nature biotechnology

Article

https://doi.org/10.1038/s41587-024-02428-4

A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations

Received: 1 February 2024

Accepted: 11 September 2024

Published online: 24 October 2024

Check for updates

Dapeng Xiong[1,2,3,12], Yunguang Qiu[4,5,12], Junfei Zhao[6,12], Yadi Zhou[4,5,12], Dongjin Lee[1,2,12], Shobhita Gupta[2,3,7], Mateo Torres[1,2,3], Weiqiang Lu[8], Siqi Liang[1,2], Jin Joo Kang[1,2,3], Charis Eng[4,9,10], Joseph Loscalzo[11], Feixiong Cheng[4,5,9,10] ✉ & Haiyuan Yu[1,2,3] ✉

To assist the translation of genetic findings to disease pathobiology and therapeutics discovery, we present an ensemble deep learning framework, termed PIONEER (Protein–protein InteractiOn iNtErfacE pRediction), that predicts protein-binding partner-specific interfaces for all known protein interactions in humans and seven other common model organisms to generate comprehensive structurally informed protein interactomes. We demonstrate that PIONEER outperforms existing state-of-the-art methods and experimentally validate its predictions. We show that disease-associated mutations are enriched in PIONEER-predicted protein–protein interfaces and explore their impact on disease prognosis and drug responses. We identify 586 significant protein–protein interactions (PPIs) enriched with PIONEER-predicted interface somatic mutations (termed oncoPPIs) from analysis of approximately 11,000 whole exomes across 33 cancer types and show significant associations of oncoPPIs with patient survival and drug responses. PIONEER, implemented as both a web server platform and a software package, identifies functional consequences of disease-associated alleles and offers a deep learning tool for precision medicine at multiscale interactome network levels.

Precision medicine has sparked major initiatives focusing on whole-genome/whole-exome sequencing (WGS/WES) and developing tools for statistical analyses, all aspiring to identify actionable variants in patients[1,2]. At the center of the vast DNA/RNA sequencing data is their functional interpretation, which largely rests on conventional statistical analyses and trait/phenotype observations[2]. Statistics is crucial for guiding the identification of disease-associated variants; however, traditional WGS/WES studies are commonly underpowered for disease risk variant/gene and drug target discoveries because very large sample sizes are generally required. Furthermore, the statistics do not directly elucidate the functional consequence of the variants. Therefore, translation of genetic and genomic findings to precision medicine is fraught with challenges using traditional statistical approaches.

Optimal information requires knowledge of the whole protein–protein interaction (PPI) network, or interactome, within which the mutant protein operates. On average, each protein interacts directly with 10–15 other proteins[3,4]; thus, the functional consequence of any mutation is not easily (if at all) predicted out of the interactome context. Previous studies[5–9] demonstrated that most disease mutations disrupt specific PPIs rather than affecting all interactions involving the mutant protein. Accurately characterizing such disruptions is essential for understanding the etiology of most disease mutations. Therefore, it is fundamentally important for precision medicine to

A full list of affiliations appears at the end of the paper. ✉e-mail: chengf@ccf.org; haiyuan.yu@cornell.edu

determine structural details, particularly the locations of interaction interfaces of all protein interactions at proteome scale. A clear limitation for this goal is that only approximately 9% of protein interactions have structural models determined by experimental or traditional homology modeling approaches (Fig. 1a and Extended Data Fig. 1a). Predicting co-complex structures of PPIs is experiencing rapid growth resulting from the advent of AlphaFold-based methods as embodied in AlphaFold-Multimer[10], AF2Complex[11] and FoldDock[12], but these methods are all time-consuming and do not scale to solve whole interactomes with hundreds of thousands of PPIs. Furthermore, it should be noted that AlphaFold2-based FoldDock can successfully generate high-quality models for only approximately 2% of human interactions without known homologous structures[12].

Here we present an ensemble deep learning pipeline, termed PIONEER (Protein–protein InteractiOn iNtErfacE pRediction), to generate the next-generation partner-specific interaction interface predictions for experimentally determined PPIs. By using the available atomic resolution co-crystal structures along with homology models, we established a comprehensive multiscale structurally informed human interactome, which consists of 282,095 interactions from humans and seven other commonly studied organisms, including all 146,138 experimentally validated PPIs for 16,232 human proteins (Fig. 1a and Extended Data Fig. 1a). Through this resource, we investigated the network effects of disease-associated mutations at amino acid resolution within the structurally informed interactome of PPI interfaces. We further explored the widespread perturbations of PPIs in human diseases and their significant impact on tumor prognosis and drug responses. This newly constructed structurally informed interactome database was then combined with disease-associated mutations and functional annotations to create an interactive, dynamic web server (https://pioneer.yulab.org) for genome-wide functional genomics studies. It also allows users to perform on-demand interface predictions using the PIONEER framework. Furthermore, we converted the PIONEER framework into a software package that is available to the community to accelerate biological research.

Results

The hybrid deep learning architecture of PIONEER

To date, an overwhelming majority of interactions (~91%) still lack reliable structural information (Fig. 1a). To address this key limitation, we built the PIONEER pipeline to generate partner-specific protein–protein interface predictions for all interactions without structural information. We constructed our labeled dataset for training, validation and testing of our classifiers (Supplementary Data 1): we especially prioritize instances where the same protein interacts with multiple interaction partners using distinct interfaces in our labeled dataset to create a model that better predicts partner-specific interfaces (Fig. 1b); we also require that there are no homologous interactions between any two of the datasets to guarantee the robustness and generalization of our models and a fair performance evaluation.

We used a comprehensive set of single-protein and interaction-partner-specific features for interface prediction (Fig. 1c–f), and both groups of features combine biophysical, evolutionary, structural and sequence information for in-depth characterization of interfaces. Specifically, the single-protein features consist of diverse biophysical features, evolutionary sequence conservation and protein structure properties. However, although these single-protein features capture the characteristics of all possible interface residues, they cannot distinguish interface residues for a protein interacting with different partner proteins through which a protein can perform different biological functions. Previously, we illustrated the importance of encompassing partner-specific features for partner-specific interface predictions[5]. Here, our interaction-partner-specific features include co-evolution of amino acid sequences, protein–protein docking and pair potential. Moreover, we incorporate AlphaFold2-predicted single protein structures[13] into PIONEER to significantly increase the coverage of structure-based features for proteins lacking experimentally determined structures.

To address the non-random missing feature problem, which cannot be adequately resolved by commonly used imputation methods[5], PIONEER's framework is structured as an ensemble of four deep learning architectures, including Structure–Structure, Structure–Sequence, Sequence–Structure and Sequence–Sequence models (Fig. 1c–f and Supplementary Figs. 1 and 2). The Structure–Structure model is used for interactions in which both proteins have structural information, whereas the Sequence–Sequence model is used for proteins without solved structural information. Otherwise, the Structure–Sequence or Sequence–Structure model is used, depending on which protein in the interaction has structural information. This maximizes the amount of information available for each interaction to yield the best possible interface predictions while avoiding potential ascertainment biases that can lead to overfitting.

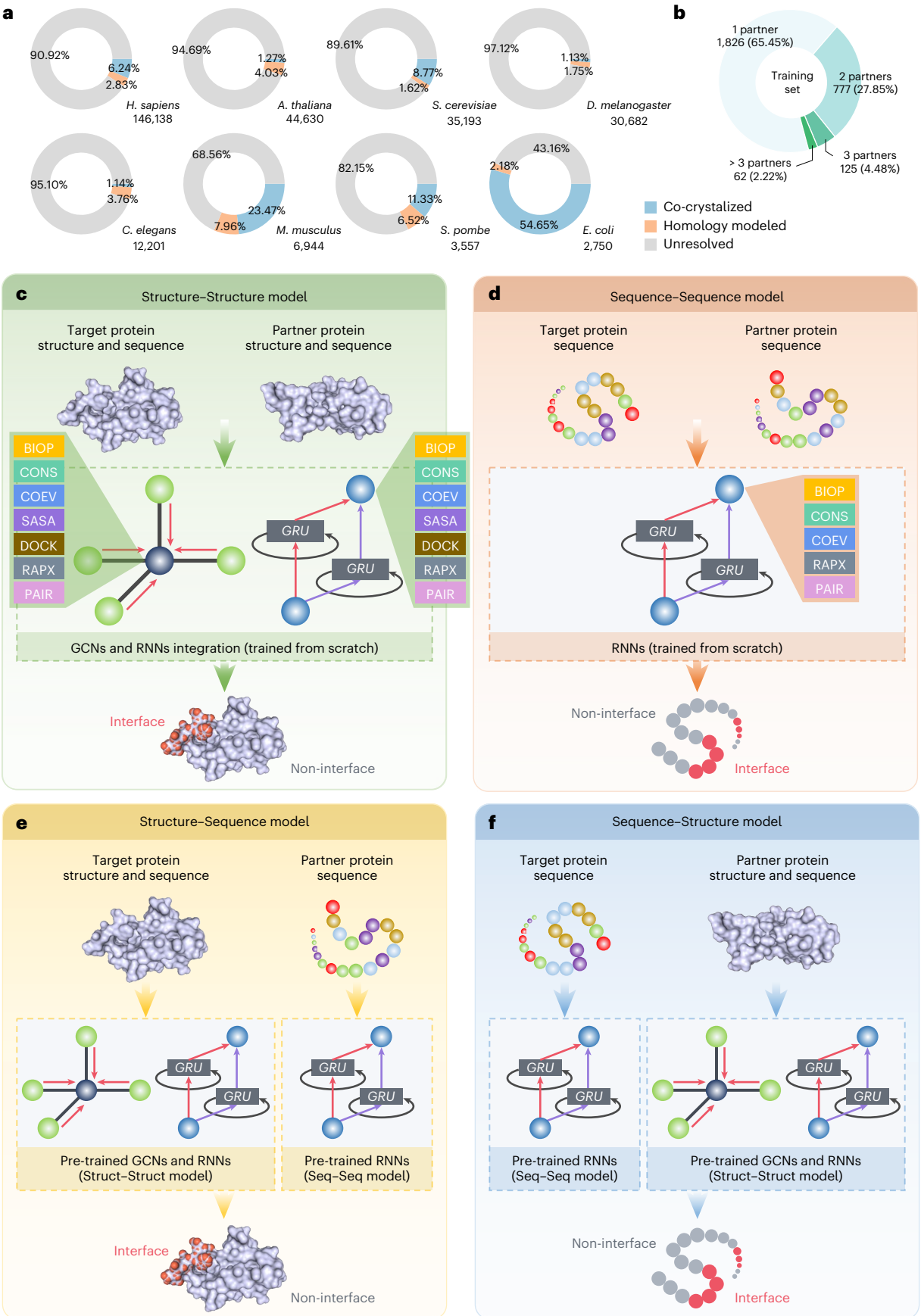
For a protein with available structures, PIONEER uses a hybrid architecture to integrate both structural information embedded through graph convolutional networks (GCNs) with auto-regressive moving average (ARMA) filters[14] and sequence information embedded through bidirectional recurrent neural networks (RNNs) with gated recurrent units (GRUs)[15]. For proteins without high-quality structure models, only sequence information is embedded via RNNs with GRUs. Using transfer learning[16], the pre-trained GCNs and RNNs in the Structure–Structure model and RNNs in the Sequence–Sequence model are deployed in the Structure–Sequence model and the Sequence–Structure model for the processing of proteins with and without structural information, respectively. Furthermore, for each residue in a target protein, our unique architecture integrates embeddings for each residue, overall protein and overall partner protein to make the most accurate interface predictions.

Benchmark evaluation of PIONEER

Our evaluation shows that PIONEER outperforms all other available methods for predicting interfaces of proteins with and without structural information (Fig. 2a,b, Supplementary Figs. 3 and 4 and Supplementary Tables 1–5). We first used the same exact test set that the Structure–Structure model used to evaluate all models for a fair comparison. We can see that the incorporation of structural information clearly improves the performance (Supplementary Table 1). We then compared PIONEER with both partner-specific and non-partner-specific methods to carry out comprehensive evaluation against current state-of-the-art methods. Methods (such as PeSTo[17], ScanNet[18] and MaSIF-site[19]) that are not partner-specific will produce identical interface predictions regardless of the interaction partners, even if they bind at distinct sites of the protein. Our evaluation of all

Fig. 1 | Overview of the PIONEER framework. a, The current size of PPIs from the eight common model organisms with the coverage of experimentally determined co-crystal structures, homology models and the unresolved interactions. **b**, The partner-specific interactions are prioritized in our training dataset for solving partner-specific interface predictions. **c**–**f**, PIONEER architecture consists of an ensemble of four deep learning models that ensures that every residue in the interactomes can be predicted with the maximal amount of available information, and it uses a comprehensive set of biophysical, evolutionary, structural and sequence features for in-depth feature characterization. The **c** and **d** models are used for interactions in which both proteins and neither protein has structural information available, respectively. The GCNs and RNNs are used for structure and sequence information embeddings, respectively. The **e** and **f** models are used for interactions in which only one protein has structure information available. The transfer learning was used in **e** and **f**. Specifically, the pre-trained GCNs and RNNs in **c** and RNNs in **d** were deployed in **e** and **f** as the starting points for model training.

https://doi.org/10.1038/s41587-024-02428-4



methods is partner-specific, because we think that the partner-specific information can be very important for many biological and biomedical applications. It is worth noting that our Sequence–Sequence model, which relies solely on sequence information, has better prediction performance than all recent state-of-the-art structure-based methods that we evaluated, such as PeSTo, ScanNet, BIPSPI+ (ref. 20), MaSIF-site, DeepPPISP[21], SASNet[22] and PIPGCN[23] (Fig. 2a,b and Supplementary Tables 2 and 3). Most of these methods already use cutting-edge deep learning models, illustrating the power of using a comprehensive set of single-protein and partner-specific features; it also validates our design choice to include RNNs with GRUs in a hybrid architecture, even for proteins with known structures. Interestingly, we also found that even our previous ECLAIR with structural information is significantly better than the above structure-based methods and achieves the second-best performance (Fig. 2a and Supplementary Table 2).

We next evaluated the effectiveness of our new models on the benchmark testing dataset by assessing the overall performance of PIONEER and ECLAIR. We found that PIONEER models with ECLAIR features substantially outperform ECLAIR (Supplementary Fig. 5a), confirming that our unique hybrid deep learning architecture captures more information in the features than the previous random forest-based models. Moreover, incorporating new features to PIONEER models further improves the prediction performance (Supplementary Fig. 5a), indicating the outstanding representation ability of our new features for protein interface predictions. Both improvements distinctly demonstrate that our new deep learning architectures and new features make significant contributions to PIONEER's strong ability to provide accurate PPI interface predictions. We then analyzed the feature significance to evaluate the contributions of different features in PIONEER architecture. As the Structure–Structure model uses the most comprehensive set of features, we retrained this model by iteratively removing each individual feature for the feature significance evaluation. We found that the complete PIONEER model achieves the best performance, and each individual feature contributes to the prediction. The solvent-accessible surface area (SASA) feature makes the largest contribution, and the co-evolution and conservation information also make substantial contributions, highlighting the importance of biologically derived features for the characterization of interaction interfaces (Supplementary Fig. 5b). We next assessed the comparison between relative SASA and absolute SASA for interface predictions and found that the relative SASA is more informative (Supplementary Fig. 5c). We also found that the inclusion of AlphaFold2-predicted single protein structures for the proteins without experimentally determined structures improved the PIONEER interface predictions (Supplementary Fig. 5d,e).

Recently, several AlphaFold-based methods, such as AF2Complex, FoldDock and AlphaFold-Multimer, were developed to generate structural models for multi-chain protein complexes. However, they are very computationally intensive and not scalable to whole interactomes. In comparison, PIONEER is approximately 1,000, more than 2,000 and more than 5,500 times faster than AF2Complex, FoldDock and AlphaFold-Multimer, respectively. Additionally, it requires only 21.20%, 18.24% and 15.18% of memory consumption compared to AF2Complex, FoldDock and AlphaFold-Multimer, respectively (Supplementary Fig. 6a,b). The significantly better time and resource efficiency of PIONNER ensures its applicability across entire interactomes. From the performance comparison based on different pLDDT scores (Supplementary Figs. 5c and 6c–g), we can see that the performance of AlphaFold-based methods improves with the increase of pLDDT scores. PIONEER, however, does not rely on the quality of AlphaFold2-predicted structures and can still learn valuable information from low-quality AlphaFold2-predicted structural regions, solidifying the robustness of PIONEER. It is worth noting that, different from PIONEER, which focuses only on finding the interface residues themselves, the AlphaFold-based methods are focused on predicting the structure of the entire co-complex. This means that even small shifts in modeling the interface area can impact the interface residue predictions significantly. Here, we show an interesting example (Supplementary Fig. 6h–k) in which AlphaFold-based methods place the two protein structures in different complex conformations, compared to the known experimental structure. These erroneous placements result in the misclassification of interface residues. In fact, PIONEER and AlphaFold-based methods have fundamentally different use cases: for genome-scale studies, PIONEER has been shown to have the best predictive performance over all other published methods that we evaluated to accurately predict interface residues and, to our knowledge, is the only viable option for modeling whole PPI interactomes; on the other hand, AlphaFold-based methods should be used to study specific PPIs or complexes, especially if three-dimensional (3D) atomic models are required. Also, in contrast to AlphaFold-based methods, users can easily modify and retrain our model based on their own needs on even a single GPU, which ensures PIONEER's high flexibility to researchers.

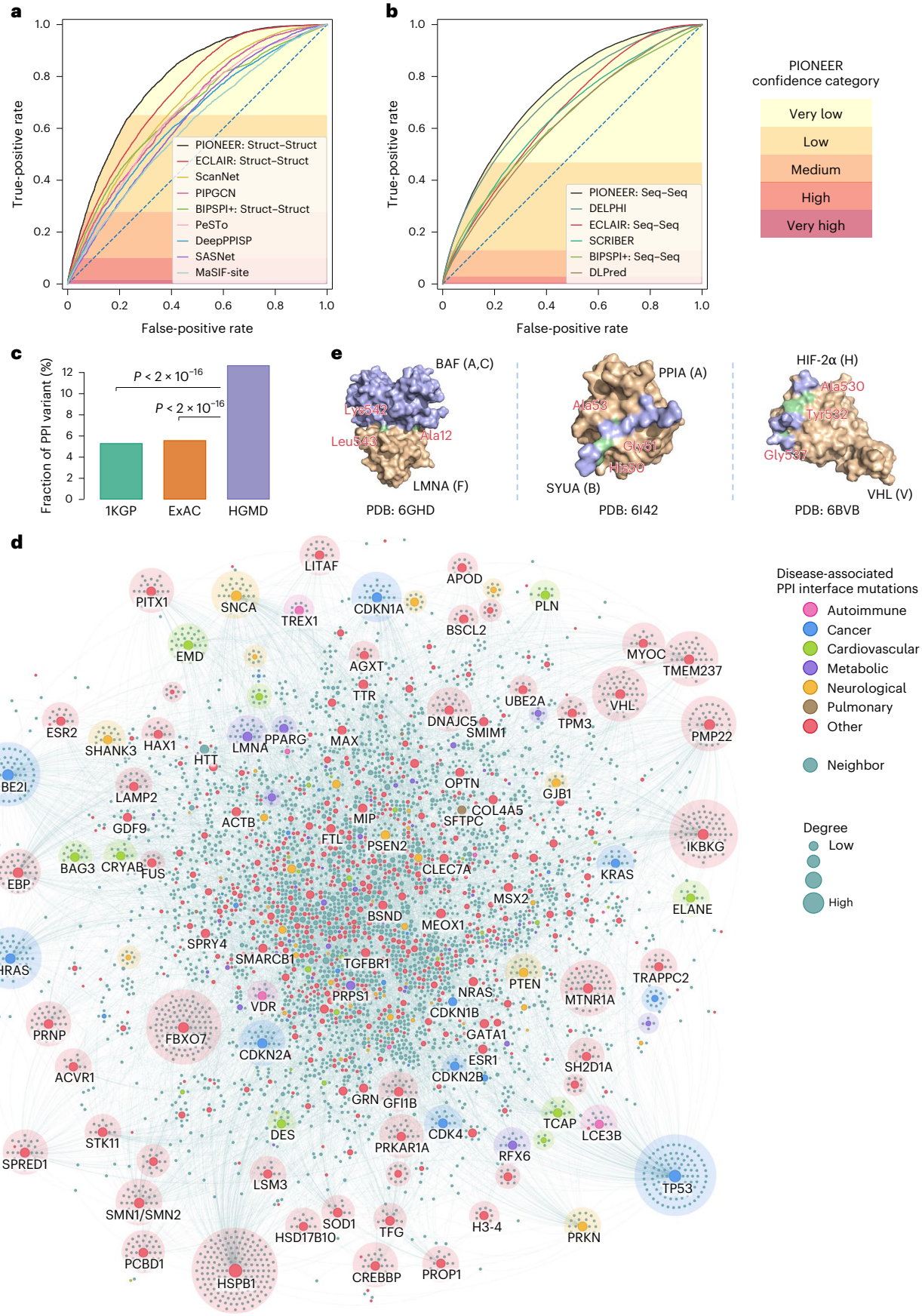
We further applied PIONEER on a widely used Critical Assessment of PRedicted Interactions (CAPRI) benchmark decoy set, Score_set[24]. This dataset contains docking models submitted by 47 participants for proteins from bacteria, yeast, vertebrates and artificial design. We removed duplicated targets as well as those without corresponding UniProt[25] sequences, resulting in 11 targets that have 15,003 decoys: 12,986 incorrect, 732 acceptable, 799 medium and 486 high-quality predictions based on CAPRI-defined criteria, respectively. We then used average PIONEER prediction score at interfaces as the measurement for the model quality evaluation to test the ability of PIONEER in assessing protein complex model quality. Extended Data Fig. 1b shows a clear distinction between any two types of decoy quality, demonstrating that PIONEER interface residue predictions provide a clear signal to model quality.

Proteome-wide interface predictions by PIONEER

Next, we compiled a comprehensive set of experimentally validated binary PPIs for humans and seven model organisms (*Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Schizosaccharomyces pombe* and *Escherichia coli*) by integrating information from commonly used databases[26], including BioGRID[27], DIP[28], IntAct[29], MINT[30], iRefWeb[31], HPRD[32] and MIPS[33]. Here, we focus on binary interactions because the concept of interface only applies if two proteins bind directly. We then used the

Fig. 2 | PIONEER-predicted PPI alleles are enriched in disease-associated mutations. a, Comparison of receiver operating characteristic (ROC) curves of PIONEER Structure–Structure model with other state-of-the-art structure-based methods. **b**, Comparison of ROC curves of PIONEER Sequence–Sequence model with other state-of-the-art sequence-based methods. **c**, Distribution of mutation burden at protein–protein interfaces for disease-associated germline mutations from HGMD in comparison with mutations from the 1KGP and the ExAC. Significance was determined by two-proportion *z*-test. **d**, PPI network with disease-associated interface mutations. Disease associations of the interface mutations were extracted from the HGMD database. Using the PIONEER-predicted high-confidence interface information, PPIs that have at least one disease-associated interface mutation from either one of the two interacting proteins were included in the network. Node colors were determined by the disease categories of their disease-associated interface mutations. Interacting proteins with no known disease-associated interface mutations were colored as 'neighbor'. The final network contains 10,753 PPIs among 5,684 proteins. The figure shows the largest connected component of the network that has 10,706 edges and 5,605 nodes. **e**, Selected structural complex pairs showing germline mutations in the PPI interface. Three disease-associated PPIs with mutations are shown: LMNA–BAF (PDB: 6GHD), PPIA–SYUA (PDB: 6I42) and VHL–HIF-2α (PDB: 6BVB). Interface mutations are shown in green.

https://doi.org/10.1038/s41587-024-02428-4



fully optimized PIONEER pipeline to predict interfaces for all 256,946 binary interactions without experimental structures or homologous models, including 132,875 human interactions (Extended Data Fig. 1a). Because we make partner-specific interface predictions for every residue of every protein, and there are, on average, more than 10 interactions per protein, we made predictions for more than 275 million residue interaction pairs. By combining PIONEER interface predictions with 25,149 interactions (~9%) with experimental or homology models, we generated a comprehensive multiscale structural human interactome, in which all interactions have partner-specific interface information at the residue level, together with atomic resolution 3D models whenever possible. We then analyzed the residue distribution within interaction interfaces based on both different groups categorized by biochemical properties[34–36] and each individual residue. We observed that charged residues are more enriched in the interfaces, and some residues, such as cysteine, tryptophan, methionine and histidine, appear less frequently in interfaces (Supplementary Fig. 7). These results agree well with previously reported statistics[37,38] and further suggest the importance of biophysiochemical properties in protein interface predictions.

To comprehensively evaluate the quality of our predicted interfaces and their biological implications, we performed large-scale mutagenesis experiments to measure the fraction of disrupted interactions by mutations in our predicted interfaces at varying confidence levels, in comparison to that of known interface and non-interface residues from co-crystal structures in the Protein Data Bank (PDB)[39]. Using our Clone-seq pipeline[40], we generated 2,395 mutations on 1,141 proteins and examined their impact on 6,754 mutation interaction pairs through a high-throughput yeast-two-hybrid (Y2H) assay for a large-scale experimental validation. We observed that mutations at PIONEER-predicted interfaces disrupt PPIs at a very similar rate to the mutations at known experimentally determined interfaces, and both of their disruption rates are significantly higher than that of known non-interfaces (Extended Data Fig. 1c). Therefore, our large-scale experiments confirm the high quality of our interface predictions and the validity of our overall PIONEER pipeline. Because interaction disruption is key to understanding the molecular mechanisms of disease mutations[8,40], our experimental results indicate that PIONEER-predicted interfaces are instrumental in prioritizing disease-associated variants and generating concrete mechanistic hypotheses.

PIONEER-informed interfaces enriched with disease mutations

Because disruption of specific PPIs is essential for the pathogenicity of many disease mutations[6,7,41], we next measured the enrichment of known disease-associated mutations from the Human Gene Mutation Database (HGMD)[42] in PIONEER-predicted interfaces and compared it to known interfaces from co-crystal structures. We found that the residues predicted by PIONEER with a high interface confidence show a very similar rate of disease mutation enrichment to those of known interfaces (Extended Data Fig. 1d). We observed that 251,368 (~98%) out of all 256,946 binary interactions have at least one or more predicted interface residues that fall into high or very high confidence categories (Supplementary Fig. 8), indicating that PIONEER provides meaningful structural information for almost all human PPIs. In fact, each bin with a higher confidence of predicted interfaces is more likely to contain disease-associated mutations than the previous bin, demonstrating the strong correlation between PIONEER prediction scores and true protein function (Extended Data Fig. 1d). We further analyzed the distribution of population genetic variants and found that their enrichment in PIONEER-predicted interfaces and non-interfaces matches well with that of known interfaces and non-interfaces, respectively (Extended Data Fig. 1e). The results also show that there is a depletion of common variants (that is, not deleterious) in both known and predicted interfaces, indicating that PIONEER predicts functionally important interface variants effectively. We also found that, compared to variants identified in individuals from the 1000 Genomes Project (1KGP)[43] and the Exome Aggregation Consortium (ExAC)[44], disease-associated mutations from HGMD are more significantly enriched in PPI interfaces of the respective proteins[7] (Fig. 2c). Moreover, as predicted by CADD[45] and FoldX[46], the population variants in PIONEER-predicted interfaces are more likely to adversely affect protein functions than those in PIONEER-predicted non-interfaces (Supplementary Fig. 9), which confirms that deleterious variants preferentially occur in protein–protein interfaces[6,8].

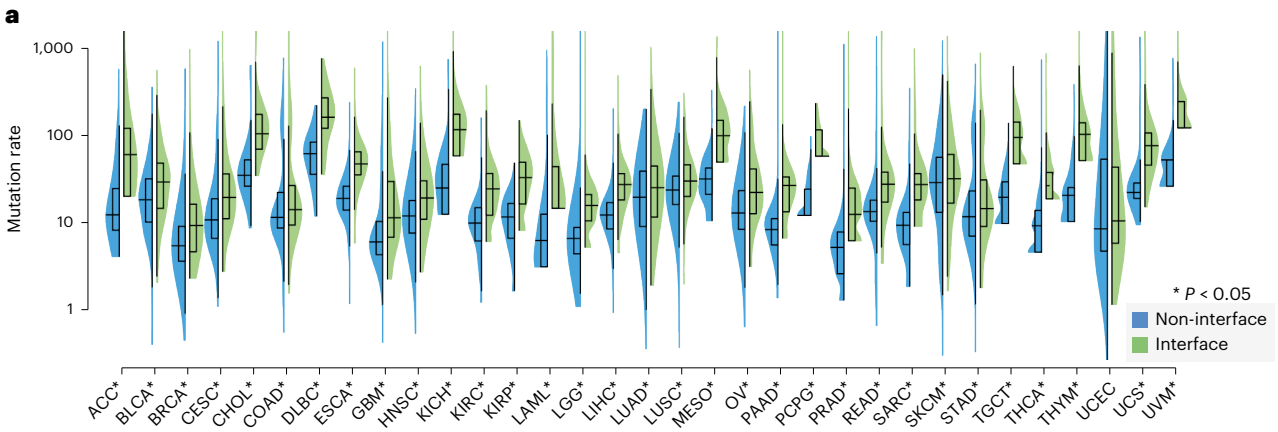
To further evaluate whether the disease-associated mutations were enriched in PIONEER-predicted PPI interfaces, we next categorized the disease-associated germline mutations from HGMD into seven major disease groups[47], including autoimmune, cancer, cardiovascular, metabolic, neurological, pulmonary and an additional 'other' category. We identified 10,753 PPIs among 5,684 proteins that had at least one disease-associated interface germline mutation (Fig. 2d and Supplementary Table 9), among which 9,795 (~91%) have such interface mutations on one protein (the other protein colored as 'neighbor') and 958 (~9%) on both interacting proteins. Overall, this network analysis shows that PIONEER-predicted PPI interfaces are altered by broad disease-associated mutations across multiple disease categories. To highlight the power of PIONEER-predicted interfaces, we examined three PPI interfaces with germline alleles. The germline mutation p.Lys542Gln on LMNA buried in the interfaces of LMNA and BAF (Fig. 2e) is associated with progeroid disease[48]. One loss-of-function PPIA mutation p.Ala53Glu in the interfaces of PPIA–SYUA (Fig. 2e) was identified in patients with early-onset Parkinson's disease[49]. The germline mutation p.Gly537Arg on HIF-2α associated with polycythemia vera[50] is located in PIONEER-predicted VHL–HIF-2α interfaces (Fig. 2e) and disrupts VHL binding via impairing ubiquitination and proteasomal degradation of HIF-2α[51,52]. Taken together, PIONEER-predicted protein–protein interface mutations convey crucial structural information in delineating the functional consequences for disease mechanisms at the atomic and allele levels.

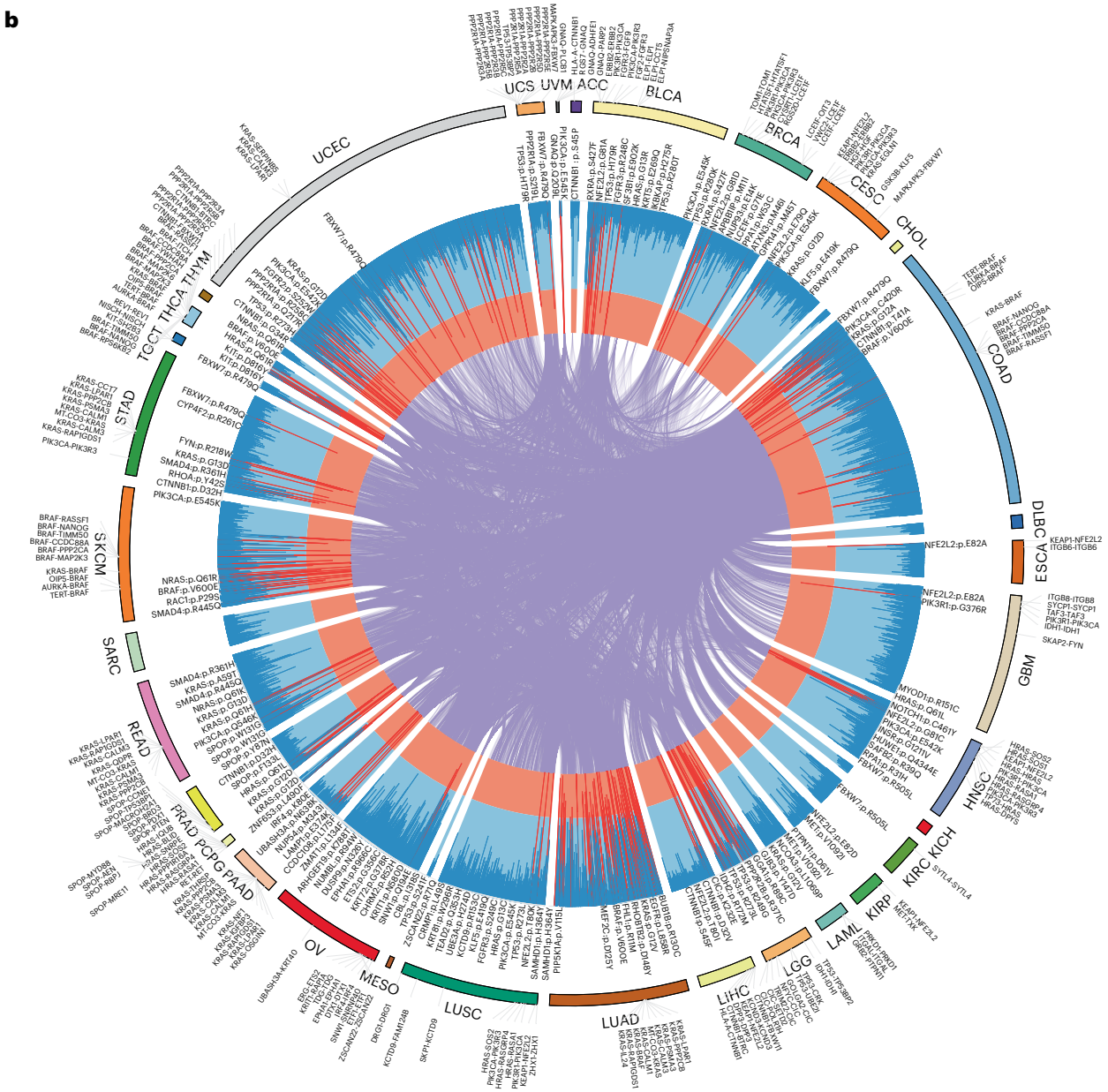
PIONEER-predicted oncoPPIs across 33 cancer types

We next investigated the somatic mutations from patients with cancer in the context of PIONEER-predicted interfaces. In total, we collected approximately 1.7 million missense somatic mutations from the analysis of approximately 11,000 tumors across 33 cancer types from The Cancer Genome Atlas (TCGA)[53]. We found significant enrichment of somatic missense mutations in PIONEER-predicted PPI interfaces compared to non-interface regions (Fig. 3a and Supplementary Data 2). Specifically, this significant enrichment was observed in 31 out of the total 33 cancer types regardless of the overall mutation burden. In lung squamous cell carcinoma, one of the cancer types that has the highest mutation

Fig. 3 | A landscape of oncoPPIs identified by PIONEER across 33 cancer types (~11,000 cancer genomes). a, Distribution of missense somatic mutations in protein–protein interfaces versus non-interfaces across 33 cancer types/subtypes from TCGA. Data are represented as violin plots with underlaid box plots, where the middle line is the median; the lower and upper edges of the rectangle are the first and third quartiles; and the lower and upper whiskers of the violin plot represent the interquartile range (IQR) × 1.5. Significance was determined by two-tailed Wilcoxon rank-sum test. The *n* numbers and *P* values are shown in Supplementary Table 10. **b**, Circos plot displaying significant putative oncoPPIs harboring a statistically significant excess number of missense somatic mutations at PPI interfaces across 33 cancer types. Putative oncoPPIs with various significance levels are plotted in the two inner layers. The links (edges, purple) connecting two oncoPPIs indicate two cancer types sharing the same oncoPPIs. Selected significant oncoPPIs and their related mutations are plotted on the outer surface. The length of each line is proportional to $-\log_{10}(P)$. All *P* values were adjusted for multiple testing using the Bonferroni correction.

https://doi.org/10.1038/s41587-024-02428-4





https://doi.org/10.1038/s41587-024-02428-4

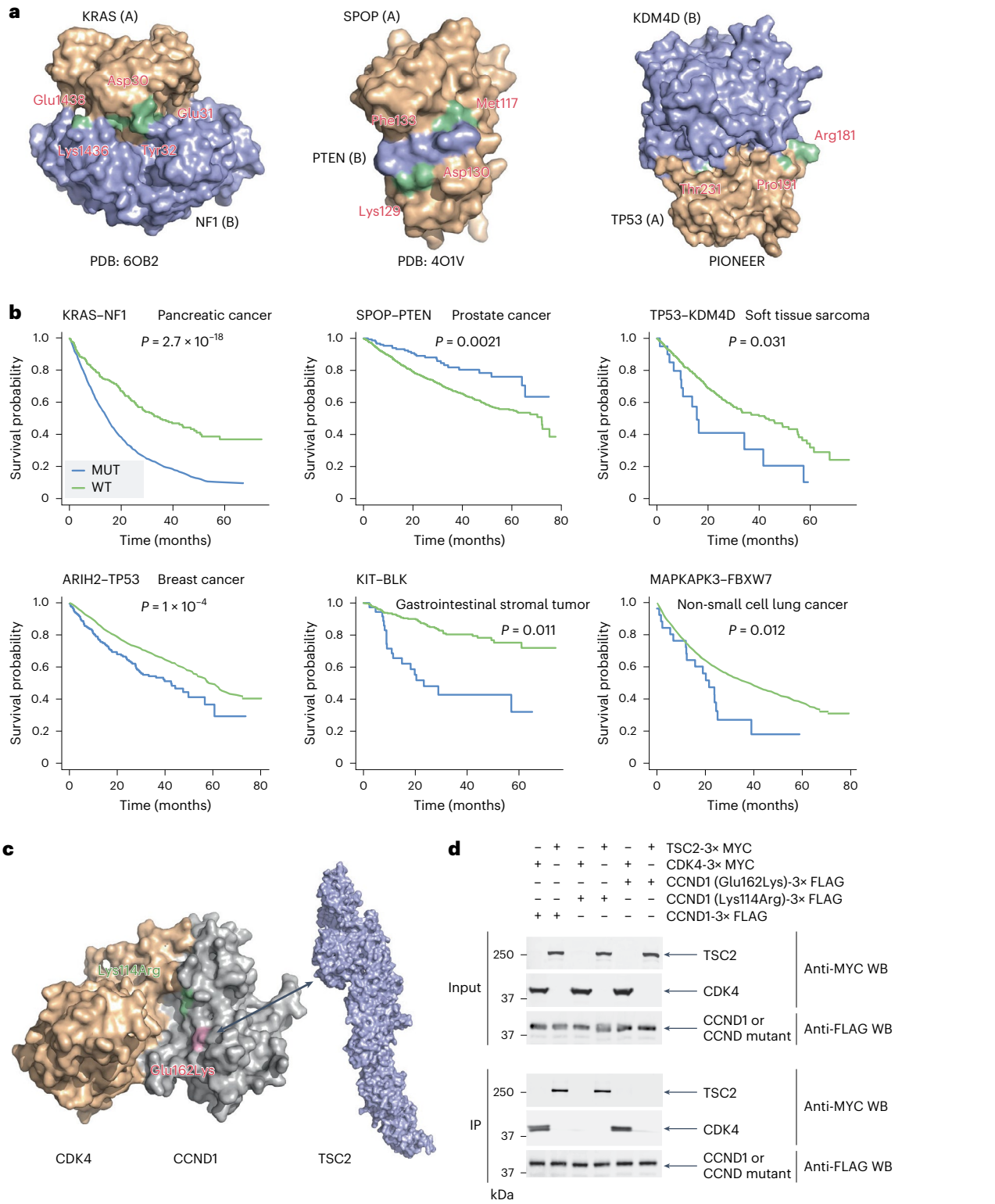


Fig. 4 | PIONEER-predicted oncoPPIs are associated with patient survival. **a**, Selected structural complex pairs showing somatic mutations in the oncoPPI interfaces. Interface mutations are shown in green. **b**, Survival analysis of six exemplary PPI-perturbing mutations in diverse cancer types. MUT, mutations. Significance was determined by two-sided log-rank test. The *n* numbers are shown in Supplementary Table 12. **c**, Example of PIONEER partner-specific interface prediction. The mutations CCND1 p.Lys114Arg and CCND1 p.Glu162Lys are shown in green and pink, respectively. **d**, Experimental validation of the partner-specific interface predictions in **c** by co-immunoprecipitation using HEK293T cells. WB, western blotting; IP, immunoprecipitation. The experiment was repeated three times independently.

load per exome, we observed 29 variants per 1 million amino acids affecting PPI interfaces, whereas the rate for non-PPI interface region is 23 ($P = 1.3 \times 10^{-11}$). For thyroid cancer, with the lowest mutation load, the difference is 27 for PPI interfaces versus 9 for the remainder of the protein sequences ($P < 10^{-16}$). To account for the potential bias in this analysis due to data sources, we divided our whole structural human

interactome into three categories—experimental structures (PDB, 6.2%), homology models (2.8%) and PIONEER predictions (90.9%)—and performed the enrichment analysis for each category separately. The results showed that the same enrichment pattern is independent of the data source, suggesting the robustness of PIONEER interface predictions (Supplementary Fig. 10). We next sought to identify PPIs significantly enriched with somatic mutations in their interfaces (named oncoPPIs) in both cancer-type-specific and pan-cancer analyses. Our analysis yielded a total of 586 statistically significant oncoPPIs across 33 cancer types (Fig. 3b and Supplementary Data 3), including KRAS–BRAF, TP53–EGLN1 and TP53–TP53BP2 across 10 cancer types.

We then turned to analyze the clinical sequencing data from MSK-MET, a pan-cancer cohort of over 25,000 patients spanning 50 different tumor types[54]. Of the 157,979 missense mutations that we investigated, 40,526 (~26%) were identified to affect 15,523 unique PPI interfaces. Focusing on the PPIs that were disturbed in at least 10 samples in a specific cancer type, we performed survival analysis to identify clinically actionable oncoPPIs whose disruption is significantly associated with patient survival. KRAS has been reported to co-mutate with NF1 in response to GTP hydrolysis[55]. We identified that mutations of KRAS–NF1 interface residues, such as Asp30 and Glu31 on KRAS (Fig. 4a), are significantly associated with poor survival rate compared to the wild-type (WT) group in pancreatic cancer ($P = 2.7 \times 10^{-18}$; Fig. 4b). SPOP plays a multifaceted role in oncogenesis and progression by mediating degradation of PTEN[56]. The SPOP MATH domain contains a mutation p.Phe133Val[57] on its PIONEER-predicted interface for binding to PTEN, which is significantly associated with survival rate in prostate cancer ($P = 0.0021$; Fig. 4b). Patients with several PIONEER-predicted interface mutations (Thr231, Pro191 and Arg181 on TP53; Fig. 4a) between TP53 and KDM4D are significantly associated with poor survival in soft tissue sarcoma ($P = 0.031$; Fig. 4b). Furthermore, oncoPPI analysis revealed that PIONEER-predicted interface mutations on ARIH2–TP53, kinase–substrate (for example, KIT–BLK), kinase–E3 ligase (for example, MAPKAPK3–FBXW7) and cyclin–E3 ligase (for example, CCND1–FBXO31) are significantly associated with survival rate in breast cancer ($P = 1 \times 10^{-4}$), gastrointestinal stromal tumor ($P = 0.011$), non-small cell lung cancer ($P = 0.012$) and endometrial cancer ($P = 0.024$), respectively (Fig. 4b and Supplementary Fig. 11). Accumulated evidence suggested that the mutations on CCND1 are associated with multiple cancer types[58]. By analyzing PIONEER-predicted oncoPPIs, we found that PIONEER-predicted interface mutations on CCND1 are significantly enriched in the CCND1–CDK4 interfaces in uterine cancer ($P = 0.012$) and low-grade glioma ($P = 0.048$). We identified that CCND1 interacts not only with CDK4 but also with TSC2 from PIONEER-predicted interfaces. Specifically, we identified that CCND1 interacts with CDK4 and TSC2 via two unique sets of interfaces (Fig. 4c). Next, we experimentally confirmed this result using co-immunoprecipitation with 293T cells. Figure 4d shows that mutation p.Lys114Arg on CCND1 specifically disrupts the interaction between CCND1 and CDK4, without disrupting its interaction with TSC2. Interestingly, mutation p.Glu162Lys on CCND1 does not disrupt its interaction with CDK4 but does disrupt its interaction with TSC2. Mutations p.Lys114Arg and p.Glu162Lys on CCND1 are associated with myeloma[59] and lung cancer[60], respectively. These results further demonstrate that PIONEER-generated structural human interactome can uncover tumorigenesis with distinctive functions corresponding to distinct interfaces, even for those on the same proteins.

PIONEER-informed alleles alter ubiquitination by E3 ligases

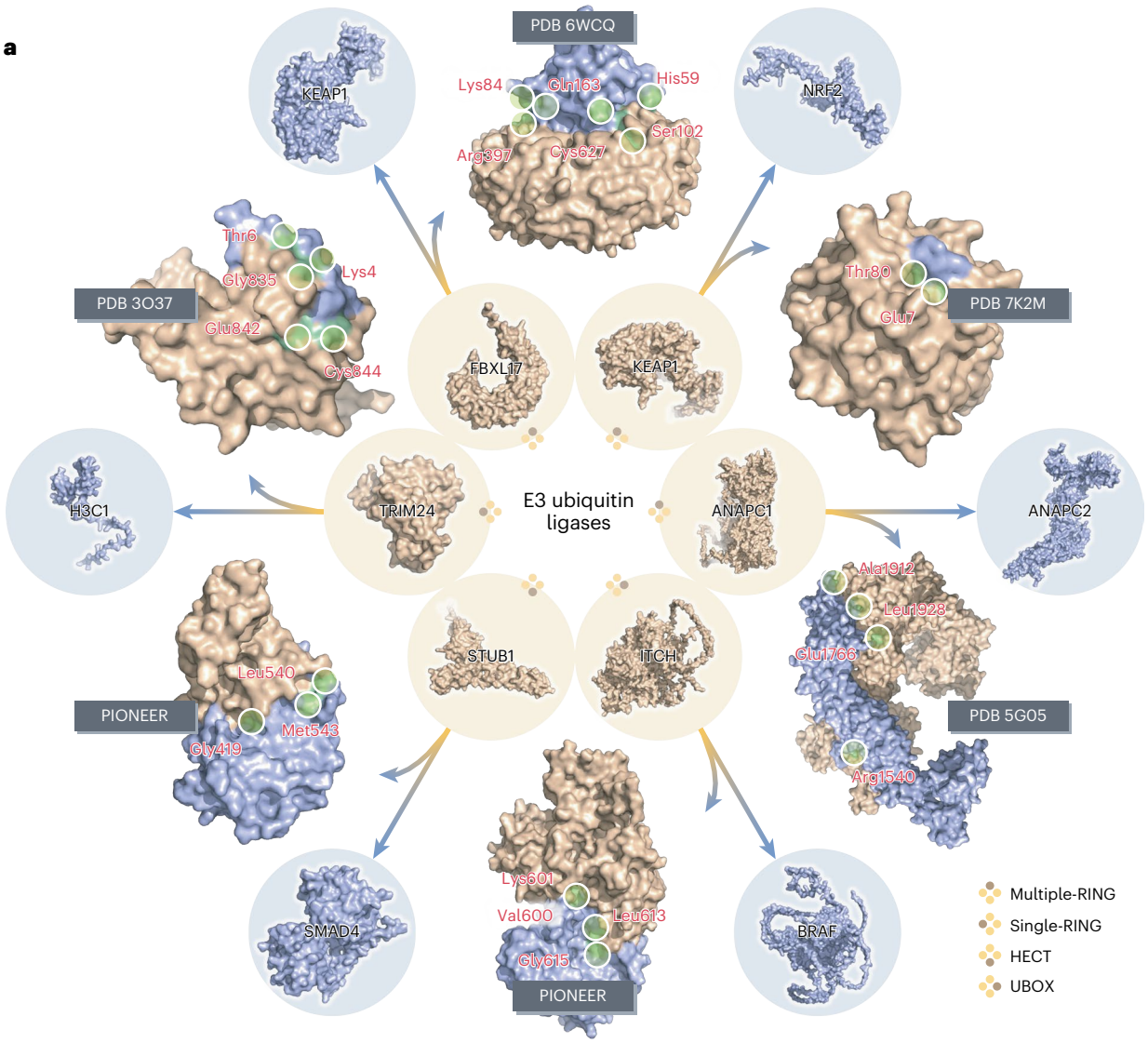
E3 ubiquitin ligases are involved in cellular transformation and tumorigenesis by targeted protein degradation[61,62]. Identifying how somatic mutations alter PPIs of E3 ligases may offer novel targets for development of targeted protein degradation therapies[63]. We investigated 4,614 PIONEER-predicted oncoPPIs connecting 355 E3 ligases annotated from E3Net[64] and UbiNet2.0 (ref. 65) databases. We next focused on 204 oncoPPIs connecting E3 ligases with significant association with patient survival rates and/or significant association with drug responses measured in tumor cell lines or patient-derived tumor xenograft (PDX) mouse models (Supplementary Table 13).

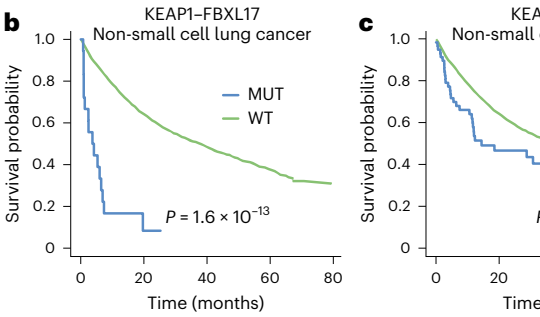
Figure 5a illustrates the selected examples of the most significant PIONEER-predicted oncoPPIs of E3 ligases (Supplementary Table 14). Among these 204 oncoPPIs, FBXW7 has the highest number of oncoPPIs (47/204; Supplementary Fig. 12 and Supplementary Table 15). FBXL17 is a multiple-RING E3 ligase that specifically recognizes and ubiquitinates the BTB proteins[66]. We found that PIONEER-predicted PPI-perturbing mutations on FBXL17–KEAP1, such as p.Ser102Leu on KEAP1, are significantly associated with poor survival in non-small cell lung cancer ($P = 1.6 \times 10^{-13}$; Fig. 5b). A multiple-RING E3 ligase complex ANAPC1–ANAPC2 (Fig. 5a) is positively regulated by the PTEN/PI3K/AKT pathway and modulates ubiquitin-dependent cell cycle progression[67]. We found that PIONEER-predicted PPI-perturbing mutations on ANAPC1–ANAPC2 is associated with resistance to a PI3K inhibitor, BKM120 ($P = 0.0043$; Supplementary Fig. 13a). ITCH, a HECT-type E3 ubiquitin ligase, has been reported to mediate BRAF kinase poly-ubiquitination and promote proliferation in melanoma cells[68]. We found that PIONEER-predicted PPI-perturbing mutations on BRAF–ITCH, such as p.Val600Glu and p.Lys601Glu on BRAF, are significantly associated with sensitivity to dabrafenib (an ATP-competitive inhibitor; $P = 1.7 \times 10^{-21}$; Supplementary Fig. 13b). STUB1, a U-box-dependent E3 ubiquitin ligase, was reported to degrade SMAD4, an intracellular signaling mediator of the TGF-β pathway[69]. Multiple PIONEER-predicted PPI-perturbing mutations on STUB1–SMAD4, including p.Gly419Arg (Trp, Val) and p.Leu540Pro (Arg) on SMAD4, are significantly associated with poor survival in colorectal cancer ($P = 0.025$; Supplementary Fig. 13c). A single-RING E3 ligase, TRIM24, is an oncogenic transcription co-factor overexpressed in breast cancer[70]. We found that PIONEER-predicted PPI-perturbing mutations on TRIM24–H3C1 (Fig. 5a) are significantly associated with resistance to GDC0941 (an EGFR signaling inhibitor; $P = 0.028$; Supplementary Fig. 13d). Treatment with an EGFR inhibitor suppresses TRIM24 expression and H3K23 acetylation and, thereby, inhibits EGFR-driven tumor growth[71], supporting the PIONEER-predicted oncoPPI findings.

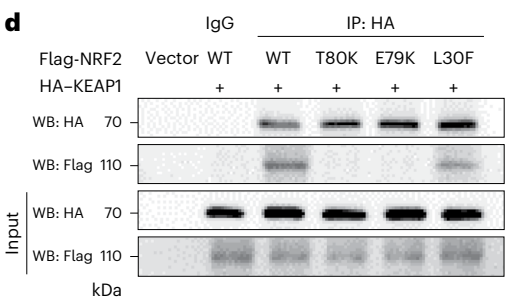
KEAP1 is an adapter of E3 ligase that senses oxidative stress by mediating degradation of NFE2L2/NRF2, a key transcription factor in multiple cancer types[72]. Patients with non-small cell lung cancer harboring PIONEER-predicted oncoPPI mutations on NRF2 have significantly worse survival than the WT ($P = 0.029$; Fig. 5c). KEAP1 recognizes NRF2

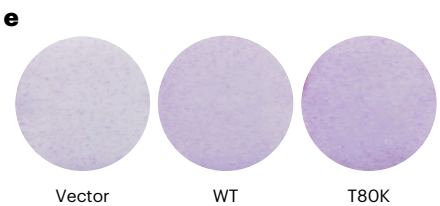
Fig. 5 | PIONEER-predicted PPI-perturbing tumor alleles in ubiquitination by E3 ligases. a, A landscape of six E3 complexes with PPI-perturbing mutations. The complex or single protein models from PDB or PIONEER modeling are shown. The protein in wheat denotes the E3 ligase, whereas the protein in blue denotes the specific substrate of E3 ligase. Interface mutations are denoted in green. **b**,**c**, Interface mutations of KEAP1–FBXL17 (**b**) and KEAP1–NRF2 (**c**) are significantly correlated with survival rate in non-small cell lung cancer. MUT, mutations. Significance was determined by two-sided log-rank test. The *n* numbers are shown in Supplementary Table 12. **d**, Experimental validation of mutation effects on p.Thr80Lys and p.Glu79Lys on NRF2 ETGE motif and p.Leu30Phe on NRF2 DLG motif on the interactions between KEAP1 and WT NRF2 was determined by co-immunoprecipitation with HEK293T cells. WB, western blotting; IP immunoprecipitation. The experiment was repeated three times independently. **e**,**f**, Colony formation assay of H1975 cells transfected with empty vectors and NRF2 (WT, p.Thr80Lys) expressing vectors. Data are represented as mean ± s.d. of three independent experiments. The dots indicate independent measurements. Significance was determined by two-tailed Student's *t*-test. **g**, Growth curves of H1975 cells transfected with empty vectors and NRF2 (p.Thr80Lys, WT) expressing vectors at day 4. Data are represented as mean ± s.d. of three independent experiments. Significance was determined by two-tailed Student's *t*-test.

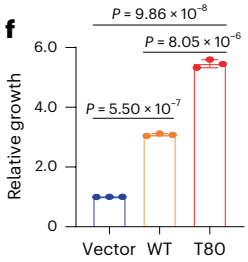
https://doi.org/10.1038/s41587-024-02428-4

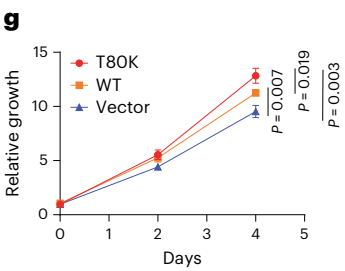












structurally through its conserved ETGE (amino acids 79–82) and DLG (amino acids 29–31) motifs[73,74]. We experimentally confirmed the association of NRF2 mutations and WT KEAP1 by co-immunoprecipitation. As shown in Fig. 5d, mutations p.Glu79Lys and p.Thr80Lys on NRF2 ETGE motif (Fig. 5a) reduce the binding of NRF2 to KEAP1, whereas mutation p.Leu30Phe on NRF2 DLG motif partially sustains the binding of NRF2 to KEAP1. The mutation p.Thr80Lys releases NRF2 from association with KEAP1 and protects NRF2 from ubiquitination and subsequent degradation. We next tested whether p.Thr80Lys on NRF2 contributes to the proliferation of non-small cell lung cancer cells. A pro-proliferative effect of p.Thr80Lys was observed in a colony formation assay (Fig. 5e,f). Overexpression of WT and p.Thr80Lys NRF2 promoted the growth of the non-small cell lung cancer H1975 cell lines harboring WT KEAP1 (Fig. 5g). In summary, PIONEER-predicted oncoPPI-perturbing tumor alleles that alter ubiquitination by E3 ligases are significantly associated with patient survival, drug responses and in vitro tumor growth.

Pharmacogenomic landscape of the PIONEER-predicted oncoPPIs

We next turned to inspect correlation between potential oncoPPIs and drug responses using high-throughput drug screening data (Extended Data Fig. 2a). The datasets include the drug pharmacogenomic profiles of more than 1,000 cancer cell lines and approximately 250 FDA-approved or clinically investigational agents from the Genomics of Drug Sensitivity in Cancer (GDSC) database and in vivo compound screens using approximately 1,000 PDX models to assess patient responses to 62 anti-cancer agents[75]. For each pair of oncoPPI and compound, the drug response characterization 50% inhibitory concentration (IC_{50}) vector was correlated with mutation status of the oncoPPIs using a linear ANOVA model. Extended Data Fig. 2b shows the landscape of the correlations between PPIs and 56 FDA-approved or clinically investigational anti-cancer drugs. In total, we identified 4,473 interface mutations that have significant correlations with drug sensitivity/resistance. Among the most significant correlations from PDX models, we found that PIONEER-predicted CDK6–BECN1 interface mutations are associated with resistance to treatment using a BYL719 plus encorafenib drug combination, whereas the mutations in PIONEER-predicted BRAF–MAP2K3 interfaces (for example, p.Val600Glu on BRAF and p.Arg152Gln on MAP2K3, both found in bladder urothelial carcinoma and glioblastoma) conferred significant drug sensitivity to encorafenib plus binimetinib treatment (Extended Data Fig. 2c). In addition, we found significant drug resistance to trastuzumab and BYL719 among those cases harboring mutations in PIONEER-predicted STK4–DDIT4L (for example, p.Arg181Gln on STK4) and ORC4–MTUS1 (for example, p.His166Tyr on ORC4) interfaces, respectively (Extended Data Fig. 2c). Taken together, PIONEER-predicted PPI interface mutations can significantly affect drug sensitivity/resistance in anti-tumor treatment using both cancer cell lines and PDX models (Supplementary Table 15).

Proteogenomic perturbation by PIONEER-informed interfaces

Recent proteogenomic study showed that somatic mutations altered protein or phosphoprotein abundance and further correlated with drug responses or survival in patients with cancer[76]. We next inspected whether PIONEER-predicted interface mutations more likely influence protein abundance in colon adenocarcinoma (COAD) and uterine corpus endometrial carcinoma (UCEC). The abundance of phosphoproteins was quantified using tandem mass tag (TMT) assays by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium. We found that PIONEER-predicted interface mutations significantly reduced phosphoprotein abundance in both COAD ($P = 0.018$) and UCEC ($P = 0.001$) (Extended Data Fig. 3a).

We next turned to inspect how the phosphorylation-associated PPI mutations identified by PIONEER perturb EGFR–RAS–RAF–MEK–ERK signaling networks in COAD (Extended Data Fig. 3b and Supplementary Table 18). The mutations involved in this signaling cascade have been suggested to regulate oncogenesis in colon and other cancers[77,78]. EGFR dimerization is activated by EGF in the extracellular domain (PDB: 3NJP and 2M20; Extended Data Fig. 3b)[79,80]. Binding of EGF triggers conformational changes in the C-terminal domain (PDB: 2GS6)[81] and results in autophosphorylation of specific tyrosine residues, such as Tyr1068 (ref. 82). The C-terminal domain of EGFR is essential for adapter protein binding to initiate signal transduction, such as by mediating GRB2/SOS1 (ref. 83). Via PIONEER, we identified that two mutations, p.Thr1021Ile and p.Thr1074Ile on EGFR C-terminal domain, may alter the phosphorylated PPI with downstream adapter protein of SOS1 (Extended Data Fig. 3b). SOS1 is a RAS activator that loads GTP (PDB: 6EPO)[84]. Its deficiency attenuates KRAS-induced leukemia in mouse model[85]. A selective SOS1–KRAS PPI inhibitor, BI 1701963, was developed for advanced KRAS-mutated solid tumors in a phase 1 clinical trial[86]. Using PIONEER, we identified two SOS1–KRAS PPI perturbation mutations: p.Tyr884His on SOS1 and p.Gln61His on KRAS (Extended Data Fig. 3b). Specifically, Tyr884 and Gln61 form strong hydrogen bond and cation-π interaction between KRAS and SOS1. We pinpointed that PIONEER-predicted SOS1–KRAS interface mutations are significantly related to trametinib resistance compared to WT group ($P = 7.6 \times 10^{-12}$; Extended Data Fig. 3c), offering potential pharmacogenomic biomarkers for trametinib (a MEK inhibitor) in KRAS-mutant colorectal cancer[87]. Binimetinib, another MEK-selective inhibitor[88], is significantly associated with resistance in PDX models harboring PIONEER-predicted SOS1–KRAS interface mutations ($P = 0.0044$; Extended Data Fig. 3c).

GTP-bound active RAS recruits RAF proteins (for example, RAF1 and BRAF) to the plasma membrane to orchestrate MAPK signaling[89]. Extended Data Fig. 3b shows PPIs of both KRAS–RAF1 and KRAS–BRAF constructed in one structure complex. Oncogenic mutations on KRAS, such as p.Gly12Val, p.Gly13Asp and p.Gln61Leu, are the most frequent mutations in common tumors[90]. PIONEER-predicted interface mutations of KRAS–RAF1, such as p.Arg59Ala and p.Asn64Ala on RAF1, are associated with significantly reduced binding affinity of the interaction[91] but not oncogenic mutations p.Gly12Val and p.Gly13Asp on KRAS (PDB: 6VJJ; Extended Data Fig. 3b). In addition, we identified that PIONEER-predicted KRAS–BRAF interface mutations are significantly associated with resistance of the MEK inhibitor refametinib[92] ($P = 4.7 \times 10^{-27}$; Extended Data Fig. 3c).

The key step for triggering the signaling cascade is that RAS-induced RAF dimerization subsequently phosphorylates MEK1/2 protein kinases[78]. Of RAF family members, BRAF shows the most potent activity[90], and the BRAF p.Val600Glu mutation confers a poor survival and prognosis in colorectal cancer[93,94]. Via PIONEER, we identified that two PPI interface mutations, p.Gly466Val and p.Asn581Ser on BRAF, may mediate how BRAF coordinates MEK1 by its C-lobes in the kinase domain (Extended Data Fig. 3b), consistent with a previous study[95]. Considering that the E3 ligase ITCH is also involved in BRAF regulation and binding to the kinase domain (Fig. 5a), we identified that PIONEER-predicted interface residue Val600 on BRAF may perturb interaction between BRAF and ITCH (Extended Data Fig. 3b). Phosphorylated MEK1 acts as upstream activators to phosphorylate ERK1/2 kinase activities in the MAPK cascade[96]. The PIONEER-predicted interface mutation p.Asp179Asn on ERK1 alters MEK1–ERK1 signaling network (Extended Data Fig. 3b)[97]. In summary, we showed that PIONEER-predicted oncoPPIs could characterize proteogenomic alterations in the EGFR–RAS–RAF–MEK–ERK signaling pathways in colon cancer and other cancer signaling pathways if broadly applied.

Construction of the PIONEER interactome web server

In total, our structurally informed interactomes cover all 282,095 experimentally determined binary interactions in the literature for

humans and seven model organisms, including 146,138 experimentally determined human interactions (Fig. 1a and Extended Data Fig. 1a). The web server is a user-friendly tool for genome-wide exploration through which users can browse multiscale structurally informed interactomes and identify functionally enriched areas in these networks (Supplementary Fig. 14). It provides rapid on-demand predictions for user-submitted interactions. Furthermore, our PIONEER web server also contains 161,244 disease-associated mutations across 10,564 disorders in HGMD and ClinVar[98] with their per-disease enrichment pre-computed on protein interaction interfaces with 3D spatial clustering at atomic (for interactions with structure models), residue and domain levels for all PPIs. By providing a user-friendly tool to visualize each protein and its given interactors with all available domain information, co-crystal structures, homology models and PIONEER-predicted interfaces coupled with all known disease mutation information, PIONEER seamlessly allows users to explore the effect of mutations on 3D structures. We think that the PIONEER web server will be instrumental in uncovering novel relationships among these mutations that help study disease mechanisms and develop personalized treatment in cancer or other diseases.

Discussion

Here we present a comprehensive, multiscale structurally informed interactome framework and web server, PIONEER, to combine seamlessly genomic-scale data with structural proteomic analyses. This resource is based on our ensemble deep learning framework, which accurately predicts partner-specific interaction interfaces for all PPIs in humans and seven model organisms. PIONEER outperforms other existing state-of-the-art methods, including our previously developed method, ECLAIR. Moreover, large-scale statistical analysis and mutagenesis experiments show that PIONEER-predicted interfaces reveal similar biological significance as those of known interfaces. Further analysis illustrates that PIONEER plays a pivotal role in dissecting the pathobiology of diseases: PIONEER-predicted interfaces are significantly enriched with both somatic cancer and germline disease mutations; and PIONEER-predicted interface mutations are highly correlated with survival of patients with cancer and anti-cancer responses in both tumor cell lines and PDX models. Our work is implemented as both a web server platform and a software package to facilitate systematic structural analysis in genomic studies, allowing the wider scientific community to adopt and further develop upon our PIONEER framework.

The experimentally determined binary human interactome is far from complete. Extensive efforts have been dedicated both experimentally (such as HuRI[99], BioPlex[100] and OpenCell[101]) and computationally (such as PrePPI[102] and HIGH-PPI[103]) to ascertaining which pairs of human proteins interact. As more protein interactions are detected for human proteins, PIONEER will be regularly updated to make interface predictions for newly released PPIs. In addition, the rate of growth of protein sequences in resources such as UniProt is much faster than that of protein structure resources such as PDB, ModBase[104] and AlphaFold2 database[13]. Even if our model can predict using solely the sequence, we have shown that the structural information greatly improves the performance. As such, a limitation is that PIONEER will not reach its full potential when the proteins do not have structural information. In the future, PIONEER's performance may be further improved. Specifically, PrePPI is a method that uses structural homology to make accurate PPI predictions. Although PrePPI does not make interface predictions (thus not comparable to PIONEER), the structural information obtained by PrePPI can be incorporated into our PIONEER pipeline as an additional feature to potentially improve our interface predictions. Furthermore, for sequence models, it could be useful to extract representations from protein language models[105]. For structural models, a promising area that is worth exploring is to develop a model that captures the geometric information of protein structures[106]. We envision that using deep learning architectures that implement geodesic invariance may improve the performance.

With rapid advances in sequencing technologies and a large number of ongoing genome/exome sequencing projects, including TCGA, cardiovascular medicine (that is, the National Heart, Lung and Blood Institute's Trans-Omics for Precision Medicine program[107]) and Alzheimer's disease sequencing project[108], we expect that our comprehensive structurally informed interactomes generated by PIONEER will help bridge the gap between genome-scale data and structural proteomic analyses. With the high-quality and comprehensive map of protein interfaces, there are numerous valuable extensions considering the biophysical effects induced by mutations in protein interfaces, such as the investigation of disease etiology and the corresponding drug prioritization and prediction of specific disease pathobiology. The partner-specific property of PIONEER-generated structurally informed interactomes also allows us to study the pleiotropic effects of genes. Therefore, the powerful and comprehensive PIONEER framework will make such extensive research possible and, more importantly, provide potentially unforeseen avenues for drug design and therapeutics.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-024-02428-4.

References

1. Nussinov, R., Jang, H., Nir, G., Tsai, C. J. & Cheng, F. Open structural data in precision medicine. *Annu. Rev. Biomed. Data Sci.* **5**, 95–117 (2022).
2. Braberg, H., Echeverria, I., Kaake, R. M., Sali, A. & Krogan, N. J. From systems to structure—using genetic data to model protein structures. *Nat. Rev. Genet.* **23**, 342–354 (2022).
3. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
4. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
5. Meyer, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107–114 (2018).
6. Wang, X. et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
7. Cheng, F. et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353 (2021).
8. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
9. Wierbowski, S. D. et al. A 3D structural SARS-CoV-2-human interactome to explore genetic and drug perturbations. *Nat. Methods* **18**, 1477–1488 (2021).
10. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2022).
11. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).
12. Burke, D. F. et al. Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023).
13. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

14. Bianchi, F. M., Grattarola, D., Livi, L. & Alippi, C. Graph neural networks with convolutional ARMA filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3496–3507 (2022).
15. Cho, K. et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) 1724–1734 (Association for Computational Linguistics, 2014).
16. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. of the IEEE* **109**, 43–76 (2021).
17. Krapp, L. F., Abriata, L. A., Cortes Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat. Commun.* **14**, 2175 (2023).
18. Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* **19**, 730–739 (2022).
19. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
20. Sanchez-Garcia, R., Macias, J. R., Sorzano, C. O. S., Carazo, J. M. & Segura, J. BIPSPI+: mining type-specific datasets of protein complexes to improve protein binding site prediction. *J. Mol. Biol.* **434**, 167556 (2022).
21. Zeng, M. et al. Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* **36**, 1114–1120 (2020).
22. Townshend, R. J. L., Bedi, R., Suriana, P. A. & Dror, R. O. End-to-end learning on 3D protein structure for interface prediction. *33rd Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2019/file/6c7de1f27f7de61a6daddfffbe05c058-Paper.pdf (NeurIPS, 2019).
23. Fout, A., Byrd, J., Shariat, B. & Ben-Hur, A. Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems 30*. https://papers.nips.cc/paper_files/paper/2017/file/f507783927f2ec2737ba40afbd17efb5-Paper.pdf (NIPS, 2017).
24. Lensink, M. F. & Wodak, S. J. Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins* **82**, 3163–3169 (2014).
25. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
26. Das, J. & Yu, H. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
27. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
28. Salwinski, L. et al. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
29. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
30. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
31. Turner, B. et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023 (2010).
32. Keshava Prasad, T. S. et al. Human protein reference database—2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
33. Mewes, H. W. et al. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220–D224 (2011).
34. Nelson, L. & Cox, M. *Lehninger Principles of Biochemistry* 7th edn (W.H. Freeman, 2017).
35. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834–838 (1985).
36. Aftabuddin, M. & Kundu, S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.* **93**, 225–231 (2007).
37. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64 (1997).
38. Ansari, S. & Helms, V. Statistical analysis of predominantly transient protein–protein interfaces. *Proteins* **61**, 344–355 (2005).
39. Burley, S. K. et al. RCSB Protein Data Bank: celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. *Protein Sci.* **31**, 187–208 (2022).
40. Wei, X. et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
41. Xiong, D., Lee, D., Li, L., Zhao, Q. & Yu, H. Implications of disease-related mutations at protein–protein interfaces. *Curr. Opin. Struct. Biol.* **72**, 219–225 (2022).
42. Stenson, P. D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
43. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
44. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
45. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
46. Schymkowitz, J. et al. The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).
47. Zhou, Y. et al. A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19. *PLoS Biol.* **18**, e3000970 (2020).
48. Plasilova, M. et al. Homozygous missense mutation in the lamin A/C gene causes autosomal recessive Hutchinson–Gilford progeria syndrome. *J. Med. Genet.* **41**, 609–614 (2004).
49. Favretto, F. et al. The molecular basis of the interaction of cyclophilin A with α-synuclein. *Angew. Chem. Int. Ed.* **59**, 5643–5646 (2020).
50. Liu, Q. et al. HIF2A germline–mutation-induced polycythemia in a patient with VHL-associated renal-cell carcinoma. *Cancer Biol. Ther.* **18**, 944–947 (2017).
51. Tarade, D., Robinson, C. M., Lee, J. E. & Ohh, M. HIF-2α-pVHL complex reveals broad genotype-phenotype correlations in HIF-2α-driven disease. *Nat. Commun.* **9**, 3359 (2018).
52. V, F. R. L. et al. Three novel *EPAS1*/*HIF2A* somatic and germline mutations associated with polycythemia and pheochromocytoma/paraganglioma. *Blood* **120**, 2080 (2012).
53. Chang, K. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
54. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* **185**, 563–575 (2022).
55. Rabara, D. et al. KRAS G13D sensitivity to neurofibromin-mediated GTP hydrolysis. *Proc. Natl Acad. Sci. USA* **116**, 22122–22131 (2019).
56. Wang, Z. et al. The diverse roles of SPOP in prostate cancer and kidney cancer. *Nat. Rev. Urol.* **17**, 339–350 (2020).
57. Song, Y. et al. The emerging role of SPOP protein in tumorigenesis and cancer therapy. *Mol. Cancer* **19**, 2 (2020).
58. Xu, J. & Lin, D. I. Oncogenic c-terminal cyclin D1 (*CCND1*) mutations are enriched in endometrioid endometrial adenocarcinomas. *PLoS ONE* **13**, e0199688 (2018).

59. Ryu, D. et al. Alterations in the transcriptional programs of myeloma cells and the microenvironment during extramedullary progression affect proliferation and immune evasion. *Clin. Cancer Res.* **26**, 935–944 (2020).
60. Zhang, M. et al. CanProVar 2.0: an updated database of human cancer proteome variation. *J. Proteome Res.* **16**, 421–432 (2017).
61. Mészáros, B., Kumar, M., Gibson, T. J., Uyar, B. & Dosztányi, Z. Degrons in cancer. *Sci. Signal.* **10**, eaak9982 (2017).
62. Yang, Q., Zhao, J., Chen, D. & Wang, Y. E3 ubiquitin ligases: styles, structures and functions. *Mol. Biomed.* **2**, 23 (2021).
63. Senft, D., Qi, J. & Ronai, Z. E. A. Ubiquitin ligases in oncogenic transformation and cancer therapy. *Nat. Rev. Cancer* **18**, 69–88 (2018).
64. Han, Y., Lee, H., Park, J. C. & Yi, G. S. E3Net: a system for exploring E3-mediated regulatory networks of cellular functions. *Mol. Cell. Proteomics* **11**, O111.014076 (2012).
65. Li, Z. et al. UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions. *Database* **2021**, baab010 (2021).
66. Mena, E. L. et al. Dimerization quality control ensures neuronal development and survival. *Science* **362**, eaap8236 (2018).
67. Wang, Q. et al. Alterations of anaphase-promoting complex genes in human colon cancer cells. *Oncogene* **22**, 1486–1490 (2003).
68. Yin, Q., Wyatt, C. J., Han, T., Smalley, K. S. M. & Wan, L. ITCH as a potential therapeutic target in human cancers. *Semin. Cancer Biol.* **67**, 117–130 (2020).
69. Li, L. et al. CHIP mediates degradation of Smad proteins and potentially regulates Smad-induced transcription. *Mol. Cell. Biol.* **24**, 856–864 (2004).
70. Tsai, W.-W. et al. TRIM24 links a non-canonical histone signature to breast cancer. *Nature* **468**, 927–932 (2010).
71. Lv, D. et al. TRIM24 is an oncogenic transcriptional co-activator of STAT3 in glioblastoma. *Nat. Commun.* **8**, 1454 (2017).
72. Cuadrado, A. et al. Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. *Nat. Rev. Drug Discov.* **18**, 295–317 (2019).
73. Furukawa, M. & Xiong, Y. BTB protein Keap1 targets antioxidant transcription factor Nrf2 for ubiquitination by the Cullin 3-Roc1 ligase. *Mol. Cell. Biol.* **25**, 162–171 (2005).
74. Fukutomi, T., Takagi, K., Mizushima, T., Ohuchi, N. & Yamamoto, M. Kinetic, thermodynamic, and structural characterizations of the association between Nrf2-DLGex Degron and Keap1. *Mol. Cell. Biol.* **34**, 832–846 (2014).
75. Gao, H. et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
76. Vasaikar, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049.e19 (2019).
77. Abi-Habib, R. J. et al. BRAF status and mitogen-activated protein/extracellular signal-regulated kinase kinase 1/2 activity indicate sensitivity of melanoma cells to anthrax lethal toxin. *Mol. Cancer Ther.* **4**, 1303–1310 (2005).
78. Roberts, P. J. & Der, C. J. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **26**, 3291–3310 (2007).
79. Endres, N. F. et al. Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell* **152**, 543–556 (2013).
80. Lu, C. F. et al. Structural evidence for loose linkage between ligand binding and kinase activation in the epidermal growth factor receptor. *Mol. Cell. Biol.* **30**, 5432–5443 (2010).
81. Liang, S. I. et al. Phosphorylated EGFR dimers are not sufficient to activate ras. *Cell Rep.* **22**, 2593–2600 (2018).
82. Bishayee, A., Beguinot, L. & Bishayee, S. Phosphorylation of tyrosine 992, 1068, and 1086 is required for conformational change of the human epidermal growth factor receptor C-terminal tail. *Mol. Biol. Cell.* **10**, 525–536 (1999).
83. Siegelin, M. D. & Borczuk, A. C. Epidermal growth factor receptor mutations in lung adenocarcinoma. *Lab Invest.* **94**, 129–137 (2014).
84. Hillig, R. C. et al. Discovery of potent SOS1 inhibitors that block RAS activation via disruption of the RAS–SOS1 interaction. *Proc. Natl Acad. Sci. USA* **116**, 2551–2560 (2019).
85. You, X. et al. Unique dependence on Sos1 in $Kras^{G12D}$-induced leukemogenesis. *Blood* **132**, 2575–2579 (2018).
86. Hofmann, M. H. et al. Trial in process: phase 1 studies of BI 1701963, a SOS1::KRAS inhibitor, in combination with MEK inhibitors, irreversible KRASG12C inhibitors or irinotecan. *Cancer Res.* **81**, CT210 (2021).
87. Huijberts, S. C. F. A. et al. Phase I study of lapatinib plus trametinib in patients with *KRAS*-mutant colorectal, non-small cell lung, and pancreatic cancer. *Cancer Chemother. Pharmacol.* **85**, 917–930 (2020).
88. Cho, M. et al. A phase I clinical trial of binimetinib in combination with FOLFOX in patients with advanced metastatic colorectal cancer who failed prior standard therapy. *Oncotarget* **8**, 79750–79760 (2017).
89. Hofmann, M. H. et al. BI-3406, a potent and selective SOS1–KRAS interaction inhibitor, is effective in KRAS-driven cancers through combined MEK inhibition. *Cancer Discov.* **11**, 142–157 (2021).
90. Liu, F., Yang, X., Geng, M. & Huang, M. Targeting ERK, an Achilles' Heel of the MAPK pathway, in cancer therapy. *Acta Pharm. Sin. B* **8**, 552–562 (2018).
91. Tran, T. H. et al. KRAS interaction with RAF1 RAS-binding domain and cysteine-rich domain provides insights into RAS-mediated RAF activation. *Nat. Commun.* **12**, 1176 (2021).
92. Patelli, G. et al. Strategies to tackle *RAS*-mutated metastatic colorectal cancer. *ESMO Open* **6**, 100156 (2021).
93. Li, Z.-N., Zhao, L., Yu, L.-F. & Wei, M.-J. *BRAF* and *KRAS* mutations in metastatic colorectal cancer: future perspectives for personalized therapy. *Gastroenterol. Rep.* **8**, 192–205 (2020).
94. Corcoran, R. B. et al. Combined BRAF, EGFR, and MEK inhibition in patients with $BRAF^{V600E}$-mutant colorectal cancer. *Cancer Discov.* **8**, 428–443 (2018).
95. Lin, Q. et al. The association between *BRAF* mutation class and clinical features in *BRAF*-mutant Chinese non-small cell lung cancer patients. *J. Transl. Med.* **17**, 298 (2019).
96. Caunt, C. J., Sale, M. J., Smith, P. D. & Cook, S. J. MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nat. Rev. Cancer* **15**, 577–592 (2015).
97. Huang, K. L. et al. Regulated phosphosignaling associated with breast cancer subtypes and druggability. *Mol. Cell. Proteomics* **18**, 1630–1650 (2019).
98. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
99. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
100. Huttlin, E. L. et al. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
101. Cho, N. H. et al. OpenCell: endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).
102. Petrey, D., Zhao, H., Trudeau, S. J., Murray, D. & Honig, B. PrePPI: a structure informed proteome-wide database of protein–protein interactions. *J. Mol. Biol.* **435**, 168052 (2023).
103. Gao, Z. et al. Hierarchical graph learning for protein–protein interaction. *Nat. Commun.* **14**, 1093 (2023).
104. Pieper, U. et al. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336–D346 (2014).

105. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
106. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. *The Twelfth International Conference on Learning Representations*. https://openreview.net/pdf?id=6MRm3G4NiU (ICLR, 2023).
107. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
108. Gary, W. B. et al. The Alzheimer's disease sequencing project: study design and sample selection. *Neurol. Genet.* **3**, e194 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2024

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [3]Center for Innovative Proteomics, Cornell University, Ithaca, NY, USA. [4]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [5]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [6]Department of Systems Biology, Herbert Irving Comprehensive Center, Columbia University, New York, NY, USA. [7]Biophysics Program, Cornell University, Ithaca, NY, USA. [8]Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China. [9]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA. [10]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA. [11]Channing Division of Network Medicine, Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [12]These authors contributed equally: Dapeng Xiong, Yunguang Qiu, Junfei Zhao, Yadi Zhou, Dongjin Lee. ✉e-mail: chengf@ccf.org; haiyuan.yu@cornell.edu

Methods

PPI interface data construction

We compiled 282,095 binary interactions for *Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Schizosaccharomyces pombe* and *Escherichia coli* in total, including 9,123 full experimentally determined binary interactions in humans. The interactions with known co-crystal structures in the PDB were used to form the training, validation and testing datasets to build PIONEER models. The homologous structures of PPIs that do not have co-crystal structures were collected from Interactome3D[109].

We calculated the partner-specific interface residues for those interactions with known co-crystal structures in the PDB. SIFTS[110] was then used to map the UniProt-indexed residues to the PDB-indexed residues. To determine the interface residues, we used NACCESS[111] to assess the change in SASA of the protein in complex and in isolation. Specifically, an interface residue is defined as a residue that is a surface residue (≥15% exposed surface) with its relative SASA decreasing by ≥1.0 Å^2 in the complex. We reviewed all available structures in the PDB for each interaction and considered a residue to be in the interface of that interaction if it had been calculated to be an interface residue in at least one of the corresponding co-crystal structures. We built the training, validation and independent benchmark testing datasets by only considering interactions for which aggregated co-crystal structures have been combined to cover at least 30% of the UniProt residues for both interacting proteins. These datasets include a random selection of 2,615, 400 and 400 interactions with known co-crystal structures, corresponding to 1,191,036, 174,739 and 186,326 residues for sufficient model training, validation and testing, respectively. The number of positive residues (interface residues) compared to negative residues (non-interface residues) in this dataset is 175,911 of 1,015,125 (17.3%), 25,641 of 149,098 (17.2%) and 27,744 of 158,582 (17.5%), respectively. It is important to note that a single residue may be labeled as positive for a specific interaction while being labeled negative for other interactions. In fact, this is the case for 58.5% of all interface residues in our study, where 86.6% of the proteins have more than one partner. Additionally, we ensured that no homologous interactions or repeated proteins existed between any of the two datasets to guarantee the robustness and generalizability of our models and a fair performance evaluation. We define homologous interactions as a pair of interactions where both proteins in one interaction are homologs of both proteins in the other interaction. Pandas (https://pandas.pydata.org) and Numpy (https://numpy.org) were used for data processing in our work, and Numba (https://numba.pydata.org) was used to speed up Numpy-based numerical functions using standard Python (https://www.python.org) programming language. Three iterations of PSIBLAST[112] with an E-value cutoff 0.001 were carried out to determine the protein homologs.

Feature characterization

Our previous pipeline, ECLAIR, employed a set of representative feature groups to describe the residues, including biophysical residue properties, evolutionary sequence conservation, co-evolution, relative SASA and docking-based metrics. While retaining all features from ECLAIR here, we also implemented two new feature groups to seek a more comprehensive and in-depth feature characterization (Supplementary Methods). Clustal Omega[113] was used for multiple sequence alignment (MSA) when calculating evolutionary sequence conservation and co-evolution. CD-HIT[114] was used to cluster protein sequences to remove the redundancy of MSA for co-evolution, with all UniProt sequences serving as the search database for MSA generation. ZDOCK[115] was used for the protein–protein docking. From each feature group, we synthesized a variation of features using scaling, by which we mean that each feature used its raw calculated values and normalized values against the average of all positions per protein.

Model building

To ensure that every residue is meticulously predicted through the maximal amount of available information from both proteins in an interaction pair, we built four deep learning models in which each model takes different interactions as input based on the availability of structures.

1. Structure–Structure model (Fig. 1c and Supplementary Figs. 1a and 2): For interactions where both proteins have structural information available, the structure and sequence information were embedded through GCNs with ARMA filters and bidirectional RNNs with GRUs, respectively. Specifically, GCN uses the structural information from graph representations of protein structures where each node represents a residue and each edge signifies that two residues are adjacent. For each node, GCN incorporates its spatial neighborhood information to generate a more comprehensive residue representation, whereas RNN explores amino acid sequences to include the sequential neighborhood information of each residue. The RNN extracts the upstream and downstream sequence information from each residue. Through the concatenation and mean aggregation, the residue embeddings of both target protein and partner protein were then converted to protein embeddings, respectively. Finally, the residue embeddings, target protein embedding and partner protein embedding were concatenated and fed into the fully connected layers to make prediction for each residue in the target protein.
2. Sequence–Sequence model (Fig. 1d and Supplementary Figs. 1b and 2b): For interactions where neither protein has structural information, the sequence information of both proteins was fed into the RNNs. Next, in a manner similar to that described in the Structure–Structure model, the residue embeddings, target protein embedding and partner protein embedding were concatenated and fed into the fully connected layers to make prediction for each residue in the target protein.
3. Structure–Sequence and Sequence–Structure models (Fig. 1e,f and Supplementary Figs. 1c,d and 2): The use of Structure–Sequence or Sequence–Structure model depends on whether target protein or partner protein has structural information, respectively. Transfer learning was used in these two models, which means that the pre-trained GCNs and RNNs in the above Structure–Structure model and RNNs in the above Sequence–Sequence model were deployed in Structure–Sequence model and Sequence–Structure model for the processing of proteins with and without structural information, respectively. Subsequently, in a manner similar to that described in Structure–Structure model and Sequence–Sequence model, the residue embeddings, target protein embedding and partner protein embedding were concatenated and fed into the fully connected layers to make prediction for each residue in the target protein.

We compiled a set of representative protein structures from the PDB, ModBase and AlphaFold2 database for each protein. For ModBase models, we only consider the models with a ModPipe Quality Score (MPQS) ≥ 1.1. The PDB structures have the highest priority, whereas the AlphaFold2-predicted structures are the lowest. The structures were then sorted by the coverage of UniProt residues based on SIFTS, excluding any homologous PDB structures of interacting protein pairs. Each residue in a target protein was then reviewed if it has structural information; if so, it was predicted using that protein's first corresponding structure that contains the structural information of that residue; otherwise, it was predicted using the sequence information. For the partner protein that has structural information, we only used the corresponding structure with the highest UniProt coverage. In particular, if a protein has multiple structures with identical coverage available, these structures were sorted by their qualities (for example,

PDB resolution and MPQS). To make our tool more practically useful and to avoid the memorization of known interfaces, we use the single protein structure that is not from co-crystal or homologous co-complex structures to train the model.

Our PIONEER framework was implemented using PyTorch (https://pytorch.org); the GCN was written based on torch-geometric (https://pytorch-geometric.readthedocs.io). To maximize model performance, we carried out comprehensive hyperparameter optimization for the neural network architectures, and the optimal set of hyperparameters was determined by maximizing area under the receiver operating characteristic (AUROC) curve on the validation set. All four models were trained with cross-entropy loss and the Adam optimizer; the kernel activation function[116] was used in GCNs and fully connected layers. The hyperparameters used for these four models can be found in our accompanying PIONEER software package. To solve the variable length inputs, we trained all four models in a mini-batch mode with only a single protein pair.

Performance evaluation

After identifying the best hyperparameters for each model, a thorough examination was performed using the benchmark testing set. Models were ordered based on their AUROCs on the validation set, which means the priorities of models are Structure–Structure, Structure-Sequence, Sequence-Structure and Sequence–Sequence, respectively. For the overall performance, the raw prediction score of each residue was taken from the results of the model with highest priority according to the availability of structures of the target protein containing that residue and its partner protein. We further compared PIONEER with numerous existing state-of-the-art methods, including ECLAIR, PeSTo, ScanNet, BIPSPI+, MaSIF-site, DeepPPISP, SASNet, PIPGCN, DELPHI[117], SCRIBER[118] and DLPred[119]. We also reported performance metrics at various discrete and comparable levels of confidence, which consist of Very low, Low, Medium, High and Very high prediction categories, by evenly separating into fifths our raw prediction scores.

Interface prediction

By further incorporating AlphaFold2-predicted structures, we predicted interface residues for the remaining 256,946 interactions not resolved by either PDB structures or homology models. Each residue was then predicted by the model of the ensemble with the highest priority according to the availability of structures of the target protein containing that residue and its partner protein.

Mutagenesis validation experiments

We performed mutagenesis experiments where we introduced random human population variants from the Exome Sequencing Project[120] into predicted interfaces, known interfaces and non-interfaces. We randomly selected mutations of predicted interfaces in each of the PIONEER prediction categories (from Very low to Very high). We also selected random mutations of known interfaces and non-interfaces in co-crystal structures in the PDB as positive and negative controls. The selected mutations were introduced into the proteins according to our Clone-seq pipeline[40]. We generated 2,395 mutations on 1,141 proteins and examined their impact on 6,754 mutation interaction pairs (either disrupting or maintaining the interactions) using our high-throughput Y2H assay.

Y2H assay

Y2H was performed as previously described[5]. Gateway LR reactions were used to transfer all WT/mutant clones into our Y2H pDEST-AD and pDEST-DB vectors. All DB-X and AD-Y plasmids were transformed into the Y2H strains MATα Y8930 and MATa Y8800, respectively. Thereafter, each of the DB-X MATα transformants (WT and mutants) was mated with corresponding AD-Y MATa transformants (WT and mutants) individually through automated 96-well procedures, including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30 °C) and replica plating onto selective Synthetic Complete media lacking histidine, leucine and tryptophan and supplemented with 1 mM 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg L^{-1} cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the LYS2::GAL1–HIS3 and GAL2–ADE2 reporter genes. The plates containing cycloheximide were used to select for cells that do not have the AD plasmid due to plasmid shuffling. Spontaneous auto-activators[121], therefore, were identified by growth on these control plates. These plates were incubated overnight at 30 °C and 'replica cleaned' the following day. Subsequently, plates were incubated for three more days, after which positive colonies were scored as those that grow on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. Disruption of an interaction by a mutation was defined as at least 50% reduction of growth consistently across both reporter genes when compared to Y2H phenotypes of the corresponding WT allele as benchmarked by two-fold serial dilution experiments. All Y2H experiments were repeated three times.

Co-immunoprecipitation

The first co-immunoprecipitation assay was conducted to validate the PIONEER partner-specific interface prediction. In specific, HEK293T cells were maintained in DMEM medium supplemented 10% FBS. Cells were seeded onto 10-cm dishes and incubated until 40–50% confluency and were transfected with a mixed solution of 3 µg of bait construct (CCND1), 3 µg of prey construct (CDK4 or TSC2), 30 µl of 1 mg ml^{-1} PEI (Polysciences, cat. no. 23966) and 1.2 ml of Opti-MEM (Gibco, cat. no. 31085-062). After 48-h incubation, transfected cells were washed three times in 10 ml of DPBS (VWR, cat. no. 14190144), resuspended in 500 µl of NP-40 lysis buffer (50 mM Tris, pH 7.5, 150 mM NaCl, 5 mM EDTA, 1.0% NP-40) and incubated on ice for 30 min. Whole lysate was sonicated on a sonifier cell disruptor (Branson, cat. no. 500-220-180) for 120 s at 40% amplitude. Extracts were cleared by centrifugation for 15 min at 16,100*g* at 4 °C. For co-immunoprecipitation, 500 µl of cell lysate per sample reaction was incubated with 15 µl of EZview Red Anti-FLAG M2 Affinity Gel (Sigma-Aldrich, cat. no. F2426) overnight 4 °C with a nutator. After incubation, bound proteins were washed three times in NP-40 lysis buffer and then eluted in 200 µl of elution buffer (10 mM Tris-Cl, pH 8.0, 1% SDS) at 65 °C for 15 min. FLAG-co-purified samples were run on 8% SDS-PAGE gel, and the proteins were transferred to PVDF membranes. Anti-FLAG (Sigma-Aldrich, cat. no. F1804) and anti-MYC (Invitrogen, cat. no. 132500) at both 1:5,000 dilutions were used for immunoblotting analysis.

We also validated mutation effects for KEAP1–NRF2 by co-immunoprecipitation assay, in which HEK293T cells were co-transfected with KEP1 (WT) expressing vector and NRF2 (WT, p.Thr80Lys, p.Glu79Lys or p.Leu30Phe) expressing vectors for 48 h. Cells were lysed with NP-40 lysis buffer (Beyotime, cat. no. P0013F) on ice, and supernatants were incubated with anti-HA antibody (Abmart, cat. no. M20003) coupled with protein A/G beads (Santa Cruz Biotechnology, cat. no. sc-2003) overnight. Immunoprecipitated complexes were washed with NP-40 lysis buffer for three times and were then eluted and subjected to western blotting.

Collection and preparation of genome sequencing data

We collected variant data across multiple sources, including TCGA, MSK-MET, 1KGP, ExAC, HGMD, Cancer Cell Line Encyclopedia and genomic profiling of PDXs from a previous study[75]. For unannotated datasets, we used VEP[122] to annotate these variants to identify the corresponding amino acid changes. We regarded one PPI as mutated if one variant affects the amino acid residue in the interfaces of either protein involved in the interaction.

Significance determination of PPI interface mutations

The significance of PPI interface mutations were tested using the method as described in our previous study[7]. A PPI in which there is significant enrichment in interface mutations in one or the other of the two protein-binding partners across individuals will be defined as an oncoPPI. For each gene g_i and its PPI interfaces, we assume that the observed number of mutations for a given interface follows a binomial distribution, binomial (T, p_{g_i}), in which T is the total number of mutations observed in one gene, and p_{g_i} is the estimated mutation rate for the region of interest under the null hypothesis that the region was not recurrently mutated. Using length(g_i) to represent the length of the protein product of gene g_i, for each interface, we computed the P value—the probability of observing >k mutations around this interface out of T total mutations observed in this gene—using the following equations:

$$P(X \geq k) = 1 - P(X < k) = 1 - \sum_{x=0}^{k-1} \left(\frac{T}{x}\right) p_{g_i}^{x} (1 - p_{g_i})^{T-x}$$

$$p_{g_i} = \frac{length\ of\ interface}{length(\ g_i)}$$

Finally, we set the minimal P value across all the interfaces in a specific protein as the representative P value of its coding gene g_i, denoted $P(g_i)$. The significance of each PPI is defined as the product of P values of the two proteins (gene products). All P values were adjusted for multiple testing using the Bonferroni correction.

PPI system construction of E3 ligases

In total, 355 E3 ubiquitin ligases were retrieved and merged from E3Net and UbiNet2.0, and 4,613 E3 ubiquitin ligase-associated oncoPPIs were analyzed after removing PPIs with homodimers or without gene name. These oncoPPIs include 198 oncoPPIs from the PDB database, 197 from homology models and 4,218 from PIONEER. The correlations between mutations in these oncoPPIs and anti-cancer drug responses in TCGA cell lines, PDX models and cancer survival rates from TCGA and MSK MetTropism datasets were then calculated (Supplementary Table 15).

Complex crystal structures (PDB: 3O37, 4O1V, 5G05, 5VZU, 6WCQ and 7K2M) were accessed from the PDB. The structures in the complex without co-crystal structures were retrieved from the AlphaFold2 database. PIONEER-predicted PPI models were constructed using HADDOCK[123]. The names, mutations and PDB IDs are shown in Supplementary Table 14.

The linear ANOVA model

We used the drug response data of human cancer cell lines from GDSC datasets and to investigate the association of PPI interface mutation with drug response. For each drug, a drug response vector consisting of IC_{50} values was modeled using the status of a genomic feature (whether a PPI interface is mutated), the tissue of origin of the cell lines, screening medium and growth properties by fitting a linear model. A genomic feature–drug pair was tested only if the final IC_{50} vector contained at least 10 positive cell lines. The effect size was quantified through Cohen's d statistic using the difference between two means divided by a pooled standard deviation for the data. The resulting P values were corrected by the Benjamini–Hochberg method. Similar to cell line drug response analysis, we also used the drug response data from high-throughput screening using PDX models to study the association of PPI interface mutation with drug response using linear model. All statistical analyses were performed using the R package (http://www.r-project.org).

Cell viability assay

H1975 cells were transfected with NRF2 (WT, p.Thr80Lys) expressing vectors or empty vectors using Lipofectamine 3000 (Thermo Fisher Scientific, cat. no. L3000001). For growth curve measurement, 3,000 cells were planted into 96-well plates, and viability was measured using CellTiter 96 AQueous MTS Reagent (Promega, cat. no. G1111) at days 0, 2 and 4.

Colony formation assay

For the colony formation assay, H1975 cells were seeded into six-well plates (2,000 cells per well). After 2 weeks, cells were fixed with 4% paraformaldehyde and stained with crystal violet. The relative growth index was analyzed using ImageJ[124].

Construction of EGFR–RAS–RAF–MEK–ERK signaling network

EGR–EGFR complex was constructed by three crystal structures (PDB: 3NJP, 2M20 and 2GS6). Membrane models were built by CHARMM-GUI[125]. SOS1–KRAS complex (PDB: 6EPO), KRAS–RAF1 complex (PDB: 6VJJ), MEK1–BRAF complex (PDB: 6Q0J), MEK1 (PDB: 3WIG) and ERK1 (PDB: 4QTB) were accessed from the PDB. Two subunits of RAF proteins are represented by RAF1 and BRAF, separately. The ITCH–BRAF complex model was generated using HADDOCK. All images were processed using PyMOL (https://www.pymol.org). The complex names, mutations and PDB IDs are shown in Supplementary Table 18.

Web server development

The PIONEER web server was developed using modern web development tools and frameworks. The details are described in the Supplementary Methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Mutation data from the TCGA study were downloaded from the National Cancer Institute's Genomic Data Commons (https://portal.gdc.cancer.gov). The MSK MetTropism dataset was downloaded from the cBioPortal (https://www.cbioportal.org/study/summary?id=msk_met_2021). Variant data from the 1000 Genomes Project were downloaded from the National Center for Biotechnology Information's FTP site (https://ftp-trace.ncbi.nih.gov/1000genomes/ftp). The ExAC dataset was downloaded from the Genome Aggregation Database (https://gnomad.broadinstitute.org/downloads#exac-variants). Variants collected by the HGMD were downloaded from https://www.hgmd.cf.ac.uk/ac/index.php. Genomic variants and drug response data of human cancer cell lines were downloaded from GDSC datasets (https://www.cancerrxgene.org/downloads/bulk_download). Genomic profiling of PDXs and drug response curve metrics of PDX clinical trials were downloaded from Supplementary Table 1 of the corresponding paper (https://www.nature.com/articles/nm.3954#Sec28). The homologous structures of PPIs that do not have co-crystal structures were collected from Interactome3D (https://interactome3d.irbbarcelona.org). The ModBase data were downloaded from https://modbase.compbio.ucsf.edu. The PDB data were downloaded from the PDB FTP site (https://files.wwpdb.org/pub/pdb/data/structures/divided/pdb). The AlphaFold2-predicted protein structures were download from the AlphaFold2 database (https://alphafold.ebi.ac.uk). All other data supporting the results in this study are available in the supplementary materials and at https://pioneer.yulab.org. Source data are provided with this paper.

Code availability

The source code of PIONEER is available at GitHub[126].

References

109. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).

110. Velankar, S. et al. SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.* **41**, D483–D489 (2013).
111. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400 (1971).
112. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
113. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
114. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
115. Pierce, B. G. et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
116. Scardapane, S., Van Vaerenbergh, S., Totaro, S. & Uncini, A. Kafnets: kernel-based non-parametric activation functions for neural networks. *Neural Netw.* **110**, 19–32 (2019).
117. Li, Y., Golding, G. B. & Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* **37**, 896–904 (2021).
118. Zhang, J. & Kurgan, L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353 (2019).
119. Zhang, B., Li, J., Quan, L., Chen, Y. & Lü, Q. Sequence-based prediction of protein–protein interaction sites by simplified long short-term memory network. *Neurocomputing* **357**, 86–100 (2019).
120. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
121. Walhout, A. J. M. & Vidal, M. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24**, 297–306 (2001).
122. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
123. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
124. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
125. Wu, E. L. et al. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J. Comput. Chem.* **35**, 1997–2004 (2014).
126. Xiong, D., Lee, D. & Liang, S. GitHub code repository for PIONEER. https://github.com/hyulab/PIONEER (2024).

Acknowledgements

This work was supported by the National Institute of General Medical Sciences (R01GM124559, R01GM125639, R01GM130885 and RM1GM139738) and the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK115398) to H.Y.; the National Institute on Aging (R01AG084250, R56AG074001, U01AG073323, R01AG066707, R01AG076448, R01AG082118, RF1AG082211 and R21AG083003) and the National Institute of Neurological Disorders and Stroke (RF1NS133812) to F.C.; and the National Human Genome Research Institute (U01HG007691), the National Heart, Lung, and Blood Institute (R01HL155107, R01HL155096, R01HL166137 and U54HL119145), the American Heart Association (AHA957729 and 24MERIT1185447) and European Union Horizon Health 2021 (101057619) to J.L. C.E. is the Sondra J. and Stephen R. Hardis Chair of Cancer Genomic Medicine at the Cleveland Clinic. This work partially used Jetstream2 at Indiana University through allocation BIO220060 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support program, which is supported by the National Science Foundation (2138259, 2138286, 2138307, 2137603 and 2138296).

Author contributions

D.X., Y.Q., J.Z., Y.Z., D.L., F.C. and H.Y. conceived and developed the project. Under close supervision of H.Y., D.X., D.L. and S.L. developed the models and conducted computational experiments; and D.X., S.G., M.T. and S.L. built the web server. W.L. and J.K. conducted biological experiments. D.X., Y.Q., J.Z., Y.Z., D.L., F.C. and H.Y. performed the analyses. D.X., Y.Q., J.Z., Y.Z., D.L., S.G., C.E., J.L., F.C. and H.Y. wrote and critically revised the manuscript. All authors discussed the results and reviewed the manuscript.

Competing interests

J.L. is co-scientific founder of Scipher Medicine, Inc., which applies network medicine strategies to biomarker development and personalized drug selection. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-024-02428-4.

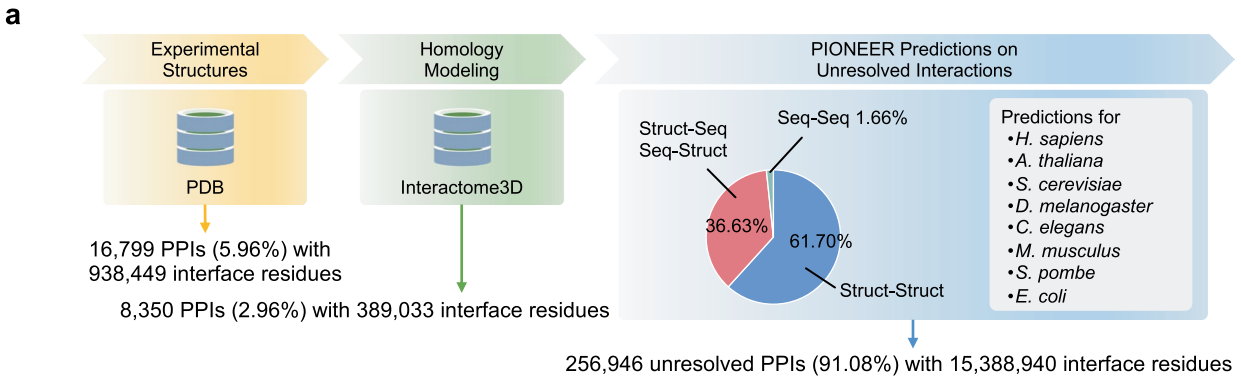
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02428-4.

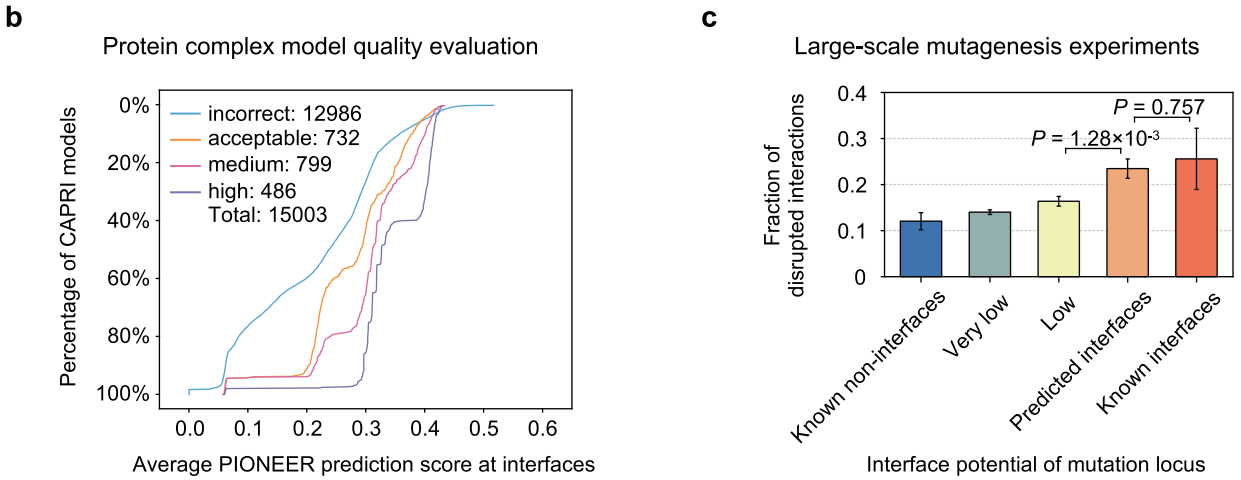
Correspondence and requests for materials should be addressed to Feixiong Cheng or Haiyuan Yu.

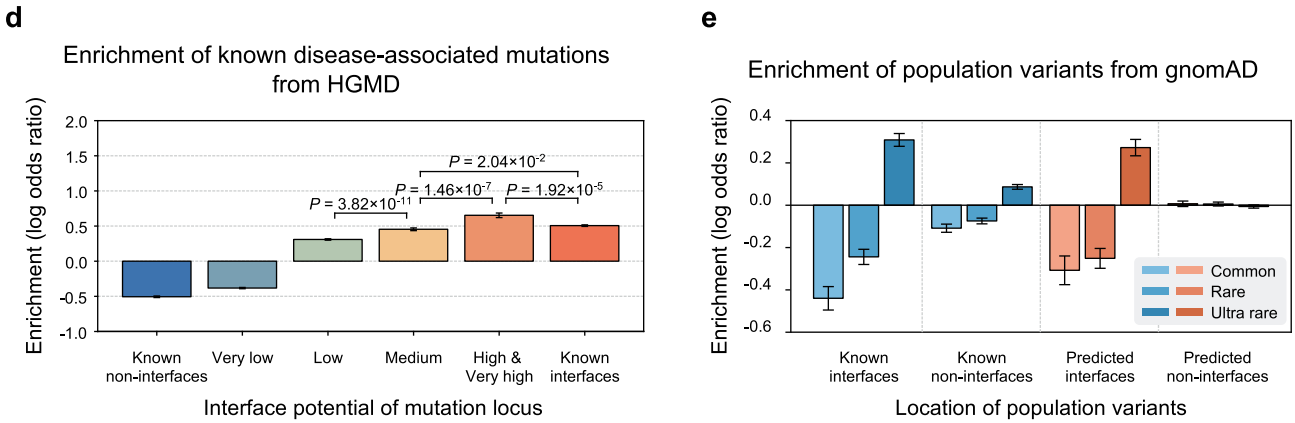
Peer review information *Nature Biotechnology* thanks Leng Han and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41587-024-02428-4

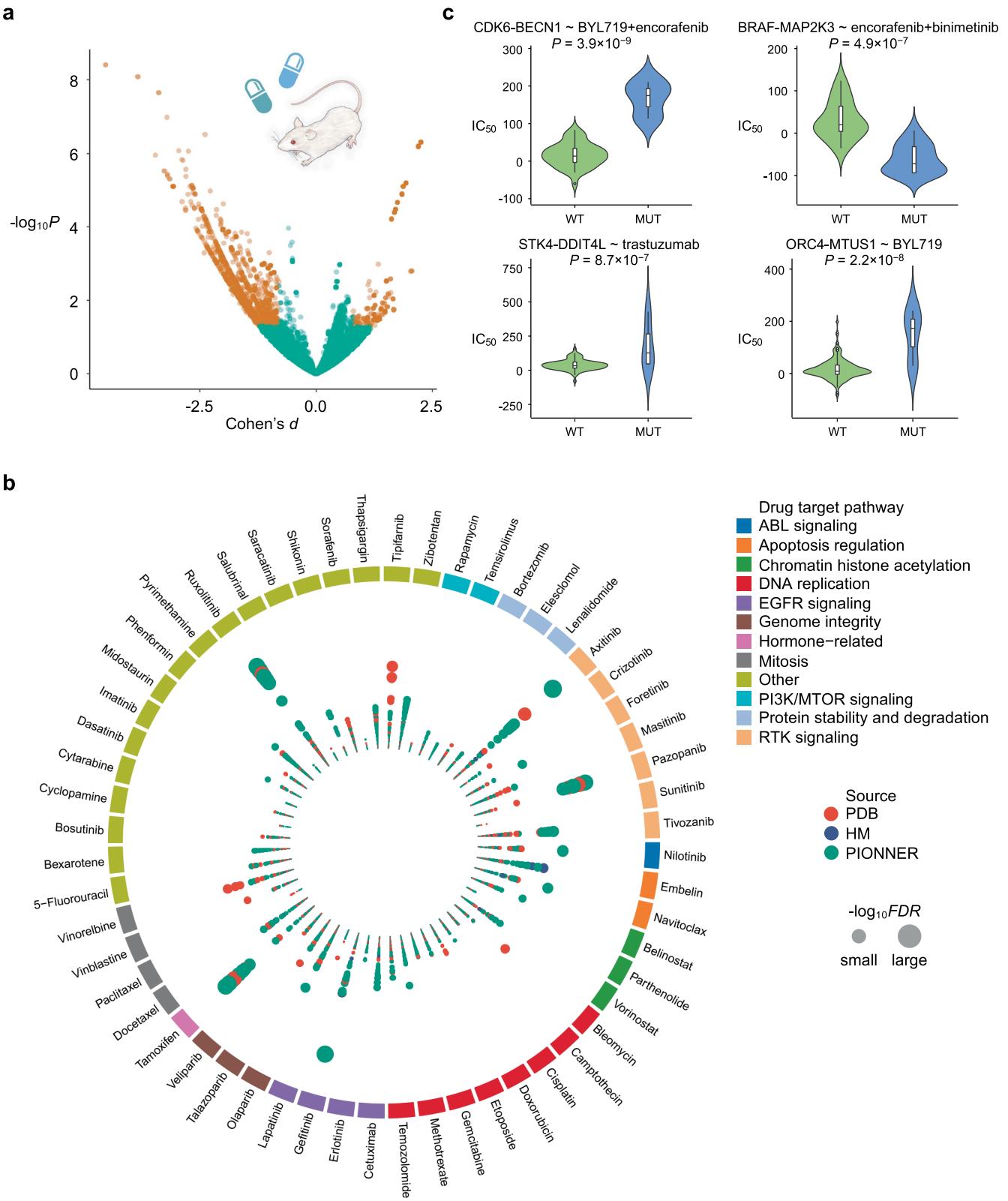






Extended Data Fig. 1 | PIONEER provides high-quality interfaces for the whole proteome. a, Workflow for compiling interactome PIONEER. The interfaces calculated from experimentally determined co-crystal structures or homology models are primarily used, the remaining unresolved interactions are predicted by PIONEER. **b**, Percentage of CAPRI decoys having a given average PIONEER prediction score at interfaces. Percentages are plotted along the y axis for 4 classes of CAPRI models. The total number of models in each class is indicated in the text in the figure. **c**, Fraction of interactions disrupted by random population variants in PIONEER-predicted and known interfaces. The error bar denotes standard error for the binomial distribution. Significance was determined by two-sided *z*-test. The *n* numbers are shown in Supplementary Table 6. **d**, Enrichment of disease-associated mutations in PIONEER-predicted and known interfaces. The error bar denotes standard error for the log odds ratio. Significance was determined by two-sided *z*-test. The *n* numbers are shown in Supplementary Table 7. **e**, Enrichment of population variants in PIONEER-predicted and known interfaces. The error bar denotes standard error for the log odds ratio. The *n* numbers are shown in Supplementary Table 8.

https://doi.org/10.1038/s41587-024-02428-4

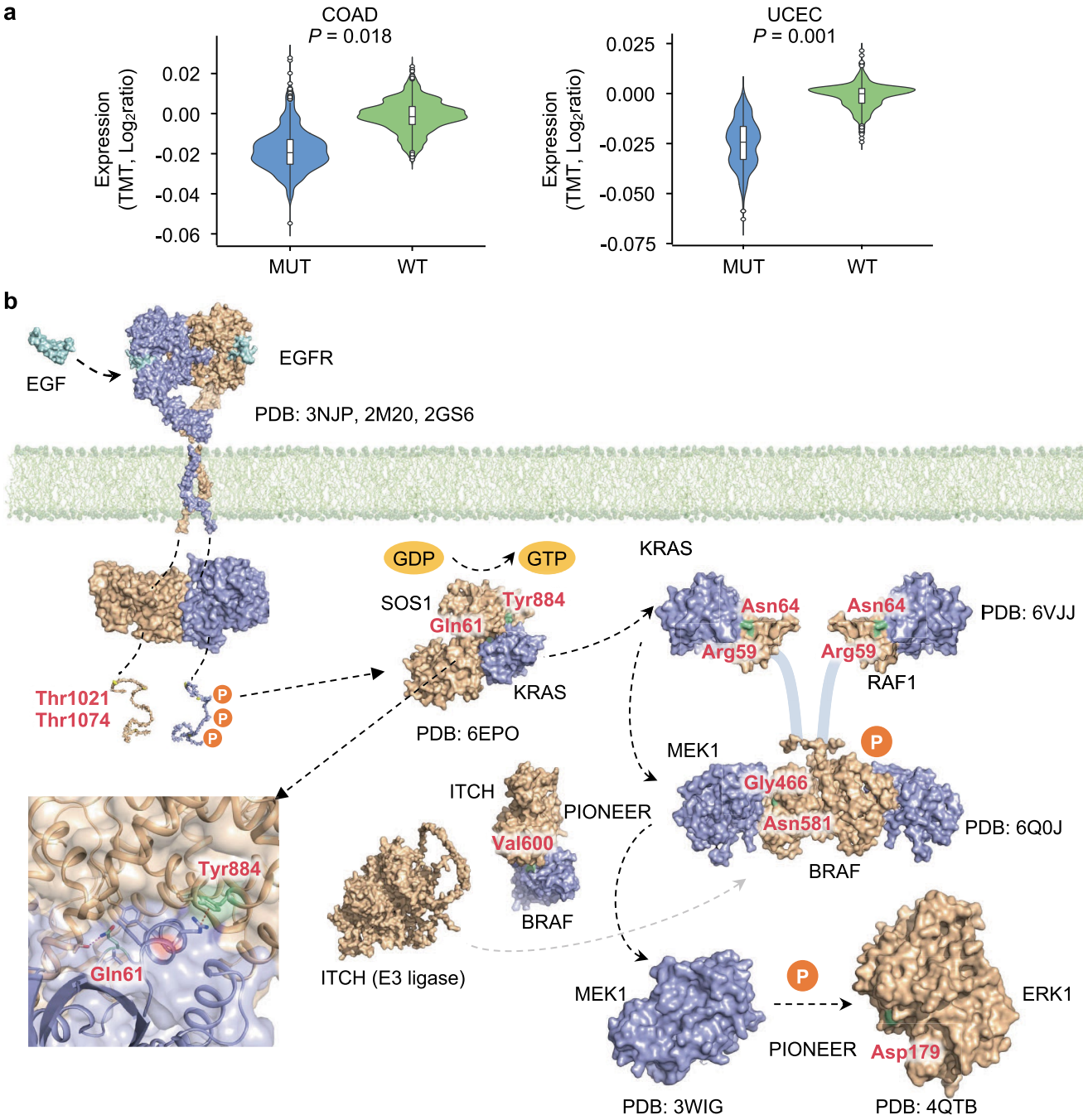


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41587-024-02428-4

Extended Data Fig. 2 | Pharmacogenomic landscape identified by the PIONEER-predicted interactome network. a, Drug responses evaluated by oncoPPIs in the PDX models. Effect size was quantified by Cohen's *d* statistic using the difference between two means divided by a pooled s.d. for the data. Significance was determined by ANOVA adjusted by Benjamini-Hochberg method. **b**, Circos plot displaying drug responses evaluated by putative PIONEER-predicted oncoPPIs harboring a statistically significant excess number of missense mutations at PPI interfaces, following a binomial distribution across selected anti-cancer therapeutic agents in cancer cell lines. Each node denotes a specific oncoPPI. Node size denotes significance determined by ANOVA. Effect size was quantified by Cohen's *d* statistic using the difference between two means divided by a pooled s.d. for the data. Node color denotes three different types of PPIs: (1) PDB: Red; (2) HM: Blue; and (3) PIONEER: Green. 'HM' represents homolog models. **c**, Highlighted examples of drug responses. Data are represented as a box plot with an underlaid violin plot in which the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent IQR × 1.5, and the dots are outlier points. Significance was determined by ANOVA. The *n* numbers are shown in Supplementary Table 16.

https://doi.org/10.1038/s41587-024-02428-4



Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41587-024-02428-4

Extended Data Fig. 3 | Proteogenomics of the PIONEER-predicted interactome network. a, Phosphorylation-associated PPI-perturbing mutations altered the proteomic changes in COAD and UCEC. The abundance of proteins was quantified using the TMT technique. Data are represented as a box plot with an underlaid violin plot in which the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent IQR × 1.5, and the dots are outlier points. Significance was determined by two-tailed Wilcoxon rank-sum test. The *n* numbers are shown in Supplementary Table 17. **b**, Phosphorylation-associated PPI-perturbing mutations in the EGFR–RAS–RAF–MEK–ERK cascade signaling pathway. The whole transmembrane EGFR structures were constructed by three crystal structures (PDB: 3NJP, 2M20, 2GS6). The membrane model is shown in green. The phosphorylation sites are indicated by the symbol 'P'. The detailed interface structure of SOS1–KRAS is also shown in the inset. The key mutated residue Gln61 on KRAS forms a hydrogen bond (purple dashed line) with residue Thr935 on SOS1, and Tyr884 on SOS1 is involved in a cation-π interaction (red dash line) with residue Arg73 on KRAS. Two subunits of RAF protein structure models were built by RAF1 and BRAF, separately (PDB: 6VJJ and 6Q0J). The two subunits are connected by a disordered loop indicated by blue cartoon lines. Two heterodimers of KRAS–RAF1 and BRAF–MEK1 constitutes the KRAS–RAF–MEK1 complex. PDB ID of each complex structure model is provided. **c**, Highlighted examples of drug responses. Data are represented as a box plot with an underlaid violin plot in which the middle line is the median, the lower and upper edges of the box are the first and third quartiles, the whiskers represent IQR × 1.5, and the dots are outlier points. Significance was determined by ANOVA. The *n* numbers are shown in Supplementary Table 16.

nature portfolio

Corresponding author(s): Feixiong Cheng and Haiyuan Yu

Last updated by author(s): Sep 9, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	python (3.8.1), numpy (1.19.2), pandas (1.2.4), numba (0.53.1), CD-HIT (4.8.1), PSIBLAST (2.10.0), Clustal Omega (1.2.4), NACCESS (2.1.1), ZDOCK (3.0.2), pytorch (1.4.0), torch-geometric (1.3.2), VEP (106), CADD (1.6), FoldX (5.0), R package (4.2.0), HADDOCK (2.4), CHARMM-GUI (3.8), ImageJ (1.53), PyMOL (2.4.2). RaptorX is available at https://github.com/realbigws/RaptorX_Property_Fast (commit 53e6e3a), ECLAIR is available at https://github.com/hyulab/ECLAIR (commit b2a177f). ScanNet is available at https://github.com/ScanNet/ScanNet (commit 3e57265). PIPGCN is available at https://github.com/fouticus/pipgcn (commit fd5f12c). BIPSPI+ is available at https://github.com/rsanchezgarc/BIPSPI (commit e0b1d01). PeSTo is available at https://github.com/LBM-EPFL/PeSTo (commit ba651aa). DeepPPISP is available at https://github.com/CSUBioGroup/DeepPPISP (commit 549b2b9). SASNet is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/H93ZKK (5.0). MaSIF-site is avialable at https://github.com/LPDI-EPFL/masif (commit 2a37051). DELPHI is available at https://github.com/lucian-ilie/DELPHI (commit 0a23ee0). SCRIBER web-server is available at http://biomine.cs.vcu.edu/servers/SCRIBER. DLPred is available upon request from the corresponding author of DLPred paper, this paper is cited in our paper. AF2Complex is available at https://github.com/FreshAirTonight/af2complex (1.4.0, commit 4c55803). FoldDock is available at https://gitlab.com/ElofssonLab/FoldDock (commit 9a1a26ce). AlphaFold-Multimer is available at https://github.com/google-deepmind/alphafold (2.3.1, commit 502d7a1). The source code of PIONEER is available at https://github.com/hyulab/PIONEER.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Mutation data from the TCGA study were downloaded from NCI genomic data commons (https://portal.gdc.cancer.gov). MSK MetTropism dataset was downloaded from cBioPortal (https://www.cbioportal.org/study/summary?id=msk_met_2021). Variant data from 1000 Genomes Project were downloaded from NCBI FTP site (https://ftp-trace.ncbi.nih.gov/1000genomes/ftp). The ExAC data set was downloaded from gnomAD (https://gnomad.broadinstitute.org/downloads#exac-variants). Variants collected by HGMD were downloaded from https://www.hgmd.cf.ac.uk/ac/index.php. Genomic variants and drug response data of human cancer cell lines were downloaded from GDSC datasets (https://www.cancerrxgene.org/downloads/bulk_download). Genomic profiling of PDXs and drug response curve metrics of PCTs were downloaded from the Supplementary Table 1 of the corresponding paper (https://www.nature.com/articles/nm.3954#Sec28). The homologous structures of PPIs that don't have co-crystal structures are collected from Interactome3D (https://interactome3d.irbbarcelona.org). The ModBase data were downloaded from https://modbase.compbio.ucsf.edu. The PDB data were downloaded from PDB FTP site (https://files.wwpdb.org/pub/pdb/data/structures/divided/pdb). The AlphaFold2-predicted protein structures were download from AlphaFold database (https://alphafold.ebi.ac.uk). All other data supporting the results in this study are available in supplementary materials, and at https://pioneer.yulab.org.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by data availability. No statistical method was used to predetermine sample size. We used all available libraries for each of the assays compared in this study.
Data exclusions	No data was excluded from analysis.
Replication	All attempts at replication were performed three times independently and were successful.
Randomization	We did the training, validation and benchmark test split randomly following the standard machine learning practice.
Blinding	Used during benchmark test following the standard machine learning practice.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: EZview Red Anti-FLAG M2 Affinity Gel (catalog no. F2426, Sigma-Aldrich), Anti-FLAG (catalog no. F1804, Sigma-Aldrich), Anti-MYC (catalog no. 132500, Invitrogen), and Anti-HA antibody (catalog no. M20003, Abmart)

Validation: All antibodies were validated according to the manufacturer's statement. Validation information is available on vendor's website. EZview Red Anti-FLAG M2 Affinity Gel (catalog no. F2426, Sigma): https://www.sigmaaldrich.com/US/en/product/sigma/f2426. Anti-FLAG (catalog no. F1804, Sigma): https://www.sigmaaldrich.com/US/en/product/sigma/f1804. Anti-MYC (catalog no. 132500, Invitrogen): https://www.thermofisher.com/antibody/product/c-Myc-Antibody-clone-9E10-Monoclonal/13-2500. Anti-HA antibody (catalog no. M20003, Abmart): https://www.ab-mart.com.cn/page.aspx?node=%2059%20&id=%20963.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s): HEK293T cells were purchased from ATCC. H1975 cells were generously supplied by the Stem Cell Bank, Chinese Academy of Sciences.

Authentication: ATCC and the Stem Cell Bank, Chinese Academy of Sciences thoroughly authenticates the cells they provided.

Mycoplasma contamination: We didn't detect any contamination issues for the used cell lines.

Commonly misidentified lines (See ICLAC register): No.

Plants

Seed stocks: Not applicable.

Novel plant genotypes: Not applicable.

Authentication: Not applicable.