG of folding stability upon mutation. Interestingly, we found 55 variants 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 in the dataset, such as TTR, SOD1, SMPD1.

Application of MAVISp for VUS characterization

We then aimed at exploring the potential of MAVISp not only in assessing variants with different functional layers, but also to classify variants. We used all the variants found in ClinVar for the proteins reported in **Table 1** for which we have data from the four main modules of the *simple mode* assessment (i.e., AHR, ATG3, ATXN3, B2M, BARD1, BCL10, BCL2, CDK4, KRAS, MAP1LC3B and PTEN) and divided them into pathogenic, benign, likely pathogenic, or VUS/conflicted evidence. It should be kept in mind

that, in the current release, not all the proteins have been necessarily analyzed for all their possible biological interactors or PTMs. Thus, if a pathogenic variant is predicted as neutral, it might mean that additional layers of evaluation are needed to achieve a full picture. This is especially true for oncogenic proteins such as KRAS, where cancer driver alterations are expected to relate to gain-of-function effects and will require the application of more MAVISp layers related to functional changes to be classified correctly. On the contrary, we would expect that most (or virtually all) the variants currently classified as benign in ClinVar to be predicted as having neutral effects, if our framework is effective.
The results are reported in **Figure 7A-B** using the example of the classification based on the STABILITY and LONG_RANGE modules of MAVISp. Interestingly, all the variants predicted as damaging with these MAVISp modules are pathogenic according to ClinVar. On the contrary, using a cutoff for pathogenicity of REVEL of 0.5 pathogenic variants in ClinVar are still captured with neutral effects (**Figure 7C**). More in-depth investigation and a much larger dataset of predictions performed with MAVISp are needed to further verify these results.
We then evaluated if MAVISp could predict any VUS or variants of conflicting evidence with a damaging effect among the 2833 investigated to shed new light on uncharacterized variants. We identified 311 potentially destabilizing variants for structural stability, 29 variants with damaging effects related to changes in phosphorylation sites, 12 variants with alterations at the binding interface for their interactors, and 776 variants with potential damaging effects through long-range perturbations.

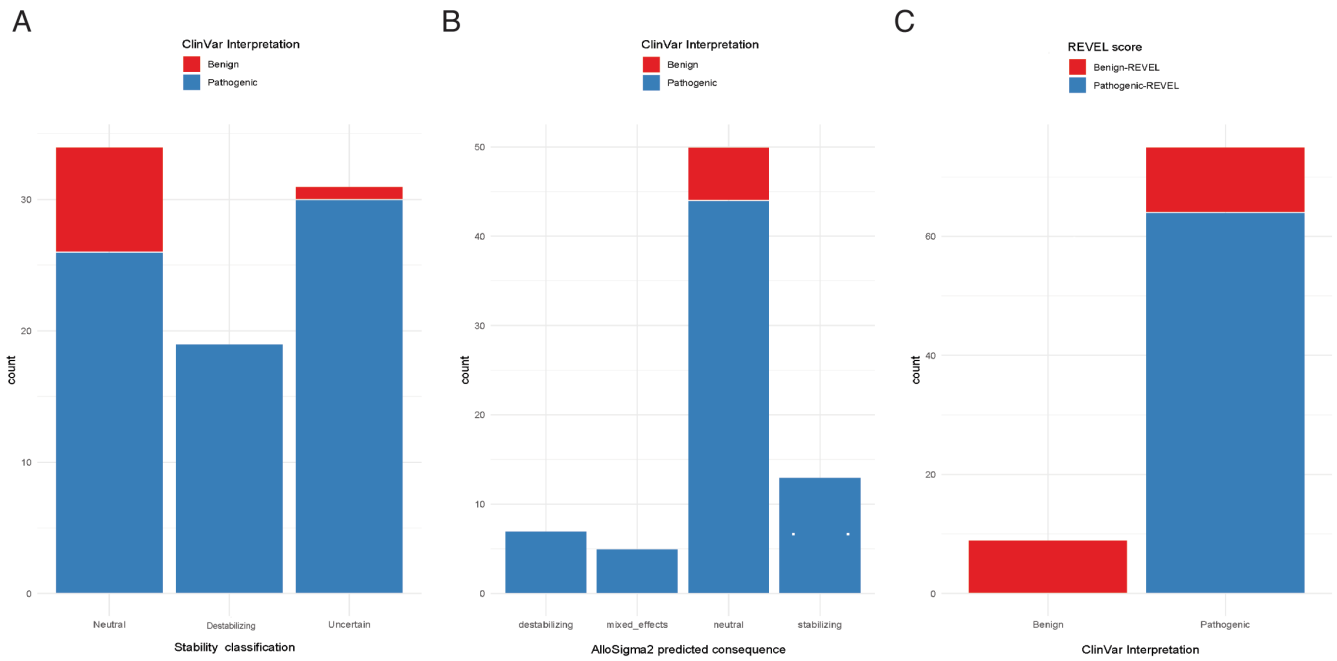


Figure 7. MAVISp prediction in comparison with ClinVar classification. A) The barplot shows the ClinVar classification for variants that are predicted as destabilizing, neutral or uncertain by MAVISp *simple mode* including only the results from the STABILITY module. **B)** The barplot shows the ClinVar classification for variants that are predicted as stabilizing, destabilizing, with mixed effects (i.e., stabilizing to certain sites and destabilizing to others) or neutral, according to the LONG_RANGE MAVISp module in the *simple mode*. C) The barplot shows the ClinVar classification for variants that are predicted as pathogenic or benign according to REVEL.

Conclusions and Future Perspective

MAVISp provides a multi-layered assessment of the effects of variants found in cancer studies using structural methods. The results can be seen as a complementary resource to available pathogenic scores. 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. MAVISp will also help to understand the effects of VUS or variants with conflict evidence at the molecular level.

The assessment provided by MAVISp could have the potential to be used for the classification of variants by itself. In this context, further studies are needed to benchmark it against variants with known effects to estimate the accuracy, sensitivity, and specificity of the different layers and their collective contributions. For example, in the case of tumor suppressor genes and loss-of-function variants, the STABILITY module might be sufficient to provide an accurate classification.

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[93]. In a part of this work, 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 and characterized also for other layers. 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 and will be adapted to the framework and included in one of the next releases.

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.

We here compared our results to scores derived from REVEL, DeMask[55], and GEMME[23] for illustrative purposes. Next, we will expand the portfolio of scores that are imported in MAVISp and can be used for comparison. For example, we aim to include EVE (Evolutionary model of variant effect)[94]. Furthermore, we will integrate experimentally derived scores and their interpretation by mining the MAVE database[95], in which score-sets of multiplex assays of variant effects (MAVE), such as deep mutational scans, are deposited.

Furthermore, approaches as the one provided by Tiemann et al.[96] could broaden the possibility of expanding the STABILITY module to transmembrane proteins. At this stage, it is possible to provide annotations for variants in regions of transmembrane proteins that are external to the membrane, as exemplified by the example of PILRA, EGFR and ATG9A in this study. 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, or 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 etc.

Methods

Initial structures for MAVISp

We used models from the AlphaFold2 database[27] 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-model the relevant cofactors upon analyses with AlphaFill[97] and where needed through a step with Modeller[98]. We used an X-ray (PDB entry 3K0N[99]) and an NMR structure (PDB entry 2KK0[100]) for PPIA and ARID3A, respectively. For AMBRA1 we used the CryoEM structure recently solved in complex with DDB1[101], isolating the chain for AMBRA1 and reconstructing missing coordinates with Modeller. The structural coverage for each protein is reported in **Table 1.** 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)[102,103].

Free energy calculations

We applied the BuildModel module of FoldX5 suite [37] averaging over five independent runs for the calculations of changes in free energy of folding upon mutation with MutateX and the foldX5 energy function. 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 [39] .

We used MutateX to calculate changes in binding free energy for the LOCAL_INTERACTION module using the Build Model and Analyze Complex 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 [39]. We also applied RaSP using the same protocol provided in the original publication[42] 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[48]. 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 [31]. We retained only variants that were exposed at the solvent (>20% relative solvent accessibility of the side chain) and 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. The distance filter has been applied using the contacts estimated by PyInteraph2 [46]and based on side-chain centers of mass, retaining only the AlloSigMA2 pairs of residues not included in the PyInteraph2 contact outputs[46].

Molecular dynamics simulations

For ARID3A, MAP1LC3B, PPIA, the DNA binding domain of p53, and $MZF1_{37-128,}$ we used previously published [30,31,58,59,104,105] all-atom molecular dynamics simulations of one microsecond with the CHARMM22star force field [106].

MAVISp classification

We aggregated the results for the MAVISp annotation related to changes in stability, which includes the results from the STABILITY module and one of the features associated with the PTM module. In a similar manner, all the features that capture effects on protein functions have been also aggregated together. This includes the results from the LOCAL_INTERACTION, LONG_RANGE modules, as well as one of the features from the PTM module. We used an approach similar to the one applied in the previous pilot projects on ULK1 and MAP1LC3B[29,30]. In particular, we assigned score of one to variants classified as: i) destabilizing or stabilizing by the STABILITY and LOCAL_INTERACTION modules, ii) damaging or potentially damaging for changes in regulation, stability, or function of post-translational modifications by the PTM module, iii) destabilizing/stabilizing/mixed_effects on distal communication sites by the LONG_RANGE module. We assigned a score of 0 to variants classified as neutral by the STABILITY, LOCAL_INTERACTION, PTM, and LONG_RANGE modules. We considered missing values variants annotated as uncertain/unknown for all the properties analyzed. The cumulative effects for stability or functions have been divided as mentioned above and then also a total contribution has been calculated. The results are then ranked for each category to identify the most damaging mutations and their corresponding altered properties to streamline the interpretation of the results.

Data Availability

The data can either be consulted through the web application provided in GitHub: https://github.com/ELELAB/MAVISp or as individual CSV files in the OSF repository: https://osf.io/ufpzm/

References

1. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science (1979)* **339**, 1546–1558 (2013).
2. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
3. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: A looking glass for cancer? *Nat Rev Cancer* **12**, 323–334 (2012).
4. 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).
5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
6. Akdel, M. *et al.* A structural biology community assessment of AlphaFold 2 applications. *biorxiv* doi:10.1101/2021.09.26.461876.
7. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nature Protocols 2016 11:10* **11**, 1782–1787 (2016).
8. 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).
9. Ollodart, A. R. *et al.* Multiplexing mutation rate assessment: determining pathogenicity of Msh2 variants in Saccharomyces cerevisiae. *Genetics* **218**, (2021).
10. Kris, M. G. *et al.* Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *JAMA* **311**, 1998–2006 (2014).
11. 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).
12. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
13. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).
14. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
15. 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.
16. 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).
17. Plazzer, J. P. *et al.* The InSiGHT database: Utilizing 100 years of insights into Lynch Syndrome. *Fam Cancer* **12**, 175–180 (2013).
18. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, (2014).
19. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835–D844 (2020).
20. 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.
21. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends Biochem Sci* **44**, 575–588 (2019).
22. 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).

23. 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).
24. 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).
25. 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).
26. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv* (2022) doi:10.1101/2021.10.04.463034.
27. 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).
28. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
29. 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).
30. Fas, B. A. *et al.* The conformational and mutational landscape of the ubiquitin-like marker for autophagosome formation in cancer. *Autophagy* 1–24 (2020).
31. 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).
32. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
33. Calderone, A., Castagnoli, L. & Cesareni, G. Mentha: A resource for browsing integrated protein-interaction networks. *Nat Methods* **10**, 690–691 (2013).
34. Jamroz, M., Kolinski, A. & Kmiecik, S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics* **30**, 2150–2154 (2014).
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. 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.
38. 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).
39. Frenz, B. *et al.* Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front Bioeng Biotechnol* **8**, (2020).
40. Tiberti, M. *et al.* MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* **23**, bbac074 (2022).
41. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. *Protein Science* **32**, e4527 (2023).
42. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *bioRxiv* (2022) doi:10.1101/2022.07.14.500157.
43. Nielsen, S. v. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* **13**, e1006739 (2017).
44. 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.
45. 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).
46. 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.
47. Tiberti, M. *et al.* PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* **54**, 1537–1551 (2014).

48. 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).
49. Guarnera, E., Tan, Z. W., Zheng, Z. & Berezovsky, I. N. AlloSigMA: Allosteric signaling and mutation analysis server. *Bioinformatics* **33**, 3996–3998 (2017).
50. Hornbeck, P. v. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512–D520 (2015).
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. 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).
54. 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).
55. Munro, D. & Singh, M. DeMaSk: A deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36**, 5322–5329 (2020).
56. 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).
57. Bushweller, J. H. Targeting transcription factors in cancer — from undruggable to reality. *Nature Reviews Cancer* vol. 19 611–624 Preprint at https://doi.org/10.1038/s41568-019-0196-7 (2019).
58. 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).
59. Nygaard, M. *et al.* The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci* **3**, 1–18 (2016).
60. 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).
61. Birkinshaw, R. W. & Czabotar, P. E. The BCL-2 family of proteins and mitochondrial outer membrane permeabilisation. *Semin Cell Dev Biol* (2017) doi:10.1016/j.semcdb.2017.04.001.
62. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. *Nat Rev Mol Cell Biol* **15**, 49–63 (2014).
63. 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).
64. Singh, S. S. *et al.* Dual role of autophagy in hallmarks of cancer. *Oncogene* **37**, 1142–1158 (2018).
65. White, E. & DiPaola, R. S. The double-edged sword of autophagy modulation in cancer. *Clinical Cancer Research* **15**, 5308–5316 (2009).
66. Thorburn, A. Autophagy and Its Effects: Making Sense of Double-Edged Swords. *PLoS Biol* **12**, 10–13 (2014).
67. Yang, Z. & Klionsky, D. J. Mammalian autophagy: Core molecular machinery and signaling regulation. *Current Opinion in Cell Biology* vol. 22 124–131 Preprint at https://doi.org/10.1016/j.ceb.2009.11.014 (2010).
68. Suzuki, H., Osawa, T., Fujioka, Y. & Noda, N. N. Structural biology of the core autophagy machinery. *Curr Opin Struct Biol* **43**, 10–17 (2017).
69. Shu-Yong, L. *et al.* GSK3-TIP60-ULK1 signaling pathway links growth factor deprivation to autophagy. *Science (1979)* **336**, 477–481 (2012).
70. Rasola, A., Neckers, L. & Picard, D. Mitochondrial oxidative phosphorylation TRAP(1)ped in tumor cells. *Trends Cell Biol* **24**, 455–463 (2014).
71. 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).
72. 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).

73. 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.
74. 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).
75. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. doi:10.1101/2022.09.02.506350.
76. Chen, Y.-G., Liu, F. & Massagué, J. Mechanism of TGFβ receptor inhibition by FKBP12. *EMBO J* **16**, 3866–3876 (1997).
77. Patel, D. *et al.* FKBP1A upregulation correlates with poor prognosis and increased metastatic potential of HNSCC. *Cell Biol Int* **46**, 443–453 (2022).
78. Yamaguchi, T., Kurisaki, A., Yamakawa, N., Minakuchi, K. & Sugino, H. FKBP12 functions as an adaptor of the Smad7–Smurf1 complex on activin type I receptor. *J Mol Endocrinol* **36**, 569–579 (2006).
79. 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).
80. 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).
81. 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).
82. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874–882 (2018).
83. 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).
84. Guo, Z. Q. *et al.* Small-Molecule Targeting of E3 Ligase Adaptor SPOP in Kidney Cancer. *Cancer Cell* **30**, 474–484 (2016).
85. Li, G. *et al.* SPOP promotes tumorigenesis by acting as a key regulatory hub in kidney cancer. *Cancer Cell* **25**, 455–468 (2014).
86. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun* **13**, (2022).
87. 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).
88. Lu, Q. *et al.* PILRα and PILRβ have a siglec fold and provide the basis of binding to sialic acid. *Proc Natl Acad Sci U S A* **111**, 8221–8226 (2014).
89. 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).
90. de Boer, E. M. J. *et al.* Genotype-phenotype correlations of KIF5A stalk domain variants. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* vol. 22 561–570 Preprint at https://doi.org/10.1080/21678421.2021.1907412 (2021).
91. Shin, D. H. *et al.* Identification of the CFAP410 pathogenic variants in a Korean patient with autosomal recessive retinitis pigmentosa and skeletal anomalies. *Korean Journal of Ophthalmology* vol. 34 500–502 Preprint at https://doi.org/10.3341/KJO.2020.0087 (2020).
92. Thacher, T. D. & Levine, M. A. CYP2R1 mutations causing vitamin D-deficiency rickets. *Journal of Steroid Biochemistry and Molecular Biology* vol. 173 333–336 Preprint at https://doi.org/10.1016/j.jsbmb.2016.07.014 (2017).
93. 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.
94. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

95. Esposito, D. *et al.* MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**, (2019).
96. 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.
97. 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).
98. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).
99. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).
100. 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).
101. 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.
102. 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).
103. 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).
104. 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).
105. 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).
106. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100**, (2011).

Author Contributions (CRediT Classification)

Conceptualization: EP; *Data Curation*: MA, LB, KD, MU, SS, TD, FM, PSI, KK, KM, JBB, ADN, JV, ML, EP. *Formal Analysis*: MA, LB, MU, TD, FM, ML, EP. *Funding Acquistion:* EP, MA, SI. *Investigation*: MA, LB, KD, JS, TD, FM, MU, SI, EP. *Methodology*: MA, LB, KD, AP, 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, ML, MT, EP. *Visualization*: MA, LB, KD, MU, TD, ML, MT. *Writing – Original Draft:* EP. *Writing – Review and Editing*: MA, LB, KD, MU, TD, KM, FM, ML, MT, EP.

Acknowledgements

Our research has been supported by Carlsberg Foundation Distinguished Fellowship (CF18-0314), Danmarks Grundforskningsfond (DNRF125), Hartmanns Fond (R241-A33877), LEO Foundation (LF17006) and 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. 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.