Allostery is a widespread cause of loss-of-function variant pathogenicity

Xiaotian Liao[1,2], Ben Lehner[1,3,4,5*]

[1] Wellcome Sanger Institute, Cambridge, UK

[2] University of Cambridge, UK

[3] Centre for Genomic Regulation (CRG), Barcelona Institute for Science and Technology (BIST), Barcelona, Spain

[4] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
*correspondence e-mail: bl11@sanger.ac.uk

Abstract

Allosteric communication between non-contacting sites in proteins plays a fundamental role in biological regulation and drug action. While allosteric gain-of-function variants are known drivers of oncogene activation, the broader importance of allostery in genetic disease and protein evolution is less clear. Here, we integrate large-scale experimental measurements and neural network models to provide evidence that allostery is a widespread cause of loss-of-function variant pathogenicity in human genetic diseases. In addition, our analyses reveal a conserved distance-dependent decay of allosteric mutational effects outside of protein active sites. As an important mechanism of pathogenicity, allostery needs to be better mapped, understood, and predicted across the human proteome.

Introduction

Allostery, the regulation of a protein's activity through changes at sites distant from an active site, underpins protein regulation[1–3]. Allostery allows proteins to function as molecular switches and microprocessors, with their enzymatic activities and molecular interactions quantitatively controlled by small molecule binding, post-translational modifications, and macromolecular interactions[4]. Allostery is central to the control of metabolic pathways, signal transduction, growth control and neural information processing in both normal physiology and disease[1,3,5–9]. Moreover, many effective therapeutics bind allosteric sites and have allosteric mechanisms-of-action, including drugs targeting ion channels, receptors, kinases, and transcription factors[2,10–14].

The importance of allostery for human genetic disease is, however, less well established[6,15]. Allosteric gain-of-function (GOF) variants are a recognized class of activating mutations in human oncogenes[2,16]. For example, the G12C and other driver mutations in KRAS allosterically increase binding to effector proteins such as the RAF kinase[17,18], and the V617F mutation within the pseudokinase domain of the JAK2 tyrosine kinase disrupts its autoinhibitory interaction with the kinase domain, leading to constitutive activation through trans-phosphorylation[19,20]. Similarly, allosteric mutations in the epidermal growth factor receptor (EGFR), including L858R and T790M, increase kinase activity and confer resistance to targeted therapies in non-small cell lung cancer[21,22]. An increasing number of licensed cancer therapeutics have allosteric mechanisms of action, including the first effective KRAS inhibitors[6,15,23].

Most pathogenic variants causing human genetic diseases are, however, loss-of-function (LOF) mutations[24,25]. LOF variants include premature stop codons (PTCs) and splice-site variants, but also a large number of protein missense variants[26,27]. Missense variants that change the amino acid (AA) sequence of a protein can have diverse molecular effects. These include destabilization of a protein's fold, with recent studies suggesting that 39-60% of clinically pathogenic missense variants reduce protein stability[28–32]. However, variants can also impair molecular activities beyond stability, such as disrupting molecular interactions or enzymatic activities[31,33–41]. We term these variants 'functional mutations' to distinguish them from variants affecting protein stability.

A small number of LOF variants are proposed to have allosteric mechanisms-of-action[6,24,42,43]. Examples include the G551D variant in the cystic fibrosis transmembrane conductance regulator (CFTR)[44–46], as well as distal variants such as K246E and L311V in adenylosuccinate lyase (ADSL), which reduce enzymatic activity and cause ADSL deficiency—a metabolic disorder linked to autism and epilepsy[47]. Nevertheless, the broader contribution of allosteric loss-of-function variants to human genetic disease remains largely uninvestigated, a shortcoming that we seek to address in the present study.

Recent comprehensive mutagenesis studies have generated the first systematic experimental maps of allosteric communication throughout proteins[39–41,48,49]. These near complete atlases for a handful of proteins have revealed that many mutations—even those distant from active sites and binding sites—can exert measurable allosteric effects on protein function. For example, 114

amino acid substitutions in the oncoprotein KRAS allosterically inhibit binding to the oncoprotein effector RAF1[41]. Even in small protein domains such as SH3 domains, many variants have small but measurable allosteric effects[39] that, in combination, can strongly impair binding to interaction partners[50]. These first allosteric maps suggest that allostery is widespread and highlight the need to reconsider the importance of allostery for variant pathogenicity and molecular evolution.

Here, we evaluate the contribution of allostery to loss-of-function human variant pathogenicity. Across extensive experimental datasets and state-of-the-art computational models, our analyses indicate that loss-of-function pathogenic variants frequently have allosteric mechanisms-of-action, disrupting protein function without destabilizing the fold or directly disrupting catalytic residues and binding sites. Across multiple analyses and structurally diverse proteins, these loss-of-function allosteric effects are enriched close to protein active sites and decay exponentially through protein structures. Our results indicate that allostery is a widespread and underappreciated mechanism of variant pathogenicity that needs to be better understood and predicted across the human proteome.

Results

Pathogenic variants often act through mechanisms beyond stability

From a structural perspective, missense mutations that impair protein function can act through three primary mechanisms: (1) globally destabilizing the protein fold, (2) directly disrupting the active or binding site, or (3) indirectly perturbing function through distal or allosteric effects. To estimate the proportion of mutations that act via destabilization, we first applied computational methods to predict the impact of each variant on protein stability. To assess the contribution of functionally disruptive mutations—those whose effects go beyond folding—we applied state-of-the-art deep learning approaches. Specifically, we used ThermoMPNN, an inverse folding neural network[51] fine-tuned on large-scale experimental data[31], to predict changes in folding free energy (ΔΔGf) for all possible missense mutations across 19,804 human proteins. This analysis encompassed 193,469,108 variants, including 113,994 annotated in the ClinVar database[52,53] **(Fig. 1A, Supplementary Table 1)**.

To quantify how much of variant pathogenicity is attributable to changes in stability, we compared ThermoMPNN-predicted ΔΔGf values to ESM1v scores—a model that estimates the functional impact of mutations from evolutionary sequence conservation[54]. Across the entire proteome, predicted ΔΔGf explained only a modest proportion of the variance in ESM1v pathogenicity scores **(Fig. 1B, median R^2 = 0.224, interquartile range (IQR) 0.125-0.298)**, suggesting that folding stability is only one of several mechanisms underlying missense variant pathogenicity.

Since most active and binding residues are solvent-exposed, we reasoned that analyses limited to buried residues would isolate stability-driven effects. However, even when excluding solvent exposed residues (relative solvent accessible surface area, rSASA > 0.2), predicted changes in stability still only partially explained predicted pathogenicity **(Fig. 1B, median R^2 = 0.158, IQR 0.096 - 0.232)**. This lower R^2 in buried regions may in part reflect a more restricted range of

ESM1v scores in the protein core, limiting the total explainable variance. The pattern held across functional classes: folding stability accounted for just ~10% of pathogenicity in voltage-gated ion channels and nuclear receptors, while explaining ~23-26% in ribosomal proteins, immunoglobulins, and T-cell receptors **(Fig. 1C)**.

Thus, state-of-the-art models suggest that stability is an important component of variant pathogenicity, but they predict that many variants are detrimental because of functional effects beyond stability changes.

Analysis of large-scale experimental abundance datasets

To go beyond computational predictions, we next turned to direct experimental measurements of changes in stability. We analysed data from two recent large-scale studies that together experimentally quantified changes in abundance for 598,338 variants in 578 different protein domains[31,32]. The first 'MegaScale' dataset used protease-sensitivity to quantify changes in folding free energy for 272,712 missense variants in 331 protein domains *in vitro*. The second 'Domainome' dataset used a protein-fragment complementation assay to quantify changes in cellular abundance for 607,081 missense variants in 522 protein domains **(Fig 1D)**. For variants with matching ESM1v scores, the experimentally-quantified abundance changes still only partly account for ESM1v-predicted pathogenicity **(Fig. 1E, median R^2 = 0.252 and 0.218, respectively)**. These results match the computational predictions: many variants appear pathogenic despite maintaining near-normal stability or expression.

To assess whether this trend holds in full-length, physiologically relevant contexts, we analysed experimental abundance measurements for nine complete human proteins (Variant Abundance by Massively Parallel sequencing - 'VAMP-seq' dataset), spanning various functions and sizes **(Supplementary Table 2)**. Across these full-length proteins, abundance changes explained a median of only 26% of ESM1v variance **(Fig. 1F, R^2 range: 13.9% – 40.2%)**, closely matching the domain-level findings.

Evaluation of ClinVar variants

Although protein language models provide a useful proxy for pathogenicity, we sought to validate our findings using clinically annotated variants. To this end, we turned to ClinVar, a curated database that classifies variants as pathogenic, benign, or of uncertain significance[52,53]. Comparing 42,576 pathogenic variants against 71,418 benign variants, we found that pathogenic variants are far more likely to be predicted as destabilizing **(Fig. 1G, 48.3% vs 8.8% odds ratio (OR) = 9.653, p < 2.2e-16, Fisher's exact test (FET))**. Predicted stability changes are thus a very useful, albeit imperfect, predictor of pathogenicity (Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) = 0.697, Precision–Recall Area Under the Curve (PR-AUC) = 0.618, n = 42,576 pathogenic, n = 71,418 benign variants).

As expected[55], mutations in the buried protein cores were more likely to be destabilizing (OR = 8.072, p < 2.2e-16, FET) and also more likely to be pathogenic (OR = 4.243, p < 2.2e-16, FET). In total, 67.4% of pathogenic variants in the core were predicted to be destabilizing compared to 32.8% of benign core variants **(Fig. 1G)**. Variations were observed across protein functional

classes, with the proportion of stability-affecting (destabilizing or stabilizing) pathogenic variants ranging from 38-43% in secreted proteins, voltage-gated ion channels, and transcription factors to 60% in metabolic proteins **(Fig. S2A)**.

Experimentally-measured stability changes presented a very similar picture: 52% of clinically pathogenic variants in protein domains and 56.3% of pathogenic variants in full-length proteins are measured to disrupt stability **(Fig. 1G)**. Thus, while experimentally-measured stability changes are predictive of pathogenicity—ROC-AUC = 0.70 and PR-AUC = 0.87 for domains (n = 471 pathogenic, 161 benign); ROC-AUC = 0.72 and PR-AUC = 0.96 for full-length proteins (n = 268 pathogenic, 24 benign)—additional pathogenic mechanisms must be widespread.

Together, these data demonstrate that while loss of stability is a common mechanism of pathogenicity, a substantial fraction of pathogenic variants must act through other mechanisms—such as disrupting active sites, perturbing binding interfaces, or exerting long-range allosteric effects.

Functional effects decay exponentially with distance

Allosteric propagation—the transmission of energetic changes from one site in a protein to another—provides a mechanistic basis for how mutations distant from active sites can still impact function[56,57]. As a baseline, in the absence of allosteric effects, one would expect no relationship between functional effect size and distance from active site. Mutations would primarily act locally, and effects would cluster strictly at active sites. However, in the presence of allostery, mutational effects are expected to propagate from active sites through the structure, and we would expect functional effects to decay with 3D distance—strongest near the active site and weaker further away. This distance-dependent allosteric decay is visible in the small number of experimental allosteric maps generated to date[39–41,48,49].

To test whether variant effects beyond stability changes include long-range allosteric mechanisms, we first analysed missense mutations across six structurally and functionally diverse full-length human allosteric proteins **(Fig. 2A–B, Supplementary Table 3)**. For each allosteric protein in our test set, we compared predicted pathogenicity (ESM1v) to predicted folding stability changes (ΔΔGf) from ThermoMPNN. We then fitted a Locally Estimated Scatterplot Smoothing (LOESS) curve and computed residuals to quantify variant effects not attributable to stability changes **(Fig. 2C, Fig. S3B)**.

Across all six proteins, many variants exhibited large residuals, indicating functional effects not accounted for by changes in fold stability. ESM1v and ΔΔGf were moderately correlated (Spearman's $\rho = -0.25$ to -0.67), with enrichment of large residuals at annotated orthosteric sites and distal regulatory positions where known allosteric regulators bind **(Fig. 2B, D, Fig. S3A, C)**.

We next asked whether these stability-independent functional effects follow a spatial pattern consistent with allosteric propagation[39–41,48,56,57]. For each residue, we computed its minimal distance to the ligand atoms and plotted the absolute median residual against this distance. In all proteins, the absolute median residuals were strongest for residues close to the active site,

with the predicted functional effects of mutations decaying approximately exponentially with 3D distance from the ligand **(Fig. 2D, Fig. S3B)**. This trend was statistically significant: in a linear model regressing the log-transformed absolute residuals on distance, increasing distance was significantly associated with weaker functional effects (Supplementary Table 3). Across proteins, the half-distances for decay ($d_{1/2}$) are comparable to the distance-dependent decay of allosteric mutational effects observed in experimental allosteric maps (GCK: 15.7 Å, PTP1B: 19.7Å, PDK1: 12.3Å, CHK1: 15.9 Å; Supplementary Table 3)[39–41,48,49].

To determine whether the absolute median residuals are increased at known allosteric sites, we compared their magnitudes at orthosteric sites, distal sites where known allosteric regulators bind, and randomly sampled control positions matched for protein abundance. In four out of the six proteins (all except PTP1B and CASP1), absolute median residuals were significantly higher at both orthosteric and known allosteric sites than at control positions **(Fig. 2E, Fig. S3D)**. In PTP1B and CASP1, allosteric sites also showed larger absolute residuals on average **(Fig. S3E)**, although the difference was not statistically significant. It is important to note that this definition of allosteric sites is likely conservative, as it is based only on the binding sites of known allosteric regulators. Most allosteric proteins will harbour additional allosteric sites that remain untested.

Notably, the distance-dependent decay is not uniform. Some positions exhibit residuals much larger than expected for their distance from the active site. We refer to these as allosteric hotspots—residues with more functional impacts than expected. These hotspots include binding sites of known allosteric regulators, such as inhibitors of protein tyrosine phosphatase 1B (PTP1B) and compound RS1, which targets pyruvate dehydrogenase kinase 1 (PDK1).[58,59].

These results suggest that allosteric mechanisms potentially underlie a substantial and underappreciated component of pathogenicity in missense variants, and that the magnitude of these effects decays in a conserved, distance-dependent manner from functional sites. To benchmark the validity of our computational predictions, we compared ESM1v–ThermoMPNN ΔΔGf residuals to experimentally measured, stability-independent activity effects from published deep mutational scanning maps[37,39,60–62]. Across all proteins for which matched in silico and experimental data were available, we observed a consistent and significant correlation between computational residuals and experimentally defined allosteric effects **(Fig. S4)**, supporting the relevance of our approach.

Experimental allosteric maps confirm distance-dependent decay

While computational predictions enable proteome-wide predictions of allosteric variant effects, they do not directly measure functional consequences of mutations. To address this limitation, we turned to experimental allosteric maps that quantify the impact of every missense mutation on both folding and binding—across a smaller but well-characterized set of proteins.

Recent advances have enabled the generation of comprehensive experimental allosteric maps for individual proteins and protein domains[39–41,48,49]. For protein–protein interaction domains, these maps report the effect of every possible missense mutation on folding free energy (ΔΔGf) and binding free energy (ΔΔGb), allowing measurement of both stability and functional

consequences of mutations at each site[39,41]. In these systems, mutations that alter binding without directly contacting the ligand must, by definition, act through an allosteric mechanism. To evaluate this, we focused on two domains with complete mutational maps: the third PDZ domain of PSD-95 and the SH3 domain of GRB2 **(Fig. 3A)**.

As observed in our proteome-wide analyses, folding stability changes alone explained only a modest fraction of predicted pathogenicity: $R^2 = 0.214$ for PSD-95 and $R^2 = 0.359$ for GRB2 **(Fig. 3F)**. To isolate the stability-independent component of pathogenicity, we fitted a LOESS curve between ΔΔGf and ESM1v predicted pathogenicity for each domain and computed residuals, quantifying the extent to which mutations disrupt function beyond what would be expected from their impact on folding stability alone **(Fig. 3B, Fig. S5A)**.

As expected, the largest absolute median residuals were observed at orthosteric binding residues and at experimentally annotated allosteric mutations (Fig. 3C). These effects also followed a spatial pattern consistent with allosteric propagation: absolute residual magnitudes were highest near the binding interface and decayed with increasing 3D distance from the ligand (Fig. 3D, PSD-95-PDZ3: slope = -0.14, P = 2.46e-07; GRB2-SH2: slope = -0.09, P = 2.06e-05). Visualizing the median ESM1v-ΔΔGf residual per site further illustrates this distance dependence of predicted functional effects away from the binding interface **(Fig. 3E)**.

We next evaluated experimentally measured ΔΔGb values to directly assess allosteric effects. For residues outside the binding interface, changes in ΔΔGb isolate the allosteric component of binding disruption. These values explained 30.2% of ESM1v variance in PSD-95 and 4.5% in GRB2 **(Fig. 3F)**. The smaller contribution of ΔΔGb to pathogenicity in GRB2-SH3 likely reflects the domain's very small size and reduced allostery.

Importantly, binding free energy changes (ΔΔGb) also showed a distance-dependent decay from the ligand **(Fig. S5C)**. Across residues, ΔΔGb values correlated with the magnitude of ESM1v–ΔΔGf residuals (Fig. S5D), with a stronger relationship in GRB2 (Spearman's $\rho = -0.70$) than in PSD-95 ($\rho = -0.31$). Moreover, models incorporating both ΔΔGf and ΔΔGb explained more pathogenicity variance than models using stability alone—even after excluding mutations within the binding interface **(Supplementary Tables 4–7)**. This further demonstrates the allosteric contribution to pathogenicity is independent of changes in fold stability.

Clinical pathogenic variants in PTEN and GCK show allosteric decay

Finally, to further evaluate the contribution of allostery to clinical pathogenicity, we analysed two human proteins with many known disease-causing variants and large-scale experimental data: the lipid phosphatase PTEN and the glucose-sensing kinase glucokinase (GCK) **(Fig. 4A, 5A)**. For both proteins, the effects of thousands of variants on abundance and enzymatic activity had been measured, enabling direct assessment of functional impact—including effects that are independent of abundance.

PTEN germline variants are associated with a spectrum of disorders, including macrocephaly, autism spectrum disorder (ASD), developmental delay, and PTEN hamartoma tumor syndrome (PHTS), which encompasses Cowden syndrome and other cancer predisposition syndromes[63].

Based on established criteria[64], we grouped PTEN variants into three phenotypic categories: (1) neurodevelopmental, (2) cancer-predisposition, and (3) combined. In addition to these inherited phenotypes, PTEN is a well-established tumour suppressor, with somatic missense mutations frequently acting as drivers in glioblastoma, endometrial, prostate, and breast cancers[65].

GCK germline mutations cause two opposing metabolic disorders depending on how they affect enzymatic function. Loss-of-function variants impair glucose sensing and lead to GCK-MODY (Maturity-Onset Diabetes of the Young, type 2), a mild familial form of hyperglycemia. In contrast, gain-of-function mutations cause congenital hyperinsulinemic hypoglycemia (HH), a neonatal disorder of excessive insulin secretion and persistent hypoglycemia[62,66,67].

In PTEN, intracellular abundance was measured for 5,083 missense variants using VAMP-seq, and phosphatase activity was measured for 7,260 variants via a yeast-based complementation assay[60,61] **(Fig. 4A)**. In GCK, abundance (8,396 variants) and catalytic activity (8,570 variants) were similarly quantified using yeast-based systems[37,62] **(Fig. 5A)**.

To isolate functional effects beyond abundance changes, we fitted LOESS curves between abundance and activity using only non-orthosteric variants to define a baseline relationship. Residuals from this fit capture changes in activity not explained by changes in protein abundance **(Fig. 4B, 5B)**.

As expected, many pathogenic ClinVar variants reduced protein abundance. However, many others reduced enzymatic activity more than expected given their abundance levels **(Fig. 4C, 5C)**. Mapping these residuals onto the protein structures revealed their enrichment in both active sites and distal allosteric regions. In PTEN, high-residual variants localized to the catalytic core and surface-exposed residues, including the WPD, P, and TI loops that form the active site pocket **(Fig. 4F)**. In GCK, pathogenic variants clustered not only in the glucose-binding pocket but also in helix 13, a known allosteric regulator of activity[68,69] **(Fig.5F)**.

In contrast, common variants from gnomAD and clinically benign variants (e.g., PTEN A79T; GCK A11T, R46T, E279Q) generally maintained wild-type-like abundance and activity, and showed minimal residuals[70] **(Fig. 4B,D; Fig. 5B,D)**. Among pathogenic variants, the activity residuals were often especially large for variants that did not strongly reduce abundance **(Fig. S6B)**. This distinction is especially relevant for PTEN, where highly stable but catalytically inactive variants are known to act as dominant negatives—interfering with the function of the wild-type protein despite retaining normal expression levels. Consistent with this, recurrent PTEN cancer hotspot mutations and known dominant negative variants showed high magnitudes of activity residuals[71] **(Fig. 4D)**, underscoring the importance of how, not just whether, phosphatase activity is lost.

To assess whether allosteric effects decay with distance from functional sites, we plotted the absolute abundance-corrected activity residuals of each variant against its distance from the bound ligand **(Fig. 4G, 5G)**. In both PTEN and GCK, we observed a clear exponential decay: variants closest to the orthosteric sites had the strongest effects, and the typical effect size dropped by half over distances of 14.5 Å (PTEN) and 15.8 Å (GCK). This distance-dependent pattern was statistically significant in both proteins. Linear regression of the log-transformed

absolute residuals against distance yielded negative slopes (PTEN: slope = –0.033, P = 2.99e-07; GCK: slope = –0.046, P = <2e-016), consistent with a conserved distance-dependent decay of stability-independent functional effects. These findings mirror the distance-dependent decay of mutational effects observed in experimental allosteric maps—including decay from binding interfaces in protein interaction domains[39] ,in KRAS[41], and in activity-altering mutations in the Src kinase[40], as well as in the allosteric phytohormone receptor PYL1[48]. They are also similar to the decay of abundance corrected mutational effects on predicted fitness presented above **(Fig. 2)**. Together, these results strongly support a conserved biophysical principle of distance-dependent allosteric decay.

Both PTEN and GCK contain clinically pathogenic variants that clustered in the active site and interface regions—as expected—but many were also located at distal positions **(Fig. 4E, 5E)**. To test whether such non-orthosteric pathogenic variants exert functional effects beyond abundance loss, we compared their activity–abundance residuals to those of abundance-matched, non-pathogenic controls. In both orthosteric and non-orthosteric regions, pathogenic variants showed significantly larger residuals than abundance-matched controls **(Fig. 4H, 5H)**, indicating greater loss of activity than expected from abundance effects alone. Many pathogenic variants in PTEN and GCK therefore likely act through allosteric mechanisms.

Discussion

Taken together, our results indicate that allostery is an important cause of loss-of-function variant pathogenicity. Although many pathogenic variants destabilize proteins[28–32], changes in stability only partially account for the pathogenicity of variants. Both state-of-the-art computational predictions **(Fig. 1,2)**, and direct experimental measurements **(Fig. 3-5)** indicate that the indirect, allosteric effects of mutations on activity are a frequent mechanism of pathogenicity. In both PTEN and GCK, for example, many known pathogenic variants have no or weak effects on stability but strong experimentally-measured allosteric effects on activity.

To date, comprehensive experimental allosteric maps have only been generated for a handful of proteins[39–41,48,49] and a limitation of our study is the extrapolation of conclusions from these proteins to the whole proteome. Nevertheless, results from our proteome-wide analyses using state-of-the-art computational predictors **(Fig. 1-2)** are highly consistent with the conclusions from the experimental allosteric maps **(Fig. 3-5)**. Ultimately, it will be important in future work to scale up the production of experimental allosteric maps, particularly for proteins that harbor large numbers of clinically observed variants[37,61,62]. Furthermore, we have only experimentally considered protein-protein binding and enzyme catalytic activity in this study; future efforts should develop methods to quantify allosteric effects on additional molecular activities, such as DNA binding and small-molecule interactions. A comprehensive understanding of pathogenic mechanisms will ultimately require experimental tools capable of capturing allosteric regulation across the full spectrum of disease-relevant molecular activities.

Additional caveats should be noted. First, ESM1v is a proxy for variant pathogenicity and evolutionary fitness derived from protein sequences; while it captures many aspects of mutational impact on protein function, it is not a direct measure of clinical pathogenicity. Second,

our computational analyses and the equivalent experimental data aim to quantify allostery but they do not identify the underlying causal molecular mechanisms, which likely involve changes in both conformation and protein dynamics[3,8,72–74].

However, despite this diversity of possible mechanisms, a conserved principle of allosteric communication appears to be the distance-dependent decay that we have highlighted here. An approximately exponential distance-dependent decay of changes in activity after correcting for abundance changes is seen in all seven experimental allosteric maps published to date[39–41,48,49] and it is also observed in the PTEN and GCK allosteric maps that we have presented here **(Fig. 4-5)**. It is also visible when plotting computational predictions of variant pathogenicity after accounting for predicted stability effects **(Fig. 2)**. Exponential distance-dependent decay has been observed for evolutionary conservation in enzyme active sites, for chemical shift perturbations in nuclear magnetic resonance (NMR) experiments and for energetic couplings in molecular dynamic simulations[56,57,75]. We believe, therefore, that distance-dependent allosteric decay is likely to be a conserved general principle of protein biophysics.

Pervasive and distance-dependent allostery has important implications for human genetics. First, many loss-of-function variants are likely to have allosteric mechanisms of action. Second, our data predict that pathogenic variants are more likely to have allosteric mechanisms when they are located structurally close to active sites. Third, and conversely, drugs targeting non-orthosteric protein pockets are more likely to have strong allosteric effects when they target pockets closer to a protein's active site. In future work, it will be important to further refine the rules of allostery, allowing improved mechanistic interpretation of variant pathogenicity and prediction of the most effective sites to target to therapeutically modulate protein function.

Figure legends

Figure 1

Pathogenic variants often act through mechanisms beyond stability.

A. Overview of datasets used for proteome-wide computational analysis. Predicted pathogenicity scores were generated using ESM1v for all missense variants in 19,804 human proteins. Predicted folding stability changes (ΔΔGf) were computed using ThermoMPNN (kcal/mol).

B. Distributions of R^2 values from linear regression between predicted ΔΔGf and ESM1v scores for all missense variants across the human proteome (left), and for protein core variants (right), defined by excluding disordered, surface-exposed and predicted binding sites. Median R^2 values are indicated.

C. Distribution of R^2 values, stratified by protein type (core variants only). Median R^2 values are labeled. Protein types were defined based on the "Function and compartment based" classification provided by the Human Protein Atlas[76–78].

D. Overview of datasets used for large-scale experimental measurements. Protein abundance changes were quantified using multiple methods: *in vivo* DHFR complementation assay (Domainome 1.0[79]), *in vitro* cDNA display (Mega-scale[31]), and VAMP-seq[61] (variant abundance by massively parallel sequencing) for full-length proteins.

E. R^2 distributions for abundance changes regressed against ESM1v scores for the Domainome 1.0 and Mega-scale datasets. Median values are labeled.

F. R^2 values between ESM1v pathogenicity scores and variant-induced abundance changes for nine full-length proteins profiled using VAMP-seq. Protein types are indicated by bar color. Number of variants (n) and R^2 values are shown.

G. Classification of ClinVar missense variants into three categories: WT-like (green), destabilizing (orange), and stabilizing (blue), based on their impact on protein abundance. Percentages and total counts are shown for all ClinVar variants included in the proteome analysis (left), protein core variants (middle-left), variants from Domainome 1.0 (middle-right), and VAMP-seq (right), stratified by benign and pathogenic clinical labels.

Figure 2

Functional effects of mutations decay exponentially with distance.

A. Overview of the analytical framework. Predicted variant pathogenicity scores from ESM1v and folding stability predictions from ThermoMPNN (ΔΔGf) were used to compute residuals (ESM1v–ΔΔGf), representing functional effects not explained by

structural destabilization. These residuals were then analyzed for spatial decay relative to distance from the substrates.

B. Crystal structures of GCK (PDB: 1V4S), PTP1B (PDB: 1PTY), PDK1 (PDB: 4RQK), and CHK1 (PDB: 2E9N), colored by the median residuals between ESM1v and ThermoMPNN (TMPNN) predicted folding free energy changes (ΔΔGf) values. Bound substrates — glucose (GLC), O-phosphotyrosine (PTR), adenosine triphosphate (ATP), and-3-(4'-hydroxybiphenyl-4-yl)-N-(4-hydroxycyclohexyl)-1,4-dihydroindeno[1,2-c]pyrazole-6-carboxamide (76A) — are shown in orange. Allosteric inhibitor RS1 is labeled in blue.

C. Scatter plots of ThermoMPNN-predicted ΔΔGf versus ESM1v scores for missense mutations in each protein. A LOESS curve (black line) models the trend, with residuals quantifying deviation between folding impact and predicted pathogenicity. Stable mutations cluster around ΔΔGf ≈ 0.

D. Per-residue relationship between the absolute median ESM1v–ΔΔGf residual and the minimal Cα distance (Å) to the substrates for each structure. Residuals are calculated using only mutations in non-orthosteric sites with negative residuals. An exponential decay model (black curve) was fitted to the data, with 95% confidence intervals shown as grey shading. The grey dashed line indicates the maximum of the minimal distances from any substrate-contacting (orthosteric) residue. Sites are colored by functional class: orthosteric sites in orange circles, non-orthosteric sites in dark green circles, and known allosteric sites (i.e., sites in physical contact with known allosteric regulators) in cyan triangles.

E. Distributions of per-site absolute median residuals between ESM1v and ΔΔGf for orthosteric, allosteric, and other sites. Each point represents the absolute median ESM1v-ΔΔGf residual for a site. For orthosteric and allosteric sites, residuals are calculated directly per annotated site. For the 'other' group, distributions are derived from 1,000 bootstrap replicates, each sampling the same number of non-orthosteric and non-allosteric sites as in the orthosteric group. Violin plots show the distribution shape; black bars and diamonds indicate medians. Sample sizes and median values are annotated. Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

Figure 3

Experimental allosteric maps confirm distance-dependent decay of functional mutational effects.

A. Overview of the functional assays used to assess the effects of missense mutations on protein abundance and binding in two human domains: PSD-95-PDZ3 and GRB2-SH3. Protein abundance and binding were quantified using protein fragment complementation (PCA) assays (abundance PCA: aPCA, binding PCA: bPCA). The residual between ESM1v scores and experimentally measured folding free energy changes (ΔΔGf)

captures functional effects not explained by stability and is used to assess distance-dependent allosteric decay.

B. Residual plots showing the relationship between ESM1v scores and experimentally measured ΔΔGf values for all PSD-95 PDZ3 mutations (left), orthosteric mutations (middle left), allosteric mutations (middle right), and all other mutations (right). LOESS regression lines (black) show the trend across all non-orthosteric mutations.

C. Distributions of the ESM1v–ΔΔGf residuals per variant group. Significance values were calculated using the two-sided Wilcoxon rank-sum test.

D. Relationship between the absolute median ESM1v–ΔΔGf residual and the minimal side chain distance to the ligand in PSD-95 PDZ3 (top) and GRB2 SH3 (bottom). Residuals were calculated from mutations with negative values only. An exponential decay model (black curve) was fitted to the data, with 95% bootstrap confidence intervals shown as grey shading. The grey dashed line marks the maximal minimal distance of orthosteric residues (i.e. the outer bound of direct ligand contacts). Residues are color-coded by functional class: orthosteric (orange) and non-orthosteric (dark green). Residues with significantly greater residuals than expected under the decay model (FDR-adjusted $p < 0.05$, one-sided Wilcoxon signed-rank test) are labeled by position number.

E. Crystal structures of PSD-95 PDZ3 bound to CRIPT (PDB: 1BE9) and GRB2-SH3 bound to GAB2 (PDB: 2VWF), colored by median ESM1v–ΔΔGf residuals for each site.

F. Linear regression models quantifying the variance in ESM1v scores explained by measured ΔΔGf (folding free energy changes), ΔΔGb (binding free energy changes), or both, across orthosteric and non-orthosteric sites. Adjusted R^2 values are shown with bootstrapped confidence intervals. Statistical significance was evaluated using ANOVA to compare each baseline model to the corresponding full model including both energy terms.

Figure 4

Clinically pathogenic variants in PTEN show allosteric decay.

A. Overview of the functional assays used to measure the effects of PTEN missense mutations on protein abundance and activity. Protein abundance was assessed using VAMP-seq, and phosphatase activity using activity-dependent yeast complementation. The residual between measured activity and abundance captures non-stability-mediated functional changes, and is used to assess distance-dependent allosteric decay.

B. Scatter plots of measured abundance versus activity fitness for all PTEN variants (left), gnomAD variants (middle), and ClinVar pathogenic variants (right). A LOESS regression curve (black line) models the relationship. Residuals from this fit quantify deviations from the abundance-activity trend. Pathogenic variants are marked with triangles, and a benign variant with a black circle. Dashed grey lines indicate wild-type abundance (0.71) and activity (-1.11) values, as determined in previous publications[60,61,71]. Two canonical

pathogenic mutations at active sites, C124R and R130G, are highlighted with black triangles.

C. Distributions of activity–abundance residuals for pathogenic and control variants. For each group, residuals reflect the deviation between observed activity and that predicted from abundance. Pathogenic variant residuals are compared to a bootstrapped null distribution generated from 1,000 abundance-matched samples of non-pathogenic variants (matching within 0.5-unit abundance bins). Violin plots show the distribution of median residuals, with medians indicated by black bars and diamonds. Sample sizes and median values are annotated. Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

D. Violin plots comparing activity-abundance residuals across variant sources (gnomAD, somatic, ClinVar pathogenic), separated into orthosteric (active/binding site) and non-orthosteric variants. Known dominant negative variants are labeled in cyan.

E. Violin plots showing the minimal Cα distance of each variant to TLA across variant classes.

F. Crystal structure of PTEN (PDB: 1D5R) colored by the median activity-abundance residual at each position. L(+)-tartaric acid (TLA), located near the catalytic site, is shown in cyan. A close-up of the P loop highlights catalytic residues with large residuals.

G. **Left:** per-residue relationship between absolute median activity-abundance residual and minimal distance to TLA for all 280 PTEN residues. An exponential decay model (black curve) was fit to the data, with 95% confidence intervals shown as grey shading. The vertical dashed line marks the maximum distance among annotated active site residues. Residues with significantly stronger effects than expected under the decay model (FDR-adjusted $p < 0.05$, one-sided Wilcoxon test) are labeled. **Right:** per-mutation relationship between absolute activity-abundance residual and distance for all variants with negative residuals. An exponential decay model (black curve) was fitted to the data, and 95% confidence intervals from bootstrapping are shown as grey shading. Dashed grey line indicates the maximal distance from any active-site residue to the ligand. Each point represents a single variant, colored by site type—active site (orange), membrane or calcium-binding site (cyan), or non-orthosteric site (dark green)—and triangle shaped/alpha-scaled by ClinVar pathogenicity classification. Labels mark pathogenic mutations at active or non-orthosteric sites with residuals that lie above the 95% confidence interval of the decay model, indicating stronger-than-expected divergence from abundance-predicted activity.

H. Distributions of activity–abundance residuals for pathogenic and control variants, stratified by whether variants are orthosteric (within active or binding sites) or non-orthosteric. Pathogenic variant residuals are compared to a bootstrapped null distribution generated from 1,000 abundance-matched samples of non-pathogenic, non-somatic variants (matching within 0.5-unit abundance bins). Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

Figure 5

Clinically pathogenic variants in GCK show allosteric decay.

A. Overview of the functional assays used to assess the effects of GCK missense mutations on protein abundance and activity. Protein abundance was measured using abundance PCA, and glucokinase activity was measured via a yeast complementation system. The residual between activity and abundance represents non-stability-mediated functional disruption and is used to assess distance-dependent allosteric effects.

B. Scatter plots of measured abundance versus activity fitness scores for all GCK variants (left), gnomAD variants (middle), and ClinVar pathogenic variants (right). A LOESS regression curve (black line) models the overall trend. Residuals from this fit capture deviations between abundance and activity. ClinVar pathogenic variants are shown as triangles, and benign variants as circles. Grey dashed lines indicate wild-type abundance (0.58) and activity (0.66) values, as determined in previous publications[37,62]. Canonical gain-of-function pathogenic variant (V452L) and loss-of-function pathogenic variant (T206M) are noted in black triangles.

C. Distributions of activity–abundance residuals for pathogenic and control variants. For each group, residuals reflect the deviation between observed activity and that predicted from abundance. Pathogenic variant residuals are compared to a bootstrapped null distribution generated from 1,000 abundance-matched samples of non-pathogenic variants (matching within 0.5-unit abundance bins). Violin plots show the distribution of median residuals, with medians indicated by black bars and diamonds. Sample sizes and median values are annotated. Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

D. Violin plots comparing activity-abundance residuals across gnomAD and ClinVar pathogenic variants, separated by whether the variant lies in orthosteric (ATP/glucose-binding) or non-orthosteric sites.

E. Violin plots comparing the minimal Cα distance to glucose across gnomAD and different ClinVar pathogenic variant classes (HH, MODY, monogenic diabetes).

F. Structure of GCK (PDB: 1V4S) colored by median activity-abundance residual per residue. Glucose is labeled in cyan. Key functional regions, including glucose-binding sites and helix 13 (residues 444–456), are highlighted in close-up views.

G. **Left:** per-residue analysis showing the relationship between absolute median activity-abundance residual and minimal distance to glucose across 445 GCK residues. An exponential decay model (black curve) was fit to the data, with 95% confidence intervals shown as grey shading. The vertical dashed line marks the maximum distance among annotated active site residues. Residues with significantly stronger effects than expected under the decay model (FDR-adjusted $p < 0.05$, one-sided Wilcoxon test) are labeled. **Right:** per-mutation analysis showing absolute residuals versus distance for

mutations with negative residuals. An exponential decay model (black curve) was fitted to the data, and 95% confidence intervals from bootstrapping are shown as grey shading. Dashed grey line indicates the maximal distance from any glucose-binding site residue to the ligand. Each point represents a single variant, colored by site type—glucose-binding site (orange), ATP-binding site (cyan), or non-orthosteric site (dark green)—and triangle shaped/alpha-scaled by ClinVar pathogenicity classification. Labels mark pathogenic mutations at glucose-binding or non-orthosteric sites with residuals that lie above the 95% confidence interval of the decay model, indicating stronger-than-expected divergence from abundance-predicted activity.

H. Distributions of activity–abundance residuals for pathogenic and control variants, stratified by whether variants are orthosteric (within active or binding sites) or non-orthosteric. Pathogenic variant residuals are compared to a bootstrapped null distribution generated from 1,000 abundance-matched samples of non-pathogenic variants (matching within 0.5-unit abundance bins). Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

Supplemental figure 1

Overview of data types, functional assays, and residual-based analysis strategies.

A. **Top panel (green):** Summary of protein-scale data used in this study. Pathogenicity scores were derived from either ESM1v or ClinVar annotations.

B. **Middle panel (blue):** Protein abundance was measured using ThermoMPNN-predicted folding free energy changes (ΔΔGf), aPCA/VAMP-seq-based fitness assays, or protease sensitivity. Binding free energy changes (ΔΔGb) was measured using bPCA in two protein domains.

C. **Middle panel (magenta):** Activity was quantified using yeast complementation assays for two full-length proteins, with abundance measured in parallel.

D. **Bottom left (blue):** Schematic of ESM1v–ΔΔGf residual analysis. Residuals quantify the divergence between predicted pathogenicity and measured folding effects. Residuals are then related to distance from the active or binding site to assess allosteric decay.

E. **Bottom right (magenta):** Schematic of activity-abundance residual analysis. The difference between measured activity and abundance (fitness) reflects functional disruption beyond stability effects. These residuals are also regressed against structural distance to active/binding sites to evaluate distance-dependent allosteric effects.

Supplemental figure 2

Protein class-specific contribution of stability to pathogenicity.

A. Bar plot showing the proportion of pathogenic ClinVar variants that are classified as stability-related (destabilizing or stabilizing), across protein functional categories.

B. Bar plot showing the proportion of benign ClinVar variants that are classified as stability-related, across protein functional categories.

C. Bar plot showing the proportion of pathogenic ClinVar variants in protein cores (excluding binding sites) that are classified as stability-related, across protein functional categories.

D. Bar plot showing the proportion of benign ClinVar variants in protein cores (excluding binding sites) that are classified as stability-related, across protein functional categories.

Supplemental figure 3

Distance-dependent allosteric decay across diverse protein types.

A. Crystal structures of CASP1 (PDB: 2FQQ) and IDH1 (PDB: 6IO0), colored by the median residuals between ESM1v and ThermoMPNN (TMPNN) predicted folding free energy changes (ΔΔGf) values. Bound substrates — PHD and NDP — are shown in orange. Allosteric inhibitor AOU is labeled in blue.

B. Scatter plots of ThermoMPNN-predicted ΔΔGf versus ESM1v scores for missense mutations in each protein. A LOESS curve (black line) models the trend, with residuals quantifying deviation between folding impact and predicted pathogenicity. Stable mutations cluster around ΔΔGf ≈ 0.

C. Per-residue relationship between the absolute median ESM1v–ΔΔGf residual and the minimal Cα distance to the substrates for each structure. Residuals are calculated using only mutations in non-orthosteric sites with negative residuals. An exponential decay model (black curve) was fitted to the data, with 95% confidence intervals shown as grey shading. The grey dashed line indicates the maximum of the minimal distances from any substrate-contacting (orthosteric) residue. Sites are colored by functional class: orthosteric sites in orange circles, non-orthosteric sites in dark green circles, and known allosteric sites (i.e., sites in physical contact with known allosteric regulators) in cyan triangles.

D. Distributions of the absolute median residuals between ESM1v and ΔΔGf per variant group. For the ‘other’ group, distributions are derived from 1,000 bootstrap replicates, each sampling the same number of non-orthosteric and non-allosteric sites as in the orthosteric group. Violin plots show the distribution shape; black bars and diamonds indicate medians. Sample sizes and median values are annotated. Statistical significance between groups was assessed using two-sided Wilcoxon rank-sum tests.

E. Distributions of the absolute median residuals between ESM1v and ΔΔGf per variant group across all six allosteric human proteins. The ‘other’ group represents the absolute residuals from all other non-orthosteric and non-allosteric sites. Significance values were calculated using the two-sided Wilcoxon rank-sum test.

Supplemental figure 4

Concordance between experimental and predicted residuals.

A. Density scatter plots showing the relationship between (left) ThermoMPNN-predicted ΔΔGf and experimentally measured abundance, and (right) ESM1v pathogenicity scores and experimentally measured activity for missense variants in GCK, PTEN, PSD-95-PDZ3, and GRB2-SH3. Point density is color-coded, and Spearman correlation coefficients (ρ) are shown for each plot.

B. Per-residue comparison between experimentally derived activity–abundance residuals and computationally derived ESM1v–ΔΔGf residuals in GCK, PTEN, PSD95-PDZ3, and GRB2-SH3. Each point represents a single residue, colored by site class. Significant spearman correlations are indicated in each panel.

Supplemental figure 5

Allosteric decay of binding effects.

A. Scatter plots showing the relationship between ESM1v pathogenicity scores and experimentally measured folding ΔΔGf for all 756 mutations in GRB2–SH3, with subsets shown for orthosteric (n = 171), allosteric (n = 53), and other (n = 532) mutation classes. Each point is colored by the ESM1v–ΔΔGf residual, and a LOESS curve (black line) models the overall trend. Example allosteric mutations are labeled.

B. Per-residue analysis of binding free energy changes (ΔΔGb) as a function of minimal side-chain distance to the ligand for PSD-95 PDZ3 (left) and GRB2–SH3 (right). Only mutations with ΔΔGb ≥ 0 were included. An exponential decay model (black curve) was fitted to the data, with 95% bootstrap confidence intervals shown as grey shading. The grey dashed line marks the maximal minimal distance of orthosteric residues (i.e. the outer bound of direct ligand contacts). Residues are color-coded by functional class: orthosteric (orange) and non-orthosteric (dark green). Residues with significantly greater ΔΔGb than expected under the decay model (FDR-adjusted $p < 0.05$, one-sided Wilcoxon signed-rank test) are labeled by position number.

C. Comparison of per-residue median ESM1v–ΔΔGf residuals with median ΔΔGb for PSD-95 PDZ3 (left, 84 residues) and GRB2–SH3 (right, 56 residues). Orthosteric residues are shown in orange and non-orthosteric residues in dark green. Spearman correlations are indicated for each plot.

Supplemental figure 6

PTEN and GCK analyses extended.

A. Violin plots showing activity–abundance residuals across variant groups in PTEN and GCK: gnomAD, and different ClinVar pathogenic categories. Residuals quantify deviation from the fitted activity–abundance relationship.

B. Violin plots comparing activity–abundance residuals among variants with wild-type–like abundance in PTEN and GCK, stratified by site type: orthosteric (left) and non-orthosteric (right).

Methods

Data processing for human proteome-wide predictions

Proteome-wide ESM1v scores were obtained from Zenodo (May 2025)[80], with detailed methods described in the accompanying publication[81]. AlphaFold predicted structures were downloaded from the AlphaFold Protein Structure Database (June 2025)[82] and filtered based on criteria established in previous publication[28], excluding proteins shorter than 16 amino acids or longer than 2700 amino acids. This yielded a final dataset of 20,144 unique, unfragmented human proteins, of which 19,804 had matching ESM1v scores and were retained for downstream analyses.

Stability changes upon mutation (ΔΔG) were predicted for all possible missense mutations using ThermoMPNN v1.0.0[51] with default settings applied to the AlphaFold-predicted structures.

Protein core positions were identified based on relative solvent-accessible surface area (rSASA), calculated using DSSP with BioPython with default settings[83]. Residues were classified as "core" if they were not predicted to be disordered or part of a binding site and had rSASA < 0.2, following criteria from previous publications[84,85]. Binding and active sites were annotated using proteome-wide predictions from the PIONEER interactome[84].

Proteins were grouped into functional classes according to the "Function and compartment based" classification provided by The Human Protein Atlas[76–78], available at https://www.proteinatlas.org/humanproteome/proteinclasses.

Clinical variant annotations were obtained from the supplementary materials of reference 28[28]. Variants were originally extracted from ClinVar[52,53,86] and processed according to established protocols[30], retaining only single-nucleotide variants with at least a one-star review status and mapped to GRCh38. For this study, variants labeled pathogenic or likely pathogenic were grouped as "pathogenic", while those labeled benign or likely benign were grouped as "benign". Variants with other labels or synonymous mutations were excluded. This yielded 71,418 benign and 42,579 pathogenic clinically annotated variants across 12,336 human proteins. ThermoMPNN-predicted missense mutations were classified as: destabilizing if ΔΔG > 1 kcal/mol, stabilizing if ΔΔG < –0.5 kcal/mol, following the thresholds defined in reference 51[51].

A subset of allosteric proteins, listed in **Supplementary Table 3**, were selected for focused analyses. These included all full-length human proteins present in two curated datasets: six proteins from the 20-protein benchmark set used to validate the allosteric network model Ohm[87], and one additional protein, isocitrate dehydrogenase 1 (IDH1), included from reference 87[88]. The enzyme prothrombin was excluded because its allosteric regulator functions as an activator

rather than an inhibitor. For glucokinase (GCK), annotations of allosteric sites were based on literature reports rather than crystal structure contacts, as its allosteric regulator (MRK) is also an activator. Orthosteric and allosteric residues were defined based on direct atomic contacts with bound substrates or modulators as observed in crystal structures. Residues were considered in contact if their atoms exhibited van der Waals overlap with atoms of the ligand, as determined using the ChimeraX VDW overlap method[89].

Data processing for Mega-scale, Domainome 1.0, and VAMP-seq datasets

ESM1v scores for variants in the Mega-scale dataset were obtained from the ProteinGym database[90]. Of the 331 natural protein domains studied, 64 had corresponding ESM1v predictions available. For the Domainome 1.0 dataset, ESM1v scores were extracted from the supplementary materials of the original publication[32]. ESM1v scores for VAMP-seq variants were derived from the proteome-wide ESM1v dataset described above. VAMP-seq abundance fitness data were collected for nine full-length human proteins: PRKN, OCT, CYP2C9, TPMT, ASPA, NUDT15, VKOR, PTEN, and GCK[37,60,91–96].

Clinical variant annotations for both the Domainome 1.0 and VAMP-seq proteins were obtained from the supplementary data of reference 28[28], following the same filtering criteria as described above.

Variant Classification

- Domainome 1.0 variants were classified based on criteria defined in the original publication. Mutations were considered stabilizing if they had a statistically significant stabilizing effect (*FDR_stabilizing* < 0.1) and a scaled fitness > 0.3. Conversely, mutations were classified as destabilizing if they had a significant destabilizing effect (*FDR_destabilizing* < 0.1) and a scaled fitness < 0.

- VAMP-seq variants were categorized based on abundance thresholds defined by ProteinGym[90]. Mutations falling below the defined abundance cutoff were considered destabilizing.

- Mega-scale variants were classified using thresholds established in the original study [31]. Note that the ΔΔG directionality in this dataset is inverted relative to others: mutations with ΔΔG > 1 kcal/mol were classified as stabilizing, while those with ΔΔG < –1 kcal/mol were classified as destabilizing.

Data processing for experimental allosteric maps

Analyses for PSD-95-PDZ3 and GRB2-SH3 were performed using variants with corresponding ESM1v scores, obtained from the human proteome-wide prediction dataset. Orthosteric sites were defined based on published structural annotations as residues within 5 Å of bound ligands[39]. Variants were classified as allosteric if they occurred outside orthosteric sites and

exhibited binding ΔΔG (ΔΔGb) values greater than or equal to the mean ΔΔGb of orthosteric-site mutations.

For PTEN and GCK, analyses were based on variants with both abundance and activity fitness measurements. In PTEN, orthosteric site annotations were manually curated from the literature[97–103]. These included the catalytic loops (arginine loop, residues 35–49; WPD loop, 88–98; P loop, 123–130; TI loop, 159–171), calcium-binding regions (residues 200–212, 226–238, and 258–268), the Cα2 loop (residues 327–335), the membrane-binding helix (residues 151–174), and the PI(4,5)P_2-binding motif (PBM, residues 1–15).

In GCK, orthosteric sites were defined based on annotated ligand-contacting residues described in previous structural studies[104–107]. These included glucose-binding residues (positions 151–153, 168–169, 204–206, 225–231, 254–258, 287, and 290) and ATP-binding residues (positions 78–85, 295–296, 331–333, 336, and 410–416).

Variants from gnomAD and ClinVar were downloaded from their respective online portals in April 2025[86,108]. Somatic variant annotations were obtained from the supplementary materials of a previous publication and filtered to retain only cancer hotspot mutations[71]. These hotspots were defined using the “curated set of non-redundant studies” available on cBioPortal[109,110].

Statistical analysis and data visualization

Linear regression models were used to analyze proteome-wide variant effect predictions, large-scale abundance datasets, and individual protein domains with experimentally mapped allosteric sites. To evaluate the relative contributions of folding ΔΔG and binding ΔΔG to ESM1v scores, both ordinary linear regression and linear mixed-effects models were employed. Linear regression was performed using the `lm()` function from the stats package in R, while mixed-effects models were fitted using the `lmer()` function from the lme4 package[111,112]. As results from both approaches were highly consistent (Supplementary Tables 4–7), linear regression outcomes are reported for interpretability. Adjusted R^2 values were extracted directly from the `summary()` output of the fitted linear models in R. To estimate variability and provide robust confidence intervals, adjusted R^2 was also bootstrapped using 1,000 resampling iterations. Model diagnostics included residual and prediction analysis, and model comparisons were performed using ANOVA, with F-statistics and p-values reported.

LOESS residual analyses were performed using only mutations located outside orthosteric sites. LOESS fitting was conducted with the `loess()` function in R, using a span of 0.7 and the family = "symmetric" setting. Distances to ligand-binding sites were computed based on Cα atom coordinates for all proteins, except for PSD-95-PDZ3 and GRB2-SH3, where distances were measured using side chain heavy atoms.

To generate abundance-matched control distributions for residual activity, we performed bootstrap resampling from non-pathogenic variants. For each protein (PTEN and GCK), pathogenic variants with measured abundance and residual activity scores were first binned by their abundance values using fixed-width bins (bin size = 0.5). For each pathogenic variant, we identified all non-pathogenic variants in the same abundance bin (excluding any known

pathogenic variants). From these bin-matched pools, one variant was sampled randomly with replacement for each pathogenic variant, and the median residual activity across the matched sample was recorded. This process was repeated 1000 times to generate a distribution of bootstrapped median residuals. The resulting distributions were compared using a two-sided Wilcoxon rank-sum test. Violin plots show the distribution of residuals in pathogenic variants and the distribution of bootstrap medians from abundance-matched controls.

To quantify the distance-dependent decay of mutational effects, nonlinear least-squares regression was applied to the LOESS residuals. The allosteric decay model included two parameters: the decay rate (−b) and the x-intercept (a). This fitting was performed on residuals from either ESM1v versus folding ΔΔG (ΔΔGf) or activity versus abundance fitness datasets, using only mutations with negative residuals that fell below the LOESS curve. Nonlinear regression was conducted with the `nls()` function in R, using initial parameter values of a = 1 and b = 0.1. 95% confidence intervals were calculated by bootstrapping the model parameter with 1,000 replicates.

To identify residues with significantly stronger mutational impacts than predicted by the distance-dependent decay model, we computed residuals between observed and expected values per residue and applied a one-sided Wilcoxon signed-rank test (alternative = "greater"). P-values were adjusted using the Benjamini-Hochberg FDR method. Residues with adjusted $p < 0.05$ were defined as allosteric "hotspots."

All statistical analyses were performed in R[112]. Experimental schematics were created using BioRender.com, while all other plots were generated using the ggplot2 package[113]. Three-dimensional protein structure visualizations were rendered using ChimeraX v1.6.1[89].

Acknowledgment

We thank Taylor Mighell, Mike Thompson, Maximilian R. Stammnitz, Albert Escobedo, Taraneh Zarin, Antoni Beltran, Xianghua Li, and other former and current members of the Lehner lab for feedback and discussions. We also thank the Sanger HumGen/GenGen Informatics for their assistance with the computational infrastructure.

Work in the lab of B.L. was funded by Wellcome (Grant reference: 220540/Z/20/A, 'Wellcome Sanger Institute Quinquennial Review 2021-2026'), a European Research Council (ERC) Advanced (883742) grant, the Spanish Ministry of Science and Innovation (LCF/PR/HR21/52410004, EMBL Partnership, Severo Ochoa Centre of Excellence), Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, 2021 SGR 01226), and the CERCA Program/Generalitat de Catalunya. X.L is funded by the Wellcome Sanger Institute PhD Graduate program.

Data availability

Proteome-wide predictions of protein stability changes and corresponding ESM1v scores are provided in Supplementary Table 1 and are available on Zenodo: https://zenodo.org/uploads/15681969. Raw ThermoMPNN predictions of protein stability changes can be accessed separately at https://zenodo.org/uploads/15586553. Large-scale experimental measurements of protein abundance changes across human domains, including datasets from the Mega-scale and Domainome studies, are accessible via their respective supplementary materials. VAMP-seq abundance data can be obtained from MaveDB and ProteinGym[90,114]. Experimental ΔΔG values for PSD95-PDZ3 and GRB2-SH3 domains are publicly available, with full methodologies described in previous publications[39,115,116]. Experimental measurements of abundance and activity-based fitness for PTEN and GCK can be accessed through the original studies[37,60–62].

Code availability

All code and metadata to reproduce the analyses are hosted on GitHub: https://github.com/lehner-lab/allostery_pathogenicity.

Author contributions

X.L. performed all analyses. X.L. and B.L. conceived the project, designed analyses, and wrote the manuscript.

Declaration of interests

B.L. is a founder and shareholder of ALLOX.

References

1. Fenton, A. W. Allostery: an illustrated definition for the ‘second secret of life’. *Trends Biochem. Sci.* **33**, 420–425 (2008).
2. Guarnera, E. & Berezovsky, I. N. Allosteric drugs and mutations: chances, challenges, and necessity. *Curr. Opin. Struct. Biol.* **62**, 149–157 (2020).
3. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).
4. Smith, J. S., Lefkowitz, R. J. & Rajagopal, S. Biased signalling: from simple switches to allosteric microprocessors. *Nat. Rev. Drug Discov.* **17**, 243–260 (2018).
5. Wu, N., Barahona, M. & Yaliraki, S. N. Allosteric communication and signal transduction in proteins. *Curr. Opin. Struct. Biol.* **84**, 102737 (2024).
6. Nussinov, R. & Tsai, C.-J. Allostery in disease and in drug discovery. *Cell* **153**, 293–305 (2013).
7. Goodey, N. M. & Benkovic, S. J. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* **4**, 474–482 (2008).
8. Gunasekaran, K., Ma, B. & Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**, 433–443 (2004).
9. Tsai, C.-J. & Nussinov, R. A unified view of ‘how allostery works’. *PLoS Comput. Biol.* **10**, e1003394 (2014).
10. Sivcev, S., Kudova, E. & Zemkova, H. Neurosteroids as positive and negative allosteric modulators of ligand-gated ion channels: P2X receptor perspective. *Neuropharmacology* **234**, 109542 (2023).
11. Keov, P., Sexton, P. M. & Christopoulos, A. Allosteric modulation of G protein-coupled receptors: a pharmacological perspective. *Neuropharmacology* **60**, 24–35 (2011).
12. Wu, P., Clausen, M. H. & Nielsen, T. E. Allosteric small-molecule kinase inhibitors.

Pharmacol. Ther. **156**, 59–68 (2015).

13. Yu, Y., Yu, Q. & Zhang, X. Allosteric inhibition of HIF-2α as a novel therapy for clear cell renal cell carcinoma. *Drug Discov. Today* **24**, 2332–2340 (2019).
14. Mighell, T. & Lehner, B. A pharmacological chaperone stabilizer rescues the expression of the vast majority of pathogenic variants in a G protein-coupled receptor. *bioRxiv* (2024) doi:10.1101/2024.11.28.625821.
15. Sheik Amamuddy, O. *et al.* Integrated Computational Approaches and Tools forAllosteric Drug Discovery. *Int J Mol Sci* **21**, (2020).
16. Nussinov, R., Jang, H., Tsai, C.-J. & Cheng, F. Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLoS Comput. Biol.* **15**, e1006658 (2019).
17. Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548–551 (2013).
18. Kwan, A. K., Piazza, G. A., Keeton, A. B. & Leite, C. A. The path to the clinic: a comprehensive review on direct KRASG12C inhibitors. *J. Exp. Clin. Cancer Res.* **41**, 27 (2022).
19. Bandaranayake, R. M. *et al.* Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nat. Struct. Mol. Biol.* **19**, 754–759 (2012).
20. Hubbard, S. R. Mechanistic insights into regulation of JAK2 tyrosine kinase. *Front. Endocrinol. (Lausanne)* **8**, 361 (2017).
21. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169–181 (2007).
22. Yun, C.-H. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2070–2075 (2008).
23. Tee, W.-V. & Berezovsky, I. N. Allosteric drugs: New principles and design approaches. *Curr Opin Struct Biol* **84**, 102758 (2024).

24. Zhang, H. *et al.* Exploring pathogenic mutation in allosteric proteins: The prediction and beyond. *bioRxiv* (2024) doi:10.1101/2024.03.23.586438.
25. Chen, B. & Altman, R. B. Opportunities for developing therapies for rare genetic diseases: focus on gain-of-function and allostery. *Orphanet J. Rare Dis.* **12**, (2017).
26. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
27. Abramowicz, A. & Gos, M. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* **59**, 253–268 (2018).
28. Cagiada, M., Jonsson, N. & Lindorff-Larsen, K. Decoding molecular mechanisms for loss of function variants in the human proteome. *bioRxiv* (2024) doi:10.1101/2024.05.21.595203.
29. Jänes, J. *et al.* Predicted mechanistic impacts of human protein missense variants. *bioRxivorg* (2024) doi:10.1101/2024.05.29.596373.
30. Tiemann, J. K. S., Zschach, H., Lindorff-Larsen, K. & Stein, A. Interpreting the molecular mechanisms of disease variants in human transmembrane proteins. *Biophys J* **122**, 2176–2191 (2023).
31. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
32. Beltran, A., Jiang, X. ’er, Shen, Y. & Lehner, B. Site-saturation mutagenesis of 500 human protein domains. *Nature* 1–10 (2025).
33. Backwell, L. & Marsh, J. A. Diverse molecular mechanisms underlying pathogenic protein mutations: Beyond the loss-of-function paradigm. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–498 (2022).
34. Taipale, M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms 1. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **97**, (2019).
35. Notbohm, J. & Perica, T. Biochemistry and genetics are coming together to improve our

understanding of genotype to phenotype relationships. *Curr Opin Struct Biol* **89**, 102952 (2024).

36. 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).
37. Gersing, S. *et al.* Characterizing glucokinase variant mechanisms using a multiplexed abundance assay. *Genome Biol.* **25**, 98 (2024).
38. 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).
39. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
40. Beltran, A., Faure, A. J. & Lehner, B. The allosteric landscape of the Src kinase. *bioRxiv* 2024.04.26.591297 (2024) doi:10.1101/2024.04.26.591297.
41. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).
42. Nussinov, R. *et al.* Allostery: Allosteric cancer drivers and innovative allosteric drugs. *J. Mol. Biol.* **434**, 167569 (2022).
43. Tee, W.-V., Guarnera, E. & Berezovsky, I. N. On the Allosteric Effect of nsSNPs and the Emerging Importance of Allosteric Polymorphism. *J Mol Biol* **431**, 3933–3942 (2019).
44. Csanády, L., Vergani, P. & Gadsby, D. C. Structure, gating, and regulation of the cftr anion channel. *Physiol. Rev.* **99**, 707–738 (2019).
45. Wang, W. *et al.* G551D mutation impairs PKA-dependent activation of CFTR channel that can be restored by novel GOF mutations. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **319**, L770–L785 (2020).
46. Lin, W.-Y., Jih, K.-Y. & Hwang, T.-C. A single amino acid substitution in CFTR converts ATP

to an inhibitory ligand. *J. Gen. Physiol.* **144**, 311–320 (2014).

47. Ariyananda, L. D. Z., Lee, P., Antonopoulos, C. & Colman, R. F. Biochemical and biophysical analysis of five disease-associated human adenylosuccinate lyase mutants. *Biochemistry* **48**, 5291–5302 (2009).
48. Stammnitz, M. R. & Lehner, B. The genetic architecture of an allosteric hormone receptor. *bioRxiv* 2025.05.30.656975 (2025) doi:10.1101/2025.05.30.656975.
49. Mighell, T. L. & Lehner, B. The genetic architecture of G-protein coupled receptor signaling. *bioRxiv* 2025.05.30.656974 (2025) doi:10.1101/2025.05.30.656974.
50. Escobedo, A., Voigt, G., Faure, A. J. & Lehner, B. Genetics, energetics and allostery during a billion years of hydrophobic protein core evolution. *bioRxiv* (2024) doi:10.1101/2024.05.11.593672.
51. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences* **121**, e2314853121 (2024).
52. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862–8 (2016).
53. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
54. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* (2021) doi:10.1101/2021.07.09.450648.
55. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
56. Rajasekaran, N., Sekhar, A. & Naganathan, A. N. A Universal Pattern in the Percolation and Dissipation of Protein Structural Perturbations. *J. Phys. Chem. Lett.* **8**, 4779–4784 (2017).
57. Rajasekaran, N., Suresh, S., Gopi, S., Raman, K. & Naganathan, A. N. A general

mechanism for the propagation of mutational effects in proteins. *Biochemistry* **56**, 294–305 (2017).

58. Rettenmaier, T. J. *et al.* A small-molecule mimic of a peptide docking motif inhibits the protein kinase PDK1. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18590–18595 (2014).
59. Wiesmann, C. *et al.* Allosteric inhibition of protein tyrosine phosphatase 1B. *Nature Structural & Molecular Biology* **11**, 730–737 (2004).
60. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
61. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
62. Gersing, S. *et al.* A comprehensive map of human glucokinase variant activity. *Genome Biol.* **24**, 97 (2023).
63. Pilarski, R. PTEN hamartoma tumor syndrome: A clinical overview. *Cancers (Basel)* **11**, 844 (2019).
64. 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).
65. Bonneau, D. & Longy, M. Mutations of the human PTEN gene. *Hum. Mutat.* **16**, 109–122 (2000).
66. Froguel, P. *et al.* Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356**, 162–164 (1992).
67. Das, S. K. *et al.* Polymorphisms in the glucokinase-associated, dual-specificity phosphatase 12 (DUSP12) gene under chromosome 1q21 linkage peak are associated with type 2 diabetes. *Diabetes* **55**, 2631–2639 (2006).
68. Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I. & Nagata, Y. Structural basis for allosteric

regulation of the monomeric allosteric enzyme human glucokinase. *Structure* **12**, 429–438 (2004).

69. Larion, M. & Miller, B. G. 23-Residue C-terminal alpha-helix governs kinetic cooperativity in monomeric human glucokinase. *Biochemistry* **48**, 6157–6165 (2009).
70. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
71. Matreyek, K. A., Stephany, J. J., Ahler, E. & Fowler, D. M. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med.* **13**, 165 (2021).
72. Laskowski, R. A., Gerick, F. & Thornton, J. M. The structural basis of allosteric regulation in proteins. *FEBS Lett.* **583**, 1692–1698 (2009).
73. Hilser, V. J., Wrabl, J. O. & Motlagh, H. N. Structural and energetic basis of allostery. *Annu. Rev. Biophys.* **41**, 585–609 (2012).
74. Tzeng, S.-R. & Kalodimos, C. G. Protein dynamics and allostery: an NMR view. *Curr. Opin. Struct. Biol.* **21**, 62–67 (2011).
75. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol.* **14**, e1002452 (2016).
76. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
77. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).
78. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, (2017).
79. Beltran, A., Jiang, X. ’er, Shen, Y. & Lehner, B. Site saturation mutagenesis of 500 human protein domains reveals the contribution of protein destabilization to genetic disease. *bioRxiv* (2024) doi:10.1101/2024.04.26.591310.
80. Livesey, B. ESM-1v predictions for human uniprot IDs. Zenodo https://doi.org/10.5281/ZENODO.8301668 (2023).

81. Livesey, B. J. & Marsh, J. A. Variant effect predictor correlation with functional assays is reflective of clinical classification performance. *bioRxiv* (2024) doi:10.1101/2024.05.12.593741.

82. 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).

83. Rost, B. & Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216–226 (1994).

84. Xiong, D. *et al.* A structurally informed human protein-protein interactome reveals proteome-wide perturbations caused by disease mutations. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02428-4.

85. Hanson, J., Yang, Y., Paliwal, K. & Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685–692 (2017).

86. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).

87. Wang, J. *et al.* Mapping allosteric communications within individual proteins. *Nat. Commun.* **11**, 3862 (2020).

88. Kannan, G. R., Hie, B. L. & Kim, P. S. Single-sequence, structure free allosteric residue prediction with protein language models. *bioRxiv* (2024) doi:10.1101/2024.10.03.616547.

89. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).

90. Notin, P. *et al.* ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* (2023).

91. Clausen, L. *et al.* PRKN-linked familial Parkinson's disease: cellular and molecular mechanisms of disease-linked variants. *Cell. Mol. Life Sci.* **81**, 223 (2024).

92. Yee, S. W. *et al.* The full spectrum of OCT1 (SLC22A1) mutations bridges transporter biophysics to drug pharmacogenomics. *bioRxiv* (2023) doi:10.1101/2023.06.06.543963.

93. Amorosi, C. J. *et al.* Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).

94. Grønbæk-Thygesen, M. *et al.* Deep mutational scanning reveals a correlation between degradation and toxicity of thousands of aspartoacylase variants. *Nat. Commun.* **15**, 4026 (2024).

95. Suiter, C. C. *et al.* Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5394–5401 (2020).

96. Chiasson, M. A. *et al.* Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**, (2020).

97. Masson, G. R. & Williams, R. L. Structural mechanisms of PTEN regulation. *Cold Spring Harb. Perspect. Med.* **10**, a036152 (2020).

98. Shenoy, S. S., Nanda, H. & Lösche, M. Membrane association of the PTEN tumor suppressor: electrostatic interaction with phosphatidylserine-containing bilayers and regulatory role of the C-terminal tail. *J Struct Biol* **180**, 394–408 (2012).

99. Nanda, H., Heinrich, F. & Lösche, M. Membrane association of the PTEN tumor suppressor: neutron scattering and MD simulations reveal the structure of protein-membrane complexes. *Methods* **77-78**, 136–146 (2015).

100. Thiriet, M. *Intracellular Signaling Mediators in the Circulatory and Ventilatory Systems*. (Springer Science & Business Media, 2012).

101. Irvine, W. A., Flanagan, J. U. & Allison, J. R. Computational Prediction of Amino Acids Governing Protein-Membrane Interaction for the PIP Cell Signaling System. *Structure* **27**, 371–380.e3 (2019).

102. Lee, J. O. *et al.* Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell* **99**, 323–334

(1999).

103. Serebriiskii, I. G. *et al.* Comprehensive characterization of PTEN mutational profile in a series of 34,129 colorectal cancers. *Nat Commun* **13**, 1618 (2022).

104. DiStefano, M. T. *et al.* Curating Clinically Relevant Transcripts for the Interpretation of Sequence Variants. *J Mol Diagn* **20**, 789–801 (2018).

105. Beck, T. & Miller, B. G. Structural basis for regulation of human glucokinase by glucokinase regulatory protein. *Biochemistry* **52**, 6232–6239 (2013).

106. Gloyn, A. L. *et al.* Insights into the structure and regulation of glucokinase from a novel mutation (V62M), which causes maturity-onset diabetes of the young. *J Biol Chem* **280**, 14105–14113 (2005).

107. Beer, N. L. *et al.* Insights into the pathogenicity of rare missense GCK variants from the identification and functional characterization of compound heterozygous and double mutations inherited in cis. *Diabetes Care* **35**, 1482–1484 (2012).

108. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

109. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, l1 (2013).

110. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

111. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* **67**, (2015).

112. Team, R. D. C. & Others. R: A language and environment for statistical computing. *R Found. Stat. Comput.* **1**, 409 (2016).

113. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).

114. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from

multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).

115. Faure, A. J. & Lehner, B. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biol* **25**, 303 (2024).

116. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol* **21**, 207 (2020).

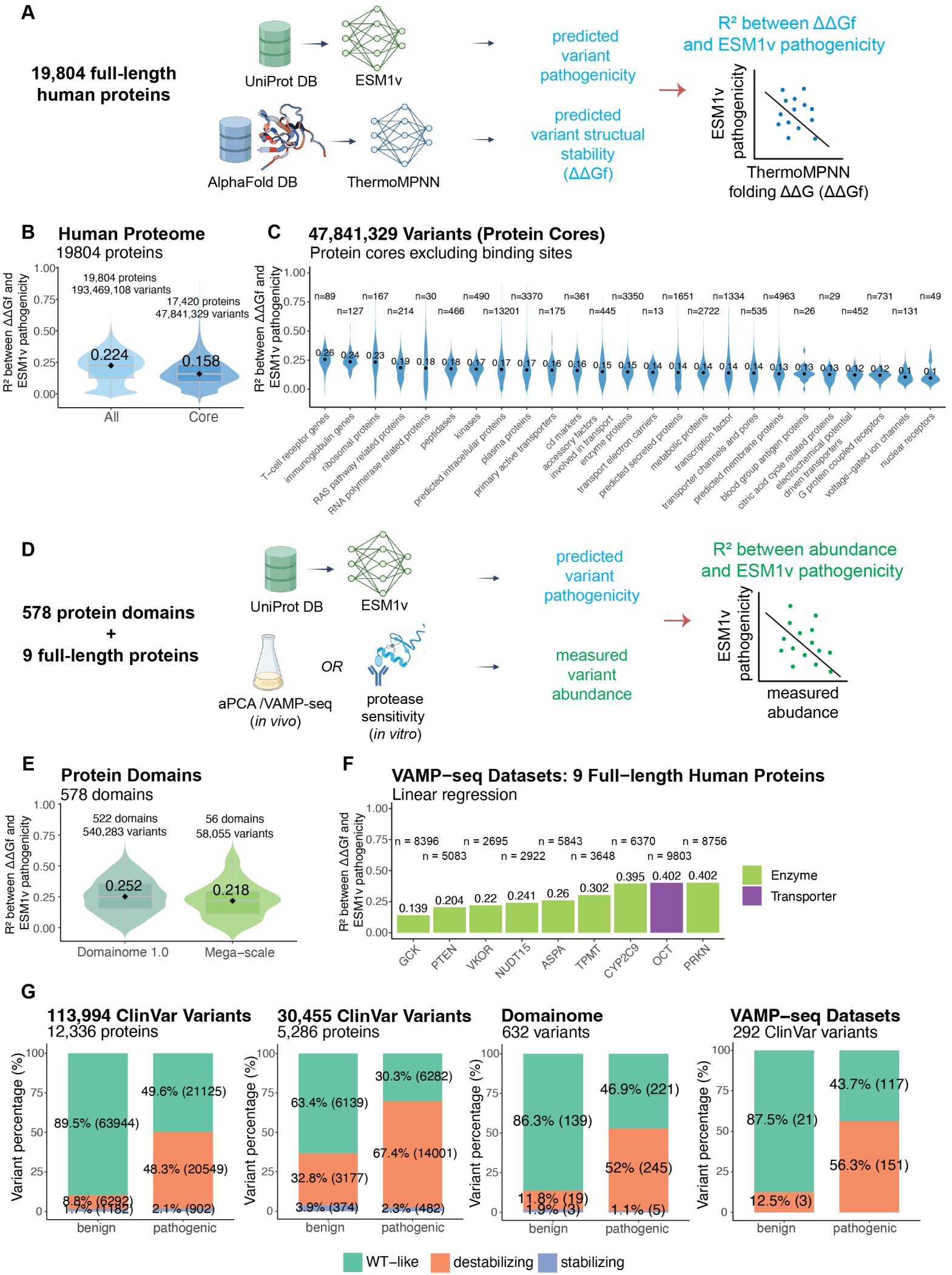

A
19,804 full-length human proteins
UniProt DB
ESM1v
AlphaFold DB
ThermoMPNN
predicted variant pathogenicity
predicted variant structual stability (ΔΔGf)
R² between ΔΔGf and ESM1v pathogenicity
ESM1v pathogenicity
ThermoMPNN folding ΔΔG (ΔΔGf)
B
Human Proteome
19804 proteins
19,804 proteins 193,469,108 variants
17,420 proteins 47,841,329 variants
0.224
0.158
All
Core
R² between ΔΔGf and ESM1v pathogenicity
C
47,841,329 Variants (Protein Cores)
Protein cores excluding binding sites
T-cell receptor genes
immunoglobulin genes
ribosomal proteins
RAS pathway related proteins
RNA polymerase related proteins
peptidases
kinases
predicted intracellular proteins
plasma proteins
primary active transporters
cd markers
accessory factors involved in transport
enzyme proteins
transport electron carriers
predicted secreted proteins
metabolic proteins
transcription factor
transporter channels and pores
predicted membrane proteins
blood group antigen proteins
citric acid cycle related proteins
electrochemical potential driven transporters
G protein coupled receptors
voltage-gated ion channels
nuclear receptors
D
578 protein domains + 9 full-length proteins
UniProt DB
ESM1v
aPCA /VAMP-seq (in vivo)
OR
protease sensitivity (in vitro)
predicted variant pathogenicity
measured variant abundance
R² between abundance and ESM1v pathogenicity
ESM1v pathogenicity
measured abudance
E
Protein Domains
578 domains
522 domains 540,283 variants
56 domains 58,055 variants
0.252
0.218
Domainome 1.0
Mega-scale
F
VAMP-seq Datasets: 9 Full-length Human Proteins
Linear regression
GCK
PTEN
VKOR
NUDT15
ASPA
TPMT
CYP2C9
OCT
PRKN
Enzyme
Transporter
G
113,994 ClinVar Variants
12,336 proteins
30,455 ClinVar Variants
5,286 proteins
Domainome
632 variants
VAMP-seq Datasets
292 ClinVar variants
Variant percentage (%)
benign
pathogenic
WT-like
destabilizing
stabilizing

Fig 1

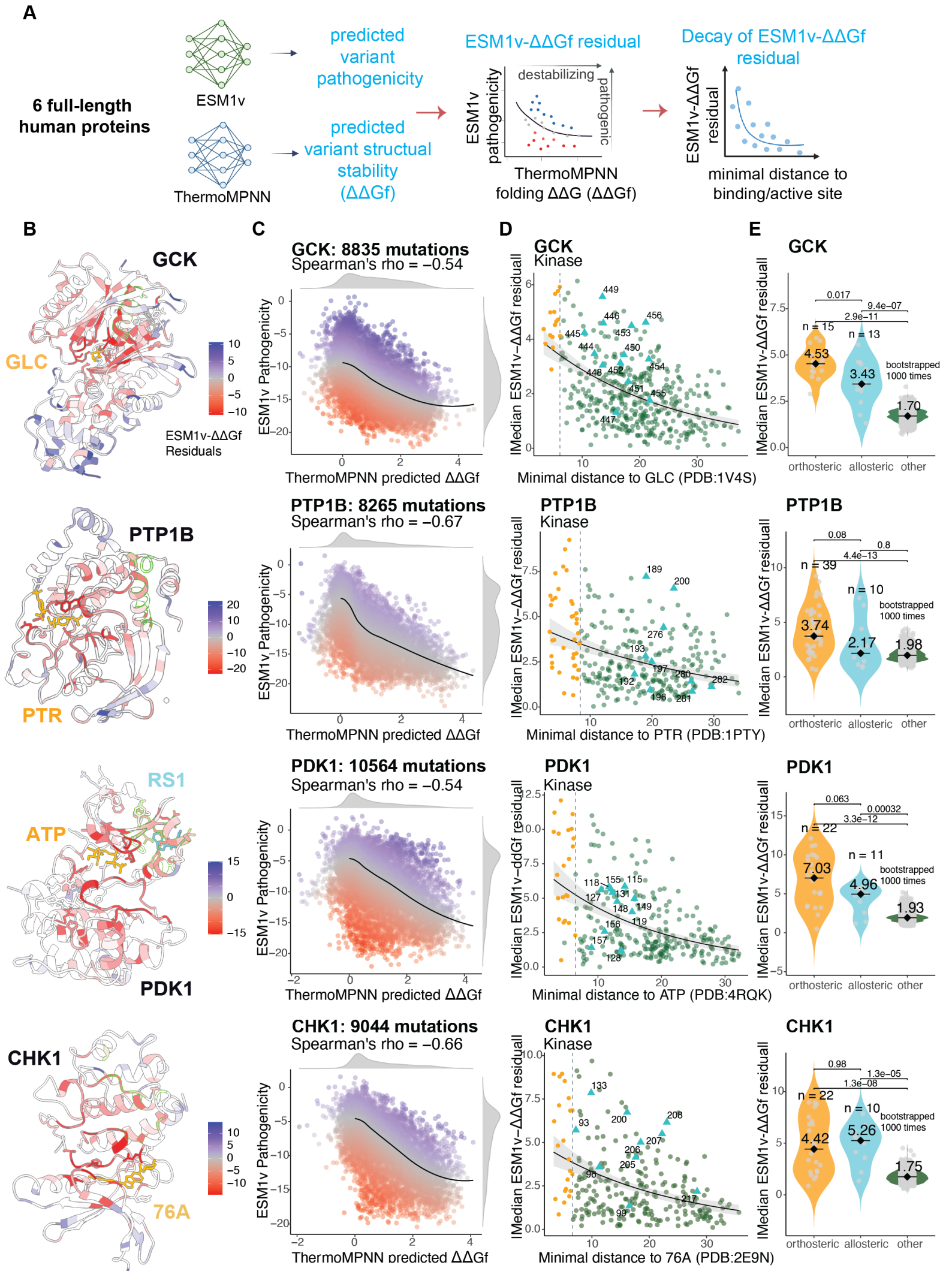

A
6 full-length human proteins
ESM1v
ThermoMPNN
predicted variant pathogenicity
predicted variant structual stability (ΔΔGf)
ESM1v-ΔΔGf residual
destabilizing
pathogenic
ESM1v pathogenicity
ThermoMPNN folding ΔΔG (ΔΔGf)
Decay of ESM1v-ΔΔGf residual
ESM1v-ΔΔGf residual
minimal distance to binding/active site
B
GCK
GLC
ESM1v-ΔΔGf Residuals
PTP1B
PTR
RS1
ATP
PDK1
CHK1
76A
C
GCK: 8835 mutations
Spearman's rho = −0.54
PTP1B: 8265 mutations
Spearman's rho = −0.67
PDK1: 10564 mutations
Spearman's rho = −0.54
CHK1: 9044 mutations
Spearman's rho = −0.66
ESM1v Pathogenicity
ThermoMPNN predicted ΔΔGf
D
GCK
Kinase
PTP1B
Kinase
PDK1
Kinase
CHK1
Kinase
IMedian ESM1v−ΔΔGf residual
IMedian ESM1v−ddGf residual
Minimal distance to GLC (PDB:1V4S)
Minimal distance to PTR (PDB:1PTY)
Minimal distance to ATP (PDB:4RQK)
Minimal distance to 76A (PDB:2E9N)
E
GCK
PTP1B
PDK1
CHK1
IMedian ESM1v-ΔΔGf residual
n = 15
n = 13
4.53
3.43
1.70
0.017
9.4e−07
2.9e−11
n = 39
n = 10
3.74
2.17
1.98
0.08
0.8
4.4e−13
n = 22
n = 11
7.03
4.96
1.93
0.063
0.00032
3.3e−12
n = 22
n = 10
4.42
5.26
1.75
0.98
1.3e−05
1.3e−08
bootstrapped 1000 times
orthosteric
allosteric
other

Fig 2

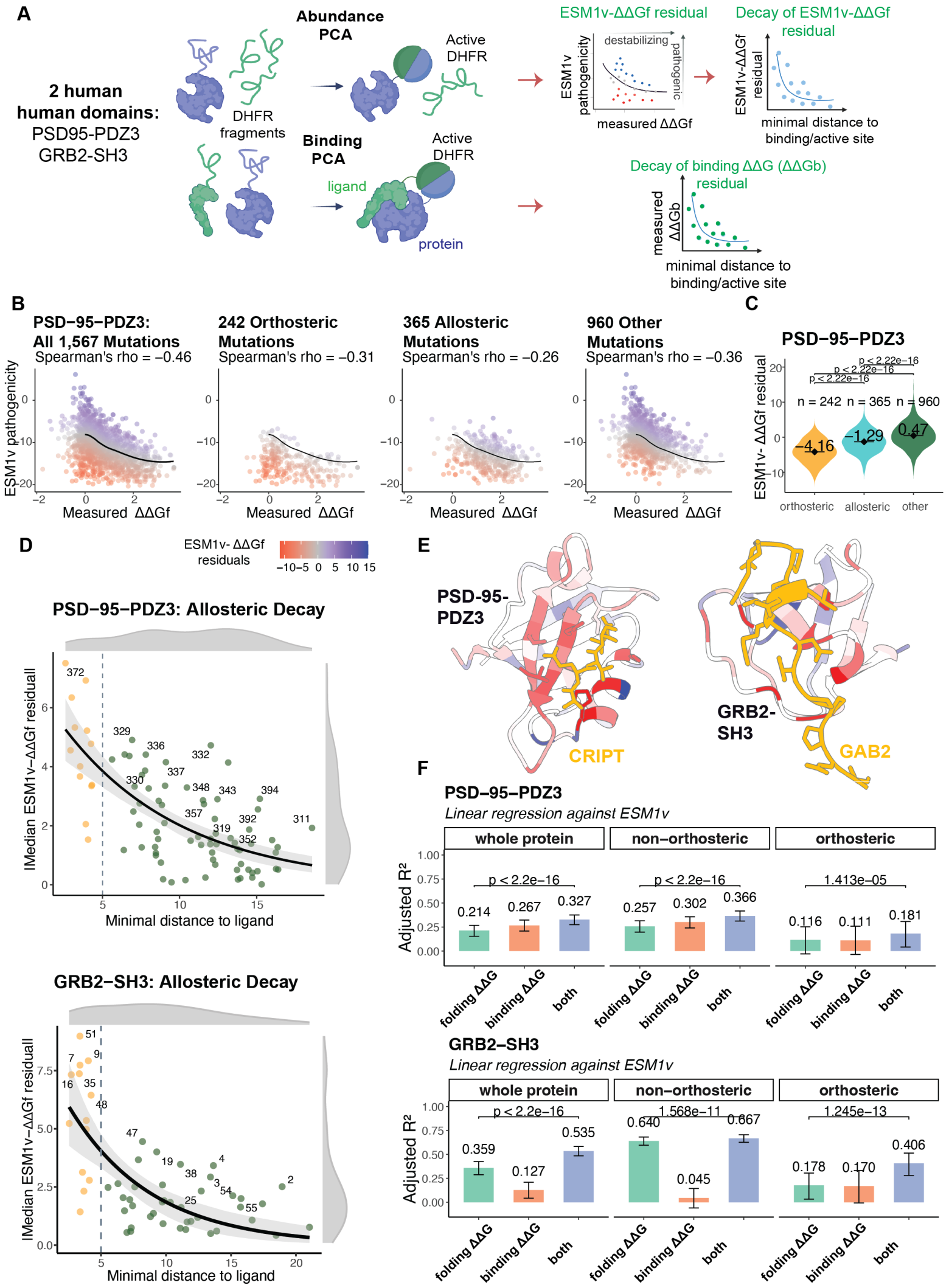

A
2 human human domains:
PSD95-PDZ3
GRB2-SH3
DHFR fragments
Abundance PCA
Active DHFR
Binding PCA
ligand
protein
ESM1v-ΔΔGf residual
destabilizing
pathogenic
ESM1v pathogenicity
measured ΔΔGf
Decay of ESM1v-ΔΔGf residual
ESM1v-ΔΔGf residual
minimal distance to binding/active site
Decay of binding ΔΔG (ΔΔGb) residual
measured ΔΔGb
minimal distance to binding/active site
B
PSD−95−PDZ3: All 1,567 Mutations
Spearman's rho = −0.46
242 Orthosteric Mutations
Spearman's rho = −0.31
365 Allosteric Mutations
Spearman's rho = −0.26
960 Other Mutations
Spearman's rho = −0.36
ESM1v pathogenicity
Measured ΔΔGf
C
PSD−95−PDZ3
ESM1v- ΔΔGf residual
p < 2.22e−16
n = 242
n = 365
n = 960
−4.16
−1.29
0.47
orthosteric
allosteric
other
D
ESM1v- ΔΔGf residuals
−10 −5 0 5 10 15
PSD−95−PDZ3: Allosteric Decay
IMedian ESM1v−ΔΔGf residuall
Minimal distance to ligand
GRB2−SH3: Allosteric Decay
IMedian ESM1v−ΔΔGf residuall
Minimal distance to ligand
E
PSD-95-PDZ3
CRIPT
GRB2-SH3
GAB2
F
PSD−95−PDZ3
Linear regression against ESM1v
whole protein
non−orthosteric
orthosteric
Adjusted R²
p < 2.2e−16
p < 2.2e−16
1.413e−05
0.214
0.267
0.327
0.257
0.302
0.366
0.116
0.111
0.181
folding ΔΔG
binding ΔΔG
both
GRB2−SH3
Linear regression against ESM1v
whole protein
non−orthosteric
orthosteric
Adjusted R²
p < 2.2e−16
1.568e−11
1.245e−13
0.359
0.127
0.535
0.640
0.045
0.667
0.178
0.170
0.406
folding ΔΔG
binding ΔΔG
both

Fig 3

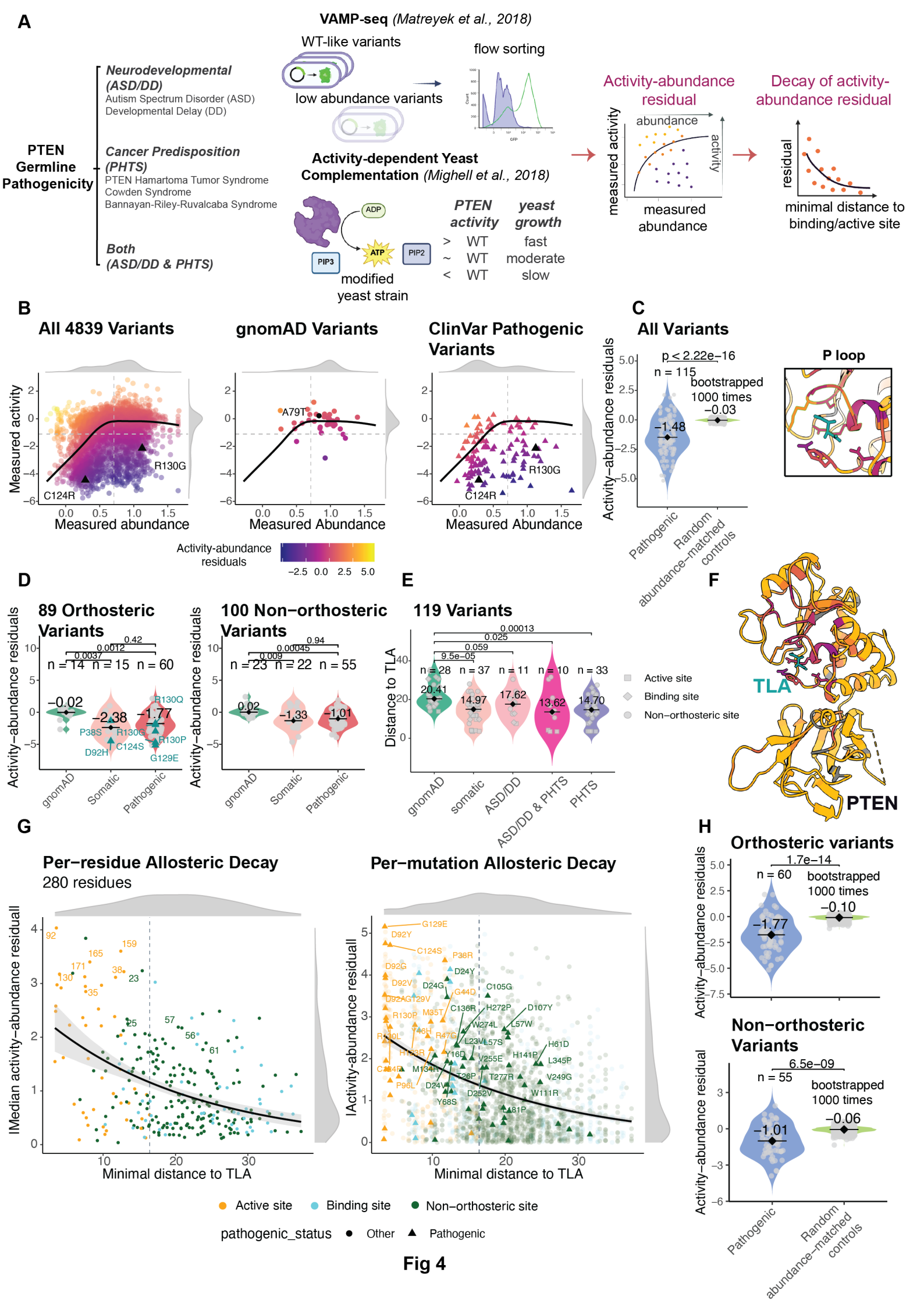

A
VAMP-seq (Matreyek et al., 2018)
WT-like variants
low abundance variants
flow sorting
PTEN Germline Pathogenicity
Neurodevelopmental (ASD/DD)
Autism Spectrum Disorder (ASD)
Developmental Delay (DD)
Cancer Predisposition (PHTS)
PTEN Hamartoma Tumor Syndrome
Cowden Syndrome
Bannayan-Riley-Ruvalcaba Syndrome
Both (ASD/DD & PHTS)
Activity-dependent Yeast Complementation (Mighell et al., 2018)
ADP
ATP
PIP3
PIP2
modified yeast strain
PTEN activity
yeast growth
> WT fast
~ WT moderate
< WT slow
Activity-abundance residual
abundance
activity
measured activity
measured abundance
Decay of activity-abundance residual
residual
minimal distance to binding/active site
B
All 4839 Variants
gnomAD Variants
ClinVar Pathogenic Variants
Measured activity
Measured abundance
R130G
C124R
A79T
Activity-abundance residuals
C
All Variants
p < 2.22e−16
n = 115
bootstrapped 1000 times
−0.03
−1.48
Activity−abundance residuals
Pathogenic
Random abundance−matched controls
P loop
D
89 Orthosteric Variants
100 Non−orthosteric Variants
n = 14 n = 15 n = 60
n = 23 n = 22 n = 55
0.42
0.0012
0.0037
0.94
0.00045
0.009
−0.02
−2.38
−1.77
0.02
−1.33
−1.01
P38S
R130G
R130Q
R130P
D92H
C124S
G129E
gnomAD
Somatic
Pathogenic
E
119 Variants
Distance to TLA
n = 28 n = 37 n = 11 n = 10 n = 33
0.00013
0.025
0.059
9.5e−05
20.41
14.97
17.62
13.62
14.70
gnomAD
somatic
ASD/DD
ASD/DD & PHTS
PHTS
Active site
Binding site
Non−orthosteric site
F
TLA
PTEN
G
Per−residue Allosteric Decay
280 residues
Per−mutation Allosteric Decay
Median activity−abundance residuall
Activity-abundance residuall
Minimal distance to TLA
Active site
Binding site
Non−orthosteric site
pathogenic_status
Other
Pathogenic
H
Orthosteric variants
1.7e−14
n = 60
bootstrapped 1000 times
−0.10
−1.77
Non−orthosteric Variants
6.5e−09
n = 55
bootstrapped 1000 times
−0.06
−1.01
Activity−abundance residuals
Pathogenic
Random abundance−matched controls

Fig 4

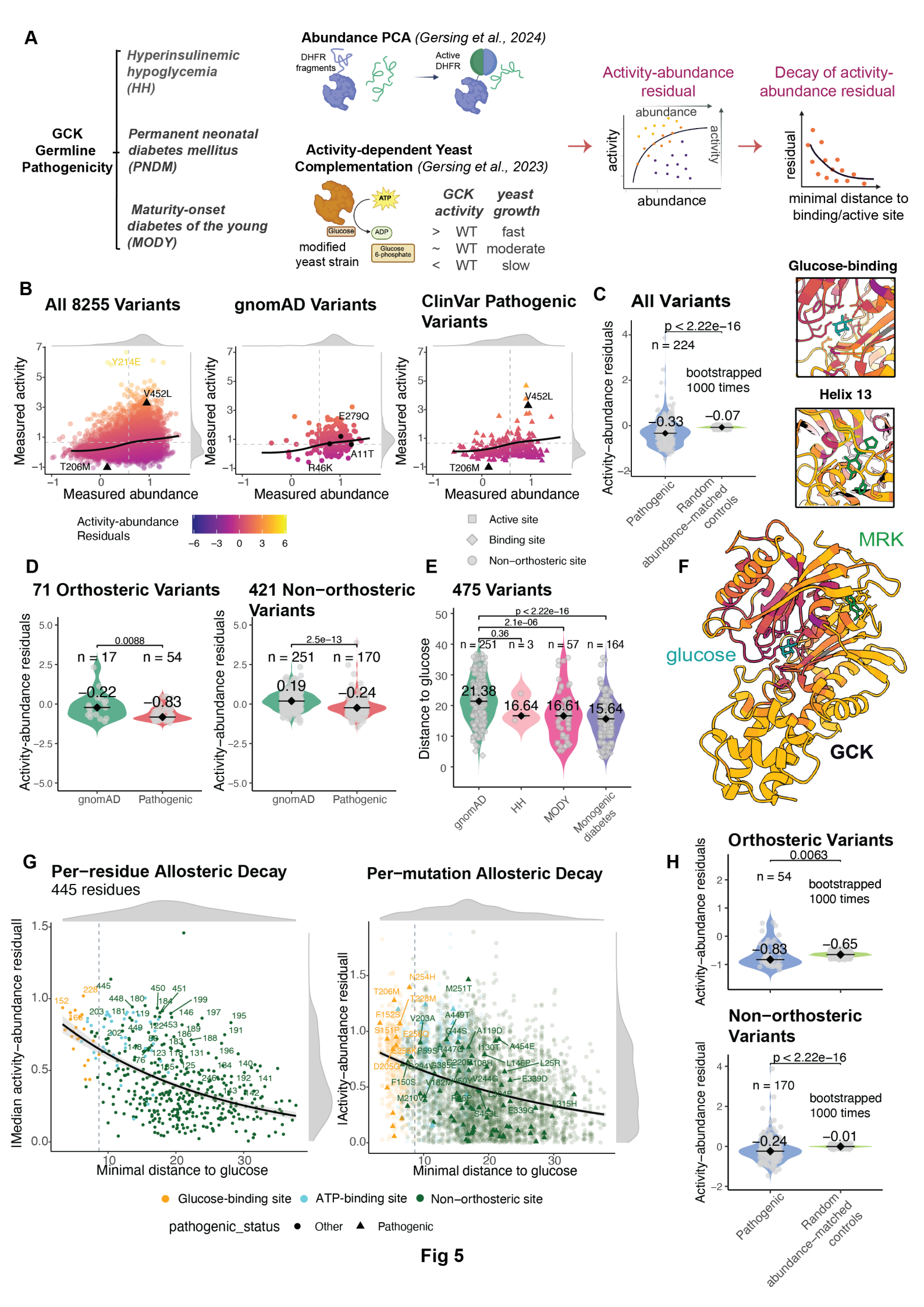
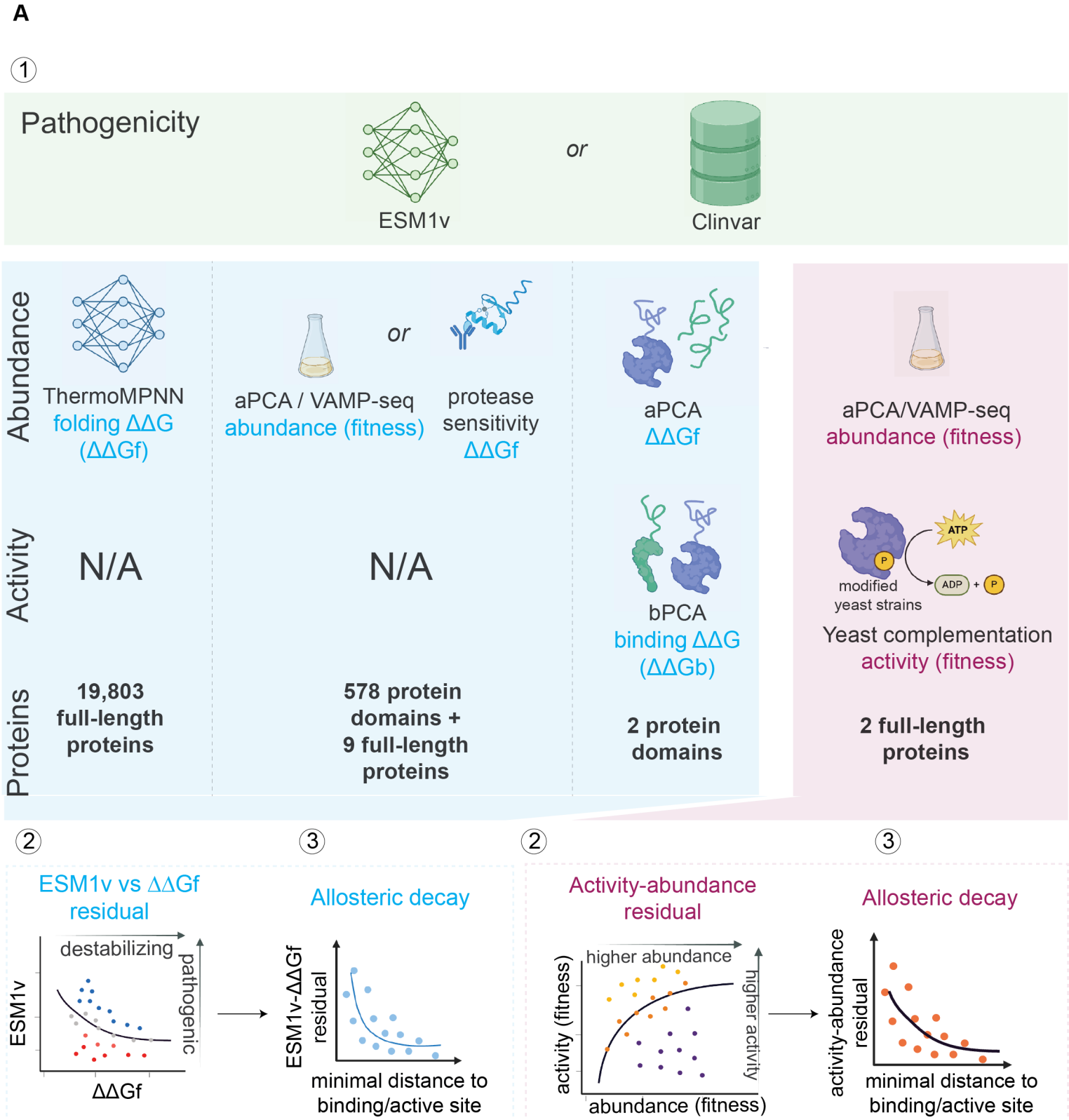
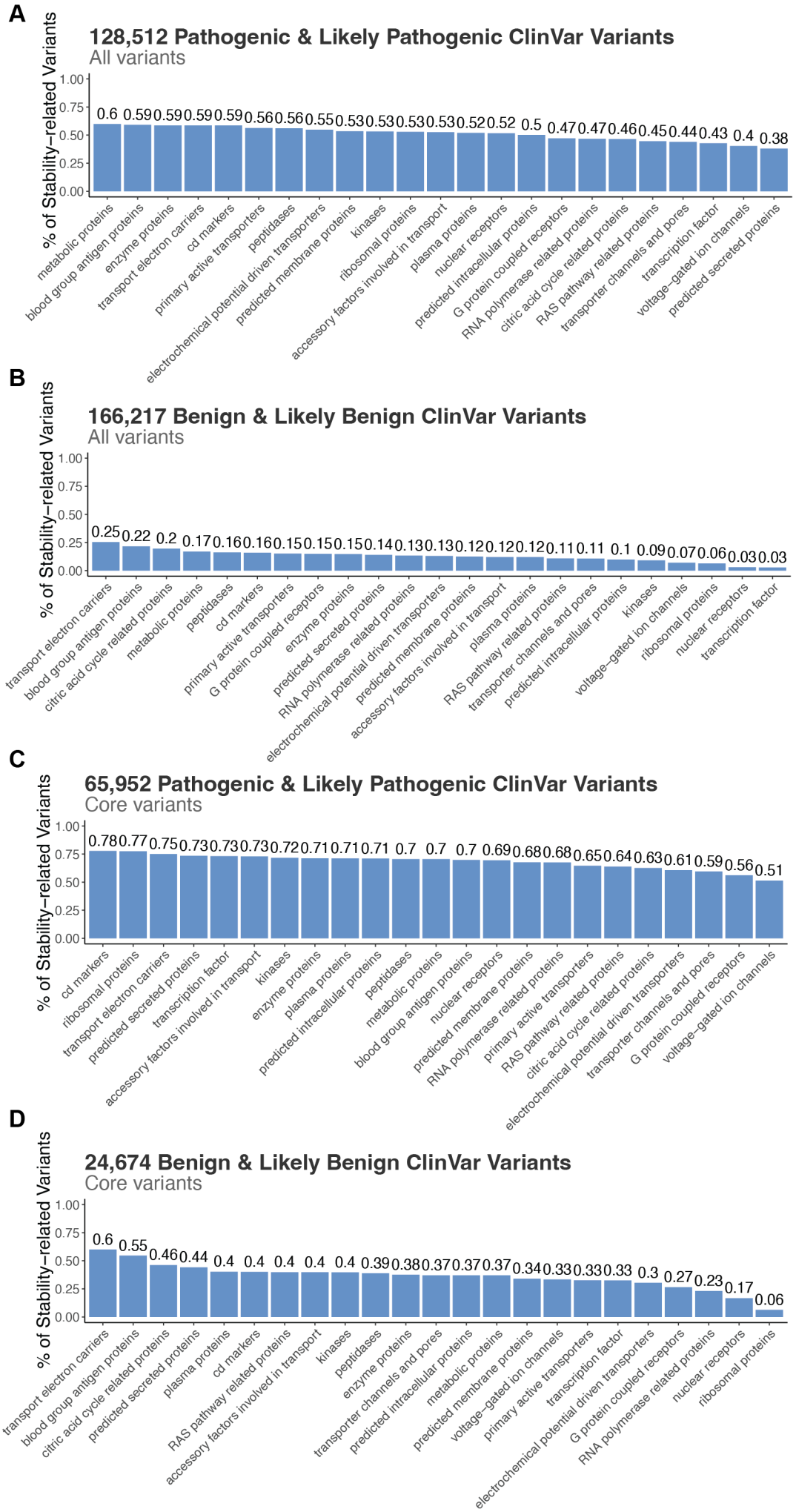

A
GCK Germline Pathogenicity
Hyperinsulinemic hypoglycemia (HH)
Permanent neonatal diabetes mellitus (PNDM)
Maturity-onset diabetes of the young (MODY)
Abundance PCA (Gersing et al., 2024)
DHFR fragments
Active DHFR
Activity-dependent Yeast Complementation (Gersing et al., 2023)
ATP
Glucose
ADP
Glucose 6-phosphate
modified yeast strain
GCK activity
yeast growth
> WT fast
~ WT moderate
< WT slow
Activity-abundance residual
abundance
activity
Decay of activity-abundance residual
residual
minimal distance to binding/active site
B
All 8255 Variants
gnomAD Variants
ClinVar Pathogenic Variants
Measured activity
Measured abundance
Y214E
V452L
T206M
E279Q
A11T
R46K
Activity-abundance Residuals
Active site
Binding site
Non-orthosteric site
C
All Variants
Activity-abundance residuals
p < 2.22e-16
n = 224
bootstrapped 1000 times
-0.33
-0.07
Pathogenic
Random abundance-matched controls
Glucose-binding
Helix 13
D
71 Orthosteric Variants
421 Non-orthosteric Variants
0.0088
n = 17
n = 54
-0.22
-0.83
2.5e-13
n = 251
n = 170
0.19
-0.24
gnomAD
Pathogenic
E
475 Variants
Distance to glucose
p < 2.22e-16
2.1e-06
0.36
n = 251
n = 3
n = 57
n = 164
21.38
16.64
16.61
15.64
gnomAD
HH
MODY
Monogenic diabetes
F
MRK
glucose
GCK
G
Per-residue Allosteric Decay
445 residues
Median activity-abundance residuall
Minimal distance to glucose
Per-mutation Allosteric Decay
Activity-abundance residuall
Glucose-binding site
ATP-binding site
Non-orthosteric site
pathogenic_status
Other
Pathogenic
H
Orthosteric Variants
0.0063
n = 54
bootstrapped 1000 times
-0.83
-0.65
Non-orthosteric Variants
p < 2.22e-16
n = 170
bootstrapped 1000 times
-0.24
-0.01
Pathogenic
Random abundance-matched controls

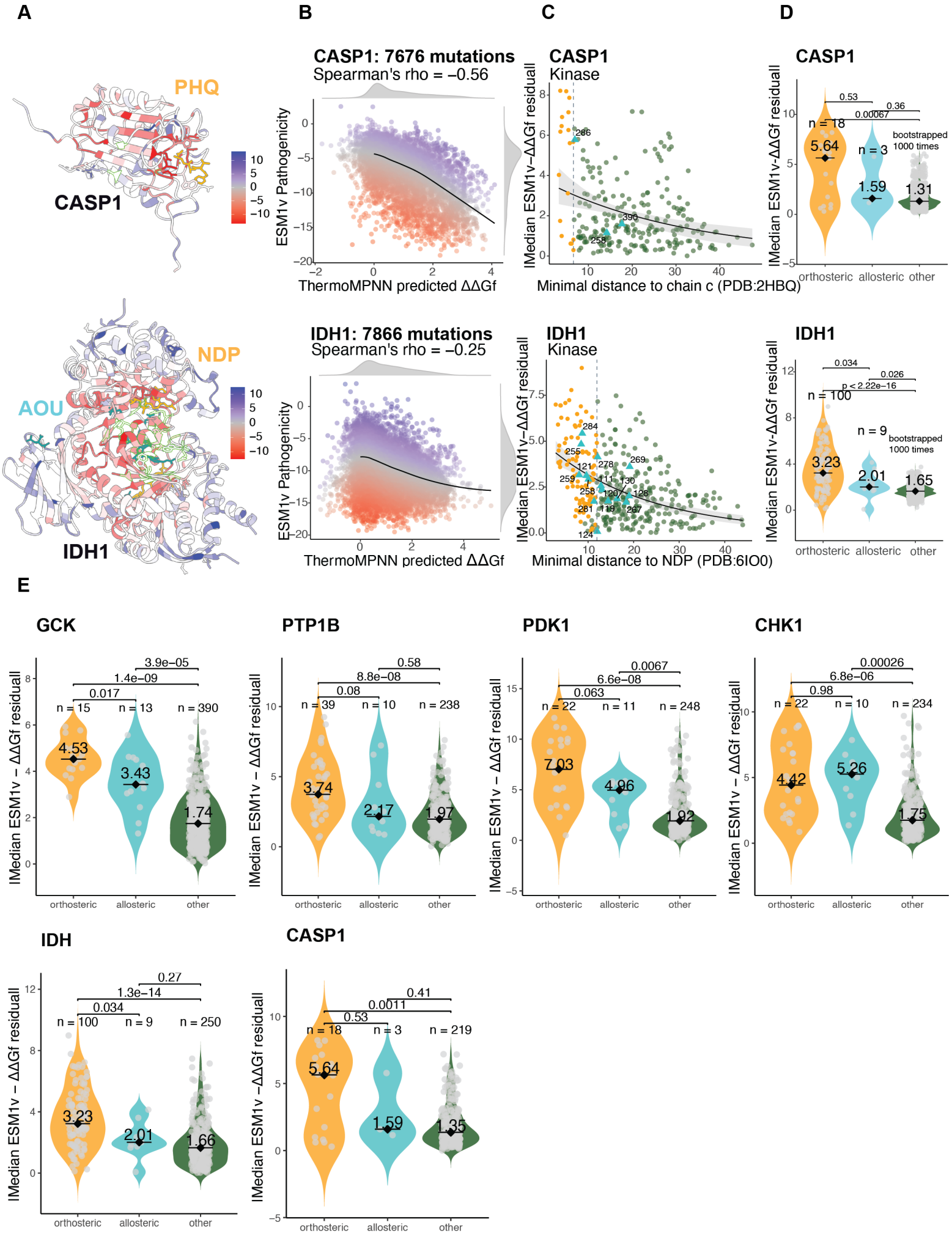
Fig 5



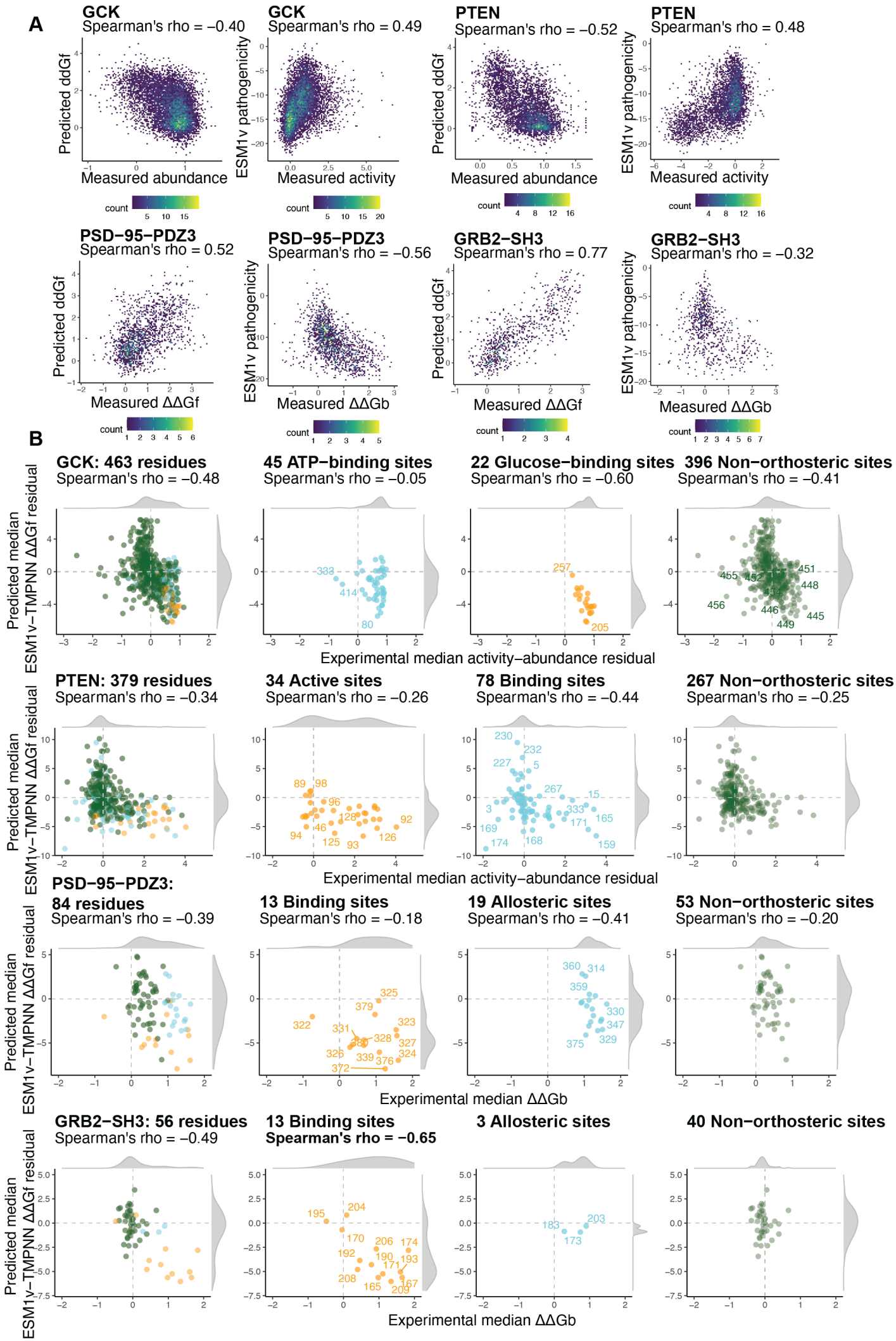
Supplemental fig 1



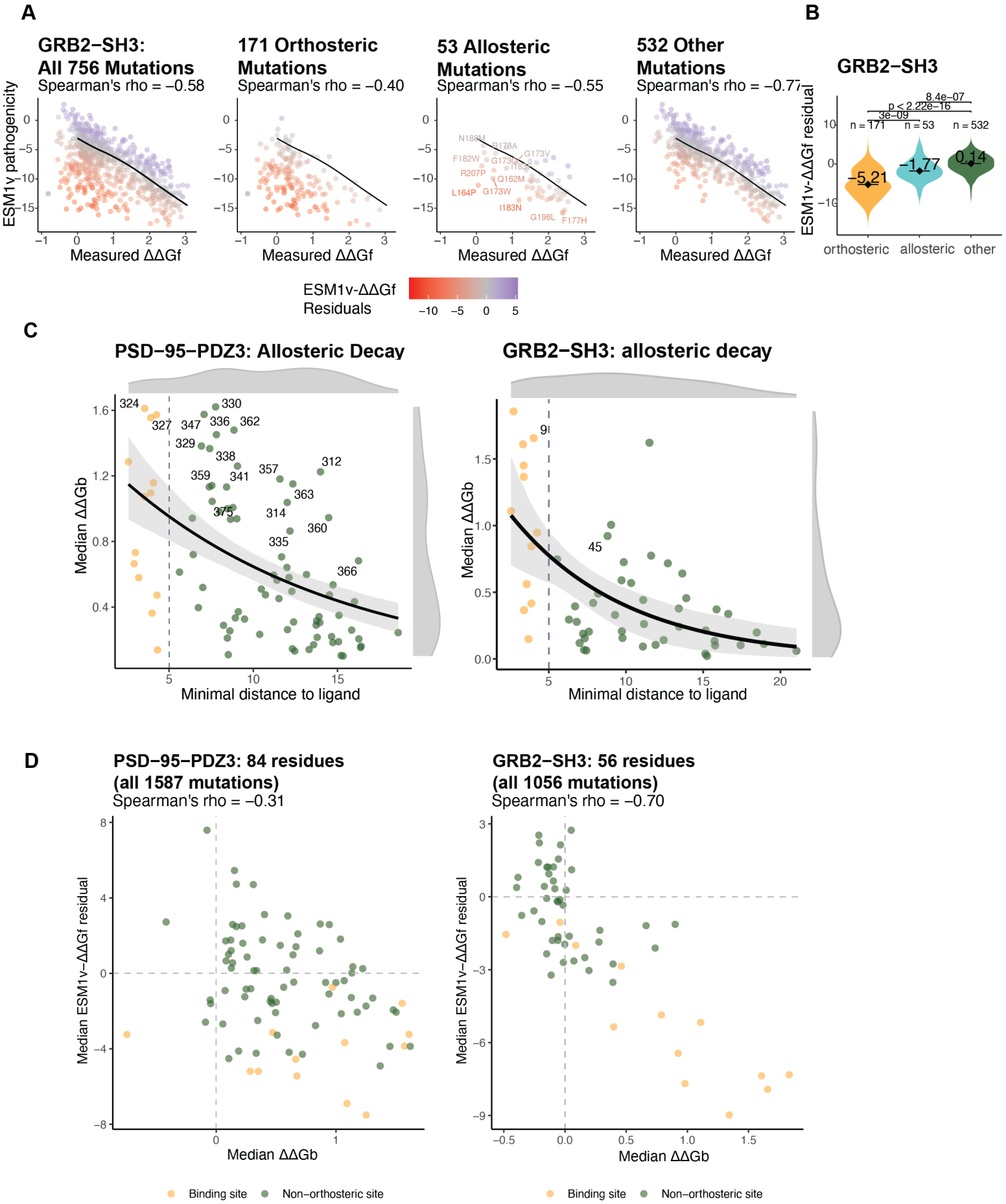
Supplemental fig 2



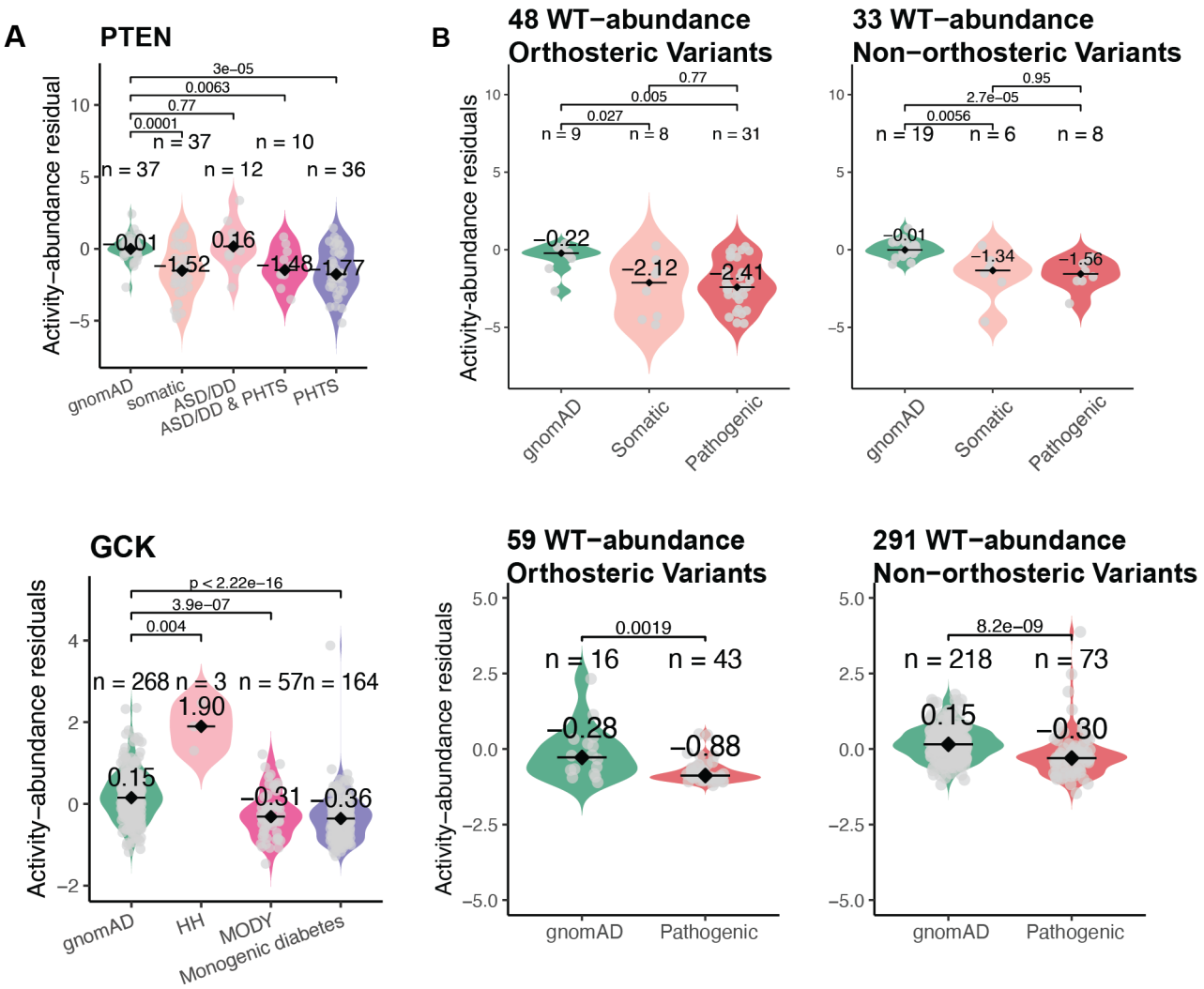
Supplemental fig 3



Supplemental fig 4



Supplemental fig 5



Supplemental fig 6

Supplemental table 2

Protein	PDB	Function	Name	No. residues with matching ESM1v scores	No. mutations with matching ESM1v scores	References
GCK	P35557	Enzyme	Glucokinase	463	8396	Gersing, S., Schulze, T. K., Cagiada, M., Stein, A., Roth, F. P., Lindorff-Larsen, K., & Hartmann-Petersen, R. (2024). Characterizing glucokinase variant mechanisms using a multiplexed abundance assay. Genome biology, 25(1), 98.
PTEN	P60484	Enzyme	Phosphatase and tensin homolog	383	5083	Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., ... & Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature genetics, 50(6), 874-882.
VKOR	Q9BQB6	Enzyme	Vitamin K expoxide reductase	162	2695	Chiasson, M. A., Rollins, N. J., Stephany, J. J., Sitko, K. A., Matreyek, K. A., Verby, M., ... & Fowler, D. M. (2020). Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. elife, 9, e58026.
NUDT15	Q9NV35	Enzyme	Nucleotide triphosphate diphosphatase	163	2922	Suiter, C. C., Moriyama, T., Matreyek, K. A., Yang, W., Scaletti, E. R., Nishii, R., ... & Yang, J. J. (2020). Massively parallel variant characterization identifies NUDT15 associated with thiopurine toxicity. Proceedings of the National Academy of Sciences, 117(10), 5394-5401.
ASPA	P45381	Enzyme	Aspartoacylase	310	5843	Grønbæk-Thygesen, M., Voutsinos, V., Johansson, K. E., Schulze, T. K., Cagiada, M., Pedersen, L., ... & Hartmann-Petersen, R. (2024). Deep mutational scanning reveals a correlation between degradation and toxicity of thousands of aspartoacylase variants. Nature Communications, 15(1), 4026.
TPMT	P51580	Enzyme	Thiopurine S-methyltransferase	241	3648	Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., ... & Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature genetics, 50(6), 874-882.
CYP2C9	P11712	Enzyme	Cytochrome P450 enzyme	486	6370	Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G., ... & Dunham, M. J. (2021). Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. The American Journal of Human Genetics, 108(9), 1735-1751.
OCT	O15245	Transporter	Organic cation transporter 1	547	9803	Yee, S. W., Macdonald, C. B., Mitrovic, D., Zhou, X., Koleske, M. L., Yang, J., ... & Coyote-Maestas, W. (2024). The full spectrum of SLC22 OCT1 mutations illuminates the bridge between drug transporter biophysics and pharmacogenomics. Molecular cell, 84(10), 1932-1947.
PRKN	O60260	Enzyme	E3 ubiquitin-protein ligase parkin	465	8756	Clausen, L., Okarmus, J., Voutsinos, V., Meyer, M., Lindorff-Larsen, K., & Hartmann-Petersen, R. (2024). PRKN-linked familial Parkinson's disease: cellular and molecular mechanisms of disease-linked variants. Cellular and Molecular Life Sciences, 81(1), 223.

Supplemental table 3

Protein	UniProt ID	Active Site Structure	Active Site Substrate	Active Site Annotation	Allosteric Site Structure	Allosteric Modulator	Allosteric Site Annotation	Type	Size	Slope	t-value	P-value	R2	Half Distance
GCK	P35557	1V4S, 3FGU	ANP,BGC, GLC	from 1V4S	1V4S,3ID8	MRK (activator)	literature (Kamata et al., 2004) (Larion & Miller, 2009)	kinase	465	-0.043143	-9.147	< 2.2e-16	0.1654	15.73543
PTP1B	P18031	1PTY	PTR	from 1PTY	1T48,1T49, 1T4J	892,BB3,FRJ (inhibitor)	consensus	phosphatase	435	-0.02548	-4.191	3.71E-05	0.05474	19.73865
PDK1	O15530	4RQK, 5LVM	ADE,ATP	consensus	3ORZ,4RQK	2A2 (activator), R1S (inhibitor)	from 4RQK	kinase	556	-0.048941	-7.152	7.56E-12	0.1519	12.28401
CHK1	O14757	2E9N, 5OOP	76A,ANP	consensus	3F9N	38M (inhibitor)	from 3F9N	kinase	476	-0.039685	-6.378	8.00E-10	0.1302	15.89498
CASP1	P29466	2HBQ, 6BZ9	C,PHQ	consensus	2FQQ	F1G (inhibitor)	from 2FQQ	protease	404	-0.024747	-3.794	0.000188	0.05308	22.64102
IDH	O75874	5YFN, 6ADG, 6B0Z, 6BKZ, 6IO0	FLC,ICT,MG, NAP,NDP	consensus from 6B0Z, 6BKZ, 6IO0	5SUN,6B0Z, 6BKZ,6BL2, 6IO0	70Q, AOU, C81, CIT, DWM, DWS (inhibitor)	consensus from 6B0Z, 6BKZ, 6IO0	oxidoreductase	414	-0.036131	-6.531	2.26E-10	0.1042	15.36862

Table 4. PSD-95-PDZ3: linear regression models

	Model Comparison	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
whole protein	ESM1v ~ ΔΔGf	1565	21,260				
whole protein	ESM1v ~ ΔΔGf + ΔΔGb	1564	18,199	1	3,061.20	263.07	< 2.2e-16 ***
non-orthosteric sites	ESM1v ~ ΔΔGf	1323	15,721				
non-orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb	1322	13,441	1	2279.3	224.18	< 2.2e-16 ***
orthosteric sites	ESM1v ~ ΔΔGf	240	1845				
orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb	239	1704.8	1	140.21	19.657	1.413e-05 ***

Table 5. GRB2-SH3: linear regression model

	Model Comparison	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
whole protein	ESM1v ~ ΔΔGf	754	7,968				
whole protein	ESM1v ~ ΔΔGf + ΔΔGb	753	5,795	1	2,173.30	282.41	< 2.2e-16 ***
non-orthosteric sites	ESM1v ~ ΔΔGf	583	3,265				
non-orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb	582	3,020	1	245.48	47.311	1.568e-11 ***
orthosteric sites	ESM1v ~ ΔΔGf	169	1864.3				
orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb	168	1343	1	521.27	65.205	1.245e-13 ***

Table 6. PSD-95-PDZ3: linear mixed models

	Model Comparison	npar	AIC	BIC	logLik	-2logLik	Chisq	Df	Pr(>Chisq)
whole protein	ESM1v ~ ΔΔGf + (1 | pos)	4	7575.2	7596.6	-3783.6	7567.2			
whole protein	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	7409.7	7436.5	-3699.9	7399.7	167.48	1	< 2.2e-16 ***
non-orthosteric sites	ESM1v ~ ΔΔGf + (1 | pos)	4	6354.3	6375.1	-3173.2	6346.3			
non-orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	6215.8	6241.7	-3102.9	6205.8	140.57	1	< 2.2e-16 ***
orthosteric sites	ESM1v ~ ΔΔGf + (1 | pos)	4	1175.3	1189.2	-583.64	1167.3			
orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	1155.6	1173	-572.78	1145.6	21.705	1	3.179e-06 ***

Table 7. GRB2-SH3: linear mixed models

	Model Comparison	npar	AIC	BIC	logLik	-2logLik	Chisq	Df	Pr(>Chisq)
whole protein	ESM1v ~ ΔΔGf + (1 | pos)	4	3339.9	3358.4	1666	3331.9			
whole protein	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	3304.1	3327.2	-1647	3294.1	37.858	1	7.609e-10 ***
non-orthosteric sites	ESM1v ~ ΔΔGf + (1 | pos)	4	2461.3	2478.8	-1226.7	2453.3			
non-orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	2434.5	2456.4	-1212.3	2424.5	28.795	1	8.047e-08 ***
orthosteric sites	ESM1v ~ ΔΔGf + (1 | pos)	4	810.36	822.92	-401.18	802.36			
orthosteric sites	ESM1v ~ ΔΔGf + ΔΔGb + (1 | pos)	5	807.81	823.52	-398.91	797.81	4.5424	1	0.03307 *