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

Matteo Arnaudi[1†], Ludovica Beltrame[1†], Kristine Degn[1†], Mattia Utichi[1,2], Alberto Pettenella[2], Simone Scrima[2], Peter Wad Sackett[1], 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

[†] 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 the response to treatment and risks of relapse in different patients. The role of genomics variants on cancer predisposition, progression, and response to treatment continues to be realized. Thanks to advances in sequencing techniques and their introduction in a clinical setting, 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. Applications of bioinformatic-based approaches to characterize the effects of these variants demonstrated their full potential thanks to advances in machine learning, comparisons between predicted effects and cellular readouts, and advances 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 the coding region of genes 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 ten different proteins and more than 4000 variants. New protein targets are routinely analyzed in batches 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. Finally, we provide guidelines for new contributors who are interested in contributing 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].

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, 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 the first ten protein targets and 4015 variants for illustrative purposes of how to work within MAVISp.

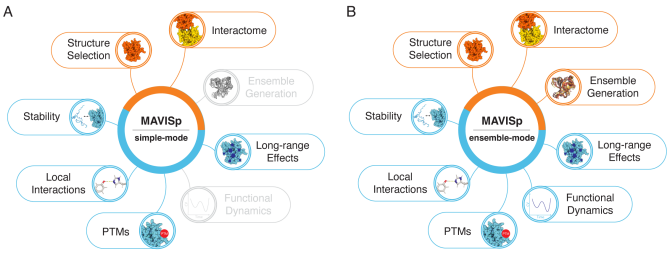

Figure 1. Overview of MAVISp components. MAVISp inlcludes different modules, each handled by workflow engines or dedicated tools. The modules in orange are the ones related to the selection/collection of the protein structures. The modules highlighted in light blue are devoted to the analyses. The procedure 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 curators can work concomitantly and independently on different proteins. The final product of each MAVISp module is one or more text files, according to specific file formats and conventions. These constitute the MAVISp database. The MAVISp web app reads such files, further processes their information, and displays it as a user-accessible website, from which the final data tables can be downloaded as simple comma-separated text files. In the future, we plan to decouple data processing and presentation so that the web app (or web server, see below) will only take care of processing the available datasets and making them accessible.

As mentioned above, we will release the results of the analyses in a Streamlit-based web server (https://github.com/ELELAB/MAVISp, **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. Both software and data are currently available on our GitHub repository.

We will add bewly 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. 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 we will provide the needed resources if needed.

Currently, the MAVISp database supports only the results from the *simple mode*, and the interface for the *ensemble mode* is under development. Therefore, the corresponding GitHub repository will contain the data for both assessment modes. In addition, we only provide the data on a single selected complex with interactors in the database, but the full-data set can be downloaded from the repository.

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 flexibility 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).

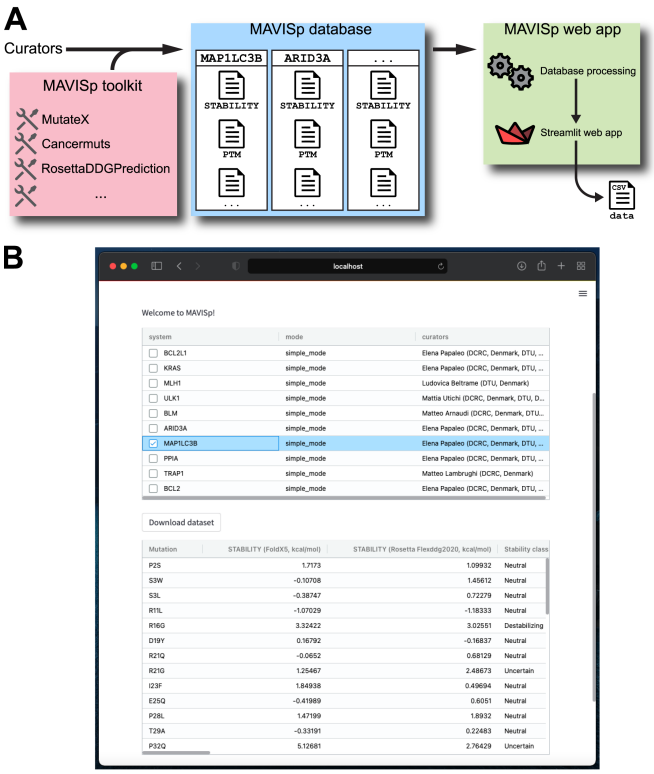

Figure 2. The MAVISp web app. (A) Scheme of the current workflow for the MAVISp database and app. Curators use specific workflows and protocols within each MAVISp module to derive structure-based prediction of the changes connected to all the variants identified for a protein target. The results are gathered into a text-file based database. The database is then imported in the 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.

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 (with CABS-flex [34,35]) or all-atom 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 module. In this step, the effect of the variants on the structural stability of the proteins is investigated using folding free energy calculations with FoldX5 suite and the foldx5 energy function[37], the cartddg2020 protocol, and the ref2015 energy function of Rosettaddg [38,39]. These steps take advantage of the recently developed workflows for high-throughput in-silico mutagenesis scans, i.e., MutateX [40] and RosettaDDGPrediction [41]. MAVISp applies a consensus approach between the two methods. Variants with changes in stability below 3 kcal/mol are not featuring marked decreased stability at the cellular level [42,43]. Thus, MAVISp defines the following classes for changes in stability: stabilizing (< - 3kcal/mol with both methods), destabilizing (> 3 kcal/mol), neutral (-2 l<x<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.

The LOCAL_INTERACTON 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 8 Å from the interaction interface using the foldx5 energy function and the flexddg protocol and talaris2014 energy function of Rosetta [44]. In this step, we also apply Mutatex and RosettaDDGPrediction.

When the ensemble mode is available, the calculations for the STABILITY and LOCAL_INTERACTION modules can be applied to a collection of 50 frames from the simulations (FoldX-based calculations) or on the 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 [45,46] 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 [47,48]. In the ensemble mode, other metrics from graph theory based on changes in shortest communication paths can also be applied using PyInteraph2 [45,46].

The FUNCTIONAL_DYNAMICS module applied metrics based on dimensionality reduction, such as Principal Component Analysis (PCA), to recapitulate the effects of changes in conformations of functional sites. This module is available only in the ensemble mode. In addition, for selected variants, an approach based on parallel bias metadynamics allows us to further verify the long-range effects of a mutation to the corresponding functional site, as recently applied to p53 [31].

Currently, the PTM modules only supports phosphorylations and it classifies the effect of variants at phosphorylatable sites reported by PhosphoSitePlus [49]. 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 [50,51]. 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'.

Table 1. Overview of the proteins and variants included in MAVISp
We indicates the protein using HUGO names for the corresponding genes and the regions with strucutral coverage used for MAVISp. In addition, the number of variants from ClinVar or retrived by Cancermuts (in COSMIC and cBioPortal) are indicated. We also annotated the MAVISp modules for which the assessment of each protein is currently available in the database.

Protein	Residues	ClinVar	COSMIC/ cBioPortal	MAVISp analysis modules
ARID3A	228-351	N.A.	44	STABILITY PTM LONG_RANGE
BCL2	1-211	N.A.	345	STABILITY PTM LOCAL_INT. LONG_RANGE
BCL2L1	1-206	N.A.	137	STABILITY LONG_RANGE
BLM	368-1290	830	597	STABILITY LONG_RANGE
KRAS	1-176	108		STABILITY LONG_RANGE
MAP1LC3B	1-120	1	**44**	STABILITY PTM LOCAL_INT. LONG_RANGE
MLH1	1-341 & 5011-756	1182	**452**	STABILITY LONG_RANGE
PPIA	2-164	N.A.	50	STABILITY PTM LONG_RANGE
TRAP1	94-704	22	**246**	STABILITY PTM LONG_RANGE
ULK1	7-279 & 837-1046	N.A.	**193**	STABILITY PTM LONG_RANGE

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 [52] for searching SLiM motifs and identifying the phospho-sites.

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 will plan regular updates of the entries in the MAVISp database to add new variants or correct previously misreported ones. The variants are retrieved using ClinVarMiner [53] and Cancermuts [52].
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 [54] using one of the tools applied in the STABILITY module [41]. The assessment for MLH1 and BLM (see below) includes some of these variants, which have been reported in COSMIC, cBioPortal or ClinVar.

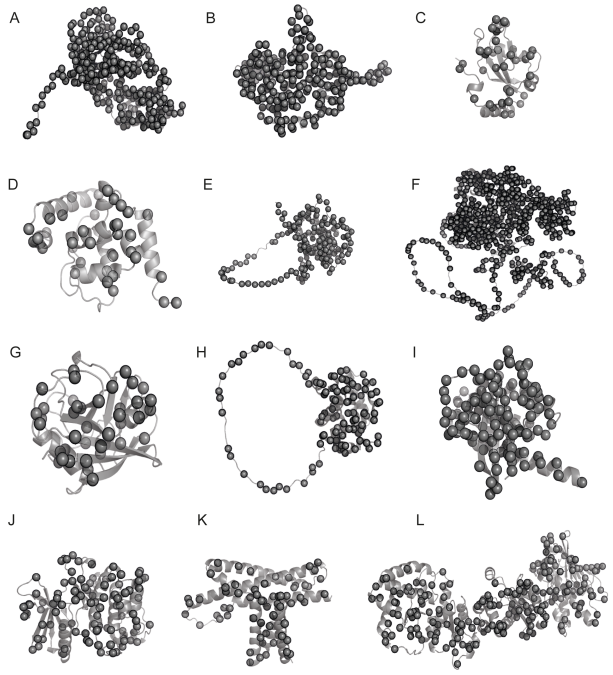

Figure 3. Overview on the proteins and variants currently included in MAVISp. All the protein structures used in the MAVISp assessment are shown as cartoon and the variants that we could classify without uncertainty with at least one of the MAVISp modules are shown as spheres. **A-B)** For MLH1 we used the AlphaFold model trimming the N-terminal (residues 1-341, A) and C-terminal (residues, 501-756, B) domains. We classified 788 variants on 317 positions for the N-terminal domain and 549 variants at 238 sites for the C-terminal domain. **C)** For MAP1LC3B, we used the AlphaFold model covering the residues 1-120, assessing without uncertainty 44 variants at 37 sites. **D)** For ARID3A, we used the NMR structure (PDB entry 2KK0, first conformer of the ensemble) for the DNA-binding domain (residues 228-351). We classified 41 variants at 36 sites. **E)** We used a model from AlphaFold for BCL2 removing the C-terminal helix that should be inserted in the membrane. Overall, we covered variants in the region 1-211 with 345 variants that could be classified at 168 different sites. **F)** For BLM we covered the region 368-1290 with the AlphaFold model and analyzed without uncertainty 1203 variants at 688 positions. **G)** We used the X-ray structure (PDB entry 3K0N, chain A) for PPIA with 49 classified variants at 41 sites. **H)** For BCL2L1, we used the Alphaold model trimmed to remove the transmembrane helix (residues 1-206) and classified 137 variants at 109 sites. **I)** For KRAS, the AlphaFold model used covered the residues 1-176 and evaluted without uncertainty 324 variants at 135 positions. **J-K)** For ULK1, two domain structures were trimmed from the full-length AlphaFold model (i.e., residues 7-279 in panel J and residues 837-1046 in panel K) to remove the connecting mostly disordered and low accuracy regions. We classified 117 variants at 98 positions and 75 variants at 63 positions, for the N-terminal and C-terminal domains, respectively. L) TRAP1 AlphaFold model covered residues 94-704 and included 250 variants without uncertainty at 211 position

Currently, MAVISp includes data on 4015 variants and ten proteins (**Table 1 and Table S1, Figure 3**). In particular, we evaluated, without uncertainty, 3832 variants. **Figure 3** illustrates the starting structures used in each assessment, i.e., X-ray (PPIA), NMR (ARID3A), and AlphaFold2 models (the remaining proteins). We only mapped the variants that we could classify with at least one of the MAVISp layers for a visual overview of the data.

The proteins and the currently available data from MAVISp are described in detail below.

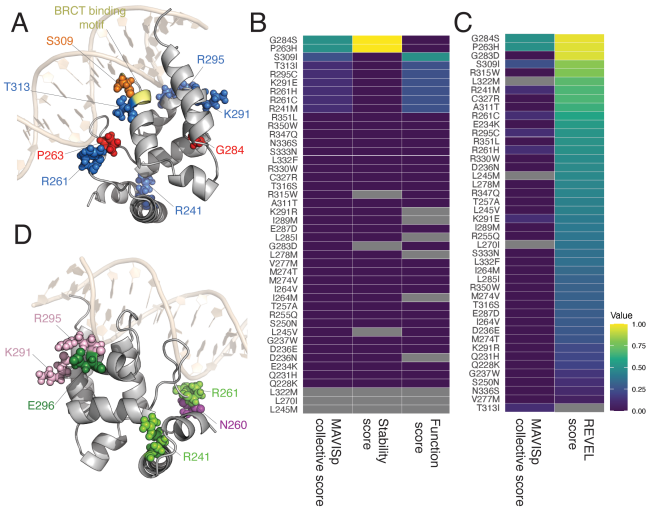

Figure 4. MAVISp *ensemble mode* to study variants in ARID3A. We characterized 44 variants in the ARID3A DNA-binding domain (residues 228-351). A) We highlighted the sites where MAVISp identified variants with effects on: i) structural stability (red spheres), ii) PTMs (orange spheres), iii) long-range 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. **B)** The heatmap shows the results of the ranking of the variants I the ARID3A DNA-binding domain based on a collective score from MAVISp (first column), along with the Stability (second column) and Function (third column) scores. Highly damaging variants are highlighted in yellow, while dark purple indicates the neutral effects. Grey tiles indicate unknown/uncertain classification. **C)** The heatmap shows the ranking of the variants comparing the MAVISp score with REVEL score . The rREVEL score ranges from 0 to 1 (dark purple – yellow) with higher scores indicating a higher likelihood of a variant of being pathogenic. MAVIp is in agreement with REVEL in pinpointing variants with marked damaging effects, allowing to link the predicted pathogenicity with possible related mechanisms. **D)** MAVISp identifies variants with long-range destabilizing effects in ARID3A. In particular, R241M and R216H (light green spheres) can perturb long range E296 (dark green spheres). Moreover, K291E and R295C can trasmit effects to the distal site N260 (purple spheres)

Transcription factors: ARID3A

Alterations in transcription factors are often linked to aberrant gene expression, including processes such as proliferation, cell death, and other cancer hallmarks[55].

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 [31,56]) or domains for protein-protein interaction (MZF1 and its SCAN domain, [57]) from a mechanistic point of view. Here we applied the *simple* and the *ensemble mode* of MAVISp to ARID3A for a comprehensive understanding of mutations found in cancer samples.

We characterized 44 mutations overall (**Table 1, Table S1, Figure 4**) 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 4A**) 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 one damaging effect on stability, long-range interactions, or post-translational regulation are also characterized by REVEL scores [58] higher than 0.5, suggesting that the assessment here provided can help to link the pathogenic effect with specific mechanisms (**Figure 4B-C**).

We also identified mutations that might have allosteric (i.e., distal) effects as R241M, R261C/H, K291E, R295C, and T313I (**Figure 4D**).

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. In this example, the predicted pathogenic variants F112C and G14E/R have altered allosteric free energy changes, whereas the variants G154D, G193E, L119P and L181P pathogenic potential seems linked to changes in structural stability, which could cause enhanced cellular degradation or aggregation.

Variant	REVEL	STABILITY	LOCAL INT. or PTM	LONG-RANGE
F112C	0.705	Uncertain	Destabilize binding to BECN1	Neutral
G14E/R	0.764	Neutral	Destab. binding to BECN1	Neutral
G154D	0.664	Destabilizing	Neutral	Neutral
G193E	0.638	Destabilizing	Neutral	Neutral
L119P	0.659	Destabilizing	Neutral	Neutral
L181P	0.754	Destabilizing	Neutral	Neutral

Pro-survival BCL2 proteins: BCL2, and BCL2L1

One of the cancer hallmarks is the ability to escape cell death (i.e., apoptosis) [59]. 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 [60,61]. We investigated the mutational landscape in cancer of this group of proteins in a previous study [62]. In that study, we were missing a comprehensive characterization of their missense mutations at the structural level. Thus, we applied MAVISp with the *simple mode* to two central pro-survival proteins BCL2, and BCL2L1 (also known as BCLxL). 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.

Mutations with long-range effects were not identified for these two proteins. Most of the variants are characterized by low pathogenicity scores (quantified according to REVEL) and neutral effects predicted by the MAVISp modules. We focused our attention on the possibly pathogenic variants 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 BCL2L1 are available by accessing the corresponding MAVISp database entry.

Autophagy proteins: MAP1LC3B and ULK1

Autophagy is a clearance mechanism with a dual role in cancer [63–65]. The autophagy pathway relies on approximately 40 proteins that belong to the core autophagy machinery [66,67].

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].

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 avoid including 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, 11 had predicted long-range destabilizing effects, six had a damaging effect on stability, and only one variant (L53P) is characterized by both effects on stability and long-range communication (**Figure 5**). All the variants interested in long-range and structural destabilizing effects (except for W834G) 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[68]. 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]. We will include in the next MAVISp release the ensemble mode for MAP1LC3B to update the results of the STABILITY module.

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 phosphorylation at T49 (T49A variant).

Chaperones: TRAP1

The tumor necrosis factor receptor-associated protein 1 (TRAP1) is a mitochondrial chaperone involved in different biological processes of relevance in cancer [69,70]. 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 [71].

We here 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). Moreover, approximately ten 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 [72]. Among these variants, a subset is expected to decrease the flexibility at the distal site (i.e., P336L, P397L, L400F, Q479H, P569Q, V589L, and Q671H).

Another group has the opposite effect and is predicted to cause enhanced flexibility at the distal sites (M141T, K324N, R368H, R602Q, R647T, and V683A).

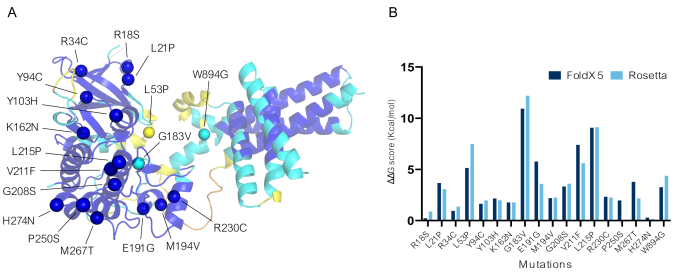

Figure 5. Destabilizing variants in the autophagy kinase ULK1. A-B) We identified 18 variants with destabilizing effects in terms of folding free energy or long-range effects.These mutations are highlighted with spheres on the AlphaFold models of the ULK1 N-terminal (Panel A, left) and C-terminal (Panel A, right) domain. B) We showed the predicted changes in folding free energy upon mutations for each of the 18 variants as calucalted by the STABILITY module of MAVISp with MutateX and RosettDDGPrediction. Interestingly, the majority of the variants that alter structural stability are located in the catalytic domain of the enzyme. This suggest potential mechanisms for ULK1 inactivation.

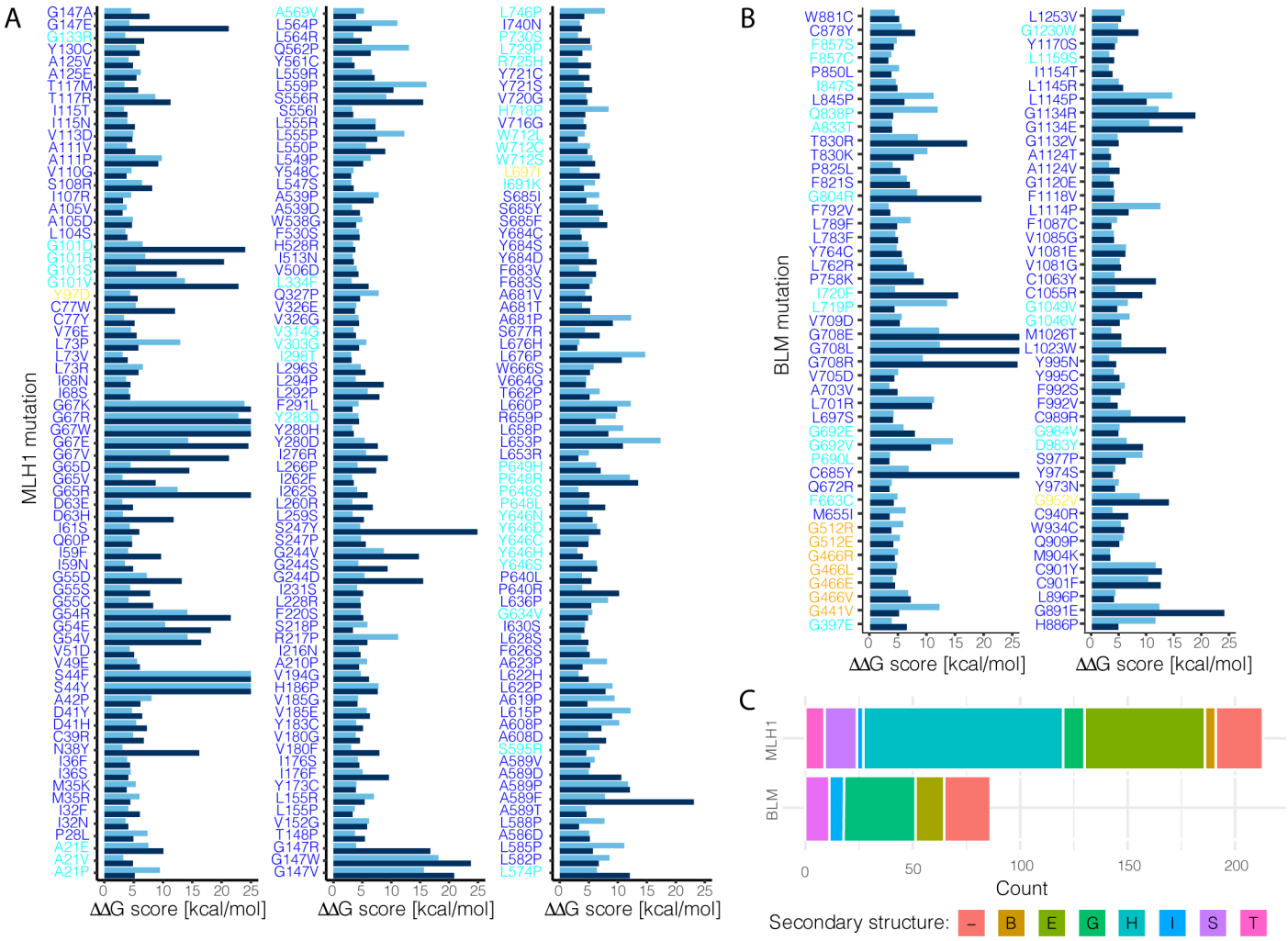

Figure 6. Variants with effects on structural stability in the tumour suppressors MLH1 and BLM. A-B) We show the 213 and 90 destabilizing variants for structural stability found in MLH1 (A) and BLM (B), respectively. We reported the results from the prediction with FoldX5 and Rosetta cartddg2020 protocol with the ref2015 energy function, as obtained with MutateX and RosettaDDGPrediction in blue and light blue, respectively. Values above 25 kcal/mol have been set to 25 kcal/mol. The labels of the variants are colored according to the AlphaFold pLDDT scores (i.e., blue: pLDDT > 90; cyan 70 <pLDDT < 90, yellow (50 <pLDDT <70) and orange (pLDDT < 50). **C)** The plot shows the distribution of the MLH1 and BLM destabilizing variants in terms of secondary structural elements as define by the DSSP dictionary (i.e, H= α-helix, B= residues in isolated β-bridge, E= extended β-strand, G= 3-helix, I= 5-helix, T= hydrogen bonded turn, S= bend and "-"= no secondary structure).

Peptidyl-prolyl cys-trans isomerases: PPIA

Prolyl cis-trans isomerases, such as cyclophilins and FK506-binding proteins play a central role in integrating protein conformers and signaling events [73]. 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), respectively. 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 [41].

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.

COSMIC Tumor Suppressor Genes and Oncogenes: MLH1, BLM, and KRAS

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 (MLH1, and BLM) and one oncogene (KRAS) reported in COSMIC.

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. We identified 90 and 213 destabilizing variants for BLM and MLH1 respectively (**Figure 6**), 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 6**).

Interestingly, in the case of 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 7**).

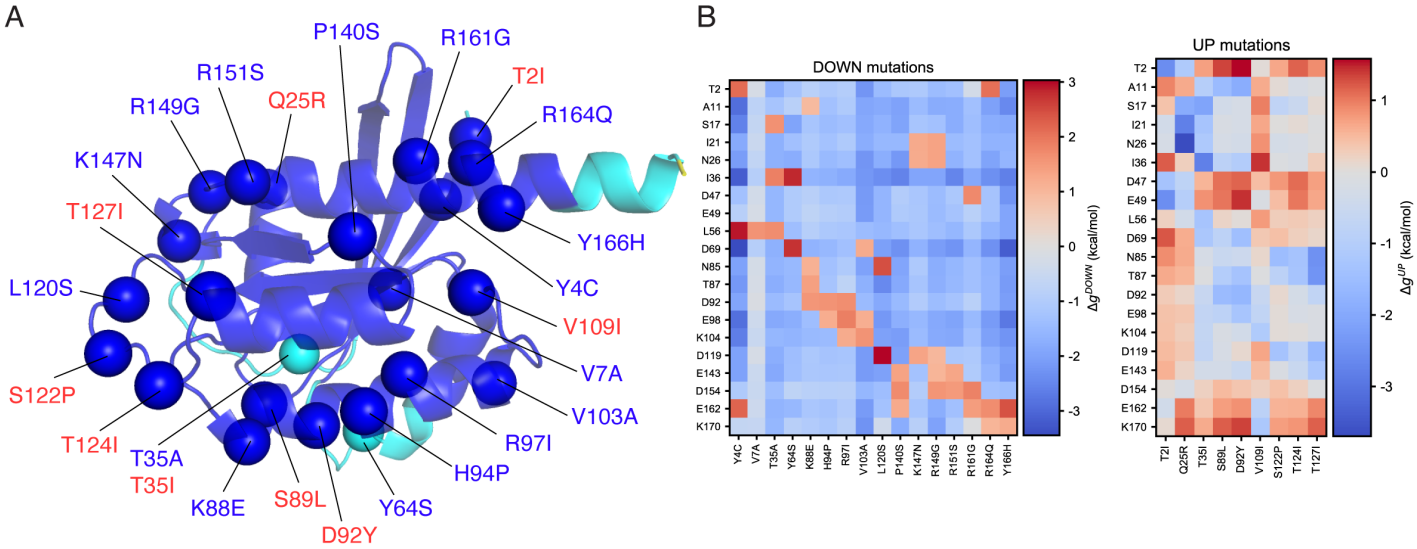

Figure 7. KRAS and allosteric mutations. A) We show the AlphaFold model of KRAS with residues color-coded according to the pLDDT score (i.e., blue: pLDDT >90, cyan 70 < pLDDT < 90, yellow 50 < pLDDT < 70 and orange pLDDT <50. The Cα atoms of the variant that can exert long-range effects are shown as spheres. The red labels indicate a mutation that decrease the flexibility of a distal site, while blue implies an increase flexibility to a distal site. B) The variant depicted in panel A are showed in the heatmap with the corresponding long-range communication sites.
We also indicates if the mutation has been modelled as a UP or DOWN effect in terms of steric hyndrance. On the x-axis, we show the variants that can have a long range impact to a distal site (i.e., at least 5 Å away from the mutation site. The distal site are shown on the y-axis. The colormap represents the changes in allosteric free energy induced upon mutation.

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 the variants from TRAP1, MAP1LC3B, and KRAS found in ClinVar and divided them into pathogenic, benign, or VUS/conflicted evidence (**Table S2**). It should be kept in mind that, in the current release, not all the proteins have been analyzed with the whole set of MAVISp modules. Thus, if a pathogenic variant is predicted as neutral at this stage, it might mean that additional layers of evaluation are needed. 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 the MAVISp layers related to functional changes to be classified correctly. On the contrary, we would expect that most of the variants currently classified as benign in ClinVar to be predicted as having neutral effects, if our framework is effective. This is what we observed for all the benign variants of TRAP1 and MAP1LC3B, even in cases that featured a high REVEL score. The only exception is Y444N (in the middle domain of TRAP1) which we predict as damaging for stability, but it is classified as likely benign in ClinVar. 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 to shed new light on uncharacterized variants. According to these predictions we identify as possibly pathogenic: W231R (TRAP1), I55S, and A134T (KRAS) in terms of altered structural stability; Q25R, Y64S, and Y64N, T124I, R161Q, R164Q (KRAS) in terms of long-range effects.

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 the currently 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 also has 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.
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 the REVEL pathogenicity score for illustrative purposes. Next, we will expand the portfolio of scores that are imported in MAVISp and can be used for comparison. We are currently working on

including EVE (Evolutionary model of variant effect) [74], DeMaSk [75], and GEMME (Global Epistatic Model for predicting Mutational Effects) [23].

Currently, MAVISp does not support transmembrane proteins, and we will implement Rosetta-based protocols for this class of proteins in future releases of RosettaDDGPrediction. For example, the protocol provided by Tiemann et al. [76] could broaden the possibility of expanding 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.

In addition, 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 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 eight target proteins (BCL2, BCL2L1, BLM, KRAS, MAP1LC3B, MLH1, TRAP1, and ULK1) and trimmed them to remove regions with pLDDT scores < 70 at the N- or C-termini or long disordered linkers between folded domains. On the contrary, we used an X-ray (PDB entry 3K0N [77]) and an NMR structure (PDB entry 2KK0 [78]) for PPIA and ARID3A, respectively.

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].

For the calculations of allosteric free energy, we used the structure-based statistical mechanic model implemented in AlloSigMA2 [47]. 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 1 kcal/mol or lower than -1 kcal/mol to sites that were at least at 5 Å of distance. The distance has been calculated using the PyInteraph2 center-of-mass PSN [45].

Molecular dynamics simulations

For ARID3A, we used previously published [56] all-atom molecular dynamics simulations of one microsecond with the CHARMM22star force field [79].

MAVISp score

To rank the variants annotated with MAVISp, we calculated a Stability score for the STABILITY module, a Function score for the impacts on PTM, LOCAL_INTERACTION, and LONG_RANGE modules, and a MAVISp collective score for the global damaging effects. The approach is based to the ones applied to MAP1LC3B in our previous pilot project [30]. We assigned a score of 1 to variants classified as: i) destabilizing or stabilizing by the STABILITY and LOCAL_INTERACTION modules, iii) damaging or potentially damaging for changes in regulation, stability, or function of post-translational modifications by the PTM module, iv) 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. The Stability score includes the score from the STABILITY module. The Function score is the summation of the three scores from the PTM module (PTM-regulation, PTM-stability, and PTM-function), and the scores for the LOCAL_INTERACTION and the LONG-RANGE module. The score is then normalized between 0 and 1. We defined the MAVISp collective score as the summation of the Stability and Function score upon normalization between 0 and 1. In future releases, we will evaluate different scoring schemes against benchmarking datasets.

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**, (2022).
41. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. doi:10.1101/2022.09.02.506350.
42. Nielsen, S. v. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* **13**, e1006739 (2017).
43. 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.
44. 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).
45. Sora, V., Tiberti, M., Mahdi Robbani, S., Rubin, J. & Papaleo, E. PyInteraph2 and PyInKnife2 to analyze networks in protein structural ensembles. *bioRxiv* 2020.11.22.381616 (2020).
46. 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).
47. 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).
48. Guarnera, E., Tan, Z. W., Zheng, Z. & Berezovsky, I. N. AlloSigMA: Allosteric signaling and mutation analysis server. *Bioinformatics* **33**, 3996–3998 (2017).
49. Hornbeck, P. v. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512–D520 (2015).
50. 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).
51. Henriques, J. & Lindorff-Larsen, K. Protein Dynamics Enables Phosphorylation of Buried Residues in Cdk2/Cyclin-A-Bound p27. *Biophys J* **119**, 2010–2018 (2020).
52. 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).
53. Henrie, A. *et al.* ClinVar Miner: Demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. *Hum Mutat* **39**, 1051–1060 (2018).
54. 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).
55. 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).
56. 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).
57. Nygaard, M. *et al.* The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci* **3**, 1–18 (2016).
58. 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).
59. 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).
60. 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.
61. 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).
62. 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).
63. Singh, S. S. *et al.* Dual role of autophagy in hallmarks of cancer. *Oncogene* **37**, 1142–1158 (2018).
64. White, E. & DiPaola, R. S. The double-edged sword of autophagy modulation in cancer. *Clinical Cancer Research* **15**, 5308–5316 (2009).
65. Thorburn, A. Autophagy and Its Effects: Making Sense of Double-Edged Swords. *PLoS Biol* **12**, 10–13 (2014).
66. 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).
67. Suzuki, H., Osawa, T., Fujioka, Y. & Noda, N. N. Structural biology of the core autophagy machinery. *Curr Opin Struct Biol* **43**, 10–17 (2017).
68. Shu-Yong, L. *et al.* GSK3-TIP60-ULK1 signaling pathway links growth factor deprivation to autophagy. *Science (1979)* **336**, 477–481 (2012).
69. Rasola, A., Neckers, L. & Picard, D. Mitochondrial oxidative phosphorylation TRAP(1)ped in tumor cells. *Trends Cell Biol* **24**, 455–463 (2014).
70. 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).
71. 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).

72. 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.
73. 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).
74. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
75. Munro, D. & Singh, M. DeMaSk: A deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36**, 5322–5329 (2020).
76. 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.
77. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).
78. 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).
79. 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, MU, SS, ML, EP. *Formal Analysis*: MA, LB, MU, ML, EP. *Funding Acquistion:* EP, MA. *Investigation*: MA, LB, MU, EP. *Methodology*: MA, LB, KD, AP, ML, MT, EP. *Project administration:* EP. *Resources*: PWS, EP. *Software*: KD, MT, PWS. *Supervision*: MT, EP. *Validation*: LB, ML, MT, EP. *Visualization*: MA, LB, KD, MU, ML, MT. *Writing – Original Draft:* EP. *Writing – Review and Editing*: MA, LB, KD, MU, 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. The work is also supported by A PhD Fellowship from the Danish Data Science Academy (DDSA) to M.A.
This work is part of Interregional Childhood Oncology Precision Medicine Exploration (iCOPE), a cross-Oresund collaboration between University Hospital Copenhagen, Rigshospitalet, Lund University, Region Skåne, and Technical University Denmark (DTU), supported by the European Regional Development Fund.