Title: A multiscale functional map of somatic mutations in cancer integrating protein structure and network topology

Authors: Yingying Zhang[1,2,3]†, Alden K. Leung[1,2]†, Tian Qiu[4], Le Li[1,2], Junke Zhang[1,2], Shayne Wierbowski[1,2], James Booth[1,5], Haiyuan Yu[1,2]*

Affiliations:

[1]Department of Computational Biology, Cornell University; Ithaca, 14853, USA.

[2]Weill Institute for Cell and Molecular Biology, Cornell University; Ithaca, 14853, USA.

[3]Department of Molecular Biology and Genetics, Cornell University; Ithaca, 14853, USA.

[4]School of Electrical and Computer Engineering, Cornell University; Ithaca, 14853, USA.

[5]Department of Statistics and Data Science, Cornell University; Ithaca, 14853, USA.

*Corresponding author. Email: haiyuan.yu@cornell.edu.

†These authors contributed equally to this work.

Abstract: A major goal of cancer biology is to understand the mechanisms underlying tumorigenesis driven by somatically acquired mutations. Existing computational approaches focus on either scoring the pathogenicity of mutations or characterizing their effects at specific scales. Here, we established a unified computational framework, NetFlow3D, that systematically maps the multiscale mechanistic effects of somatic mutations in cancer. The establishment of NetFlow3D hinges upon the Human Protein Structurome, a complete repository we first compiled that incorporates the 3D structures of every single protein as well as the binding interfaces for all known PPIs in humans. The vast majority of 3D structural information was resolved by recent deep learning algorithms. By applying NetFlow3D to 415,017 somatic protein-altering mutations in 5,950 TCGA tumors across 19 cancer types, we identified 1,656 intra- and 3,343 inter-protein 3D clusters of mutations throughout the Human Protein Structurome, of which ~50% would not have been found if using only experimentally-determined protein structures. These 3D clusters have converging effects on 377 cellular subnetworks. Compared to canonical PPI network analyses, NetFlow3D achieved a 5.5-fold higher statistical power for identifying significantly dysregulated subnetworks. The majority of identified subnetworks were previously obscured by the overwhelming background noise of non-clustered passenger mutations, including portions of non-canonical PRC1, mediator complex, MCM2-7 complex, neddylation of cullins, complement system, TRiC, etc. NetFlow3D and our pan-cancer results can be accessed from http://netflow3d.yulab.org. This work shows that mapping how individual mutations act across scales requires the integration of their local spatial organization on protein structures and their global topological organization in the PPI network.

One-Sentence Summary: Integrating the arrangement of somatic mutations on 3D protein structures and throughout the protein-protein interaction network provides unique insights of their multiscale mechanistic effects in cancer.

Main Text:

Somatically acquired mutations are one of the major sources driving tumorigenesis (*1*). Computational approaches have been developed to assign pathogenicity scores to given mutations, indicating their phenotypic effects on an organism (*2–8*). Complementary to these approaches, understanding the mechanisms driven by each mutation—from altering genomic sequences to changing key amino acid residues to dysregulating relevant cellular pathways—is key to developing effective therapeutic strategies. Efforts have been made to interpret the effects of mutations at specific scales (*9–17*). Some studies focus on the molecular effects and look for spatial clustering of mutations within critical regions of proteins (*9–15*, *18*, *19*). Others focus on cancer-related pathways and look for significantly mutated subnetworks of proteins (*16*, *17*, *20*). Studies at the molecular and pathway levels offer complementary insights into the underlying mechanisms of cancer.

At the molecular scale, the spatial organization of somatic mutations on 3D protein structures can reveal functionally important protein regions and can thus assist in identifying cancer driver mutations (*9–15*, *18*, *19*). Given that the overwhelming majority of somatic mutations in cancer are passengers (i.e., mutations that do not contribute to cancer development) (*21*), identifying spatial clusters of mutations can narrow down potential driver mutations and thus significantly boost the signal-to-noise ratio. However, current 3D clustering algorithms either limit their scope to the experimentally-determined structures (*9*, *11*, *12*, *15*), or specifically focus on single proteins (*10–12*, *14*) or protein-protein interaction (PPI) interfaces (*22*, *23*). No approach yet fully examines the 3D structures of every single protein as well as the binding interfaces of all known PPIs in humans, leaving many spatial clusters yet to be identified. The bottleneck has been the limited coverage of 3D structural information: only ~35% of single proteins and ~4% of known PPIs in humans have experimentally-determined structures (see materials and methods).

At the pathway scale, PPI network analyses are widely used to investigate the biological processes involved in cancer development (*16*, *17*). These approaches integrate the topology of PPI network with the gene-level statistics to identify significantly mutated subnetworks. Nevertheless, the smallest unit of interest in canonical PPI network analyses is one protein or gene. Therefore, the diverse functional effects of different mutations in the same protein were not fully captured and reported, leading to potential dilution of statistical power by passenger mutations. Moreover, even if the cancer proteins and pathways are known, it is still largely unknown whether an individual patient's mutations within these cancer pathways truly result in corresponding pathways dysregulation. Therefore, canonical PPI network analyses leave a translational gap between biological discoveries and clinical applications.

Bridging the functional effects of mutations on molecular- and pathway-scale is key to comprehensively understanding the mechanisms of tumor development in individual cancer patients. However, a unified computational tool that systematically maps the multiscale mechanistic effects of somatic mutations in cancer is still missing. Here, we (i) compiled the Human Protein Structurome–a comprehensive repository of 3D structural information that covers every single protein as well as all known PPIs in humans–by leveraging recent deep-learning-based protein structural information prediction strategies, and (ii) established NetFlow3D, a systematic tool that maps the multiscale functional effects of somatic mutations by integrating their local spatial organization on protein structures and their global topological organization in the PPI network. Our key idea of comprehending how mutations act across scales is based on a global view of how the submolecular spatial clusters are organized in the PPI

network, which would not have been achieved without the 3D structural details fully covering every protein and PPI in the network. Since the smallest unit of interest in NetFlow3D is mutations instead of proteins, NetFlow3D can help more accurately predict which set of cancer patients may have better responses to the treatments specific to which mechanism.

By applying NetFlow3D to a TCGA pan-cancer dataset across 19 cancer types, we generated a multiscale functional map of somatic mutations: (i) Our analysis revealed nearly double the number of spatial clusters across the Human Protein Structurome compared to those identified solely on experimentally-determined structures. This expansion of scope encompasses a vast number of potential driver mutations that have not been previously reported. (ii) By mapping the submolecular spatial clusters across the PPI network, we were able to identify hundreds of potentially dysregulated subnetworks. This approach yielded a statistical power that was 5.5 times higher than that of canonical PPI network analyses. The boosted power led to the discovery of many potentially novel mechanisms that had been blurred by the overwhelming passenger mutations, including the dysregulation of non-canonical PRC1, mediator complex, MCM2-7 complex, neddylation of cullins, complement system, TRiC, mitochondrial ribosome, spliceosome, etc. To make both the NetFlow3D tool and the findings in our TCGA pan-cancer study more accessible to the broader biomedical community, we built a user-friendly interactive web server (http://netflow3d.yulab.org/) where researchers can directly apply NetFlow3D to their own datasets and interactively browse our TCGA pan-cancer results.

NetFlow3D maps the functional effects of individual somatic mutations across different scales

We collected and processed a TCGA pan-cancer dataset of 415,017 somatic protein-altering mutations across 5,950 tumor samples spanning 19 cancer types (Fig. 1A; table S1; materials and methods). Of these mutations, 86% were expected to change only one or a few amino acid residues in the encoded proteins (i.e. missense mutations and in-frame indels), and were therefore termed as non-truncating (NT) mutations. Without further biological contexts, it is particularly difficult to interpret the varying downstream functional effects based on these subtle changes to the protein sequences.

Mounting evidence has demonstrated the efficaciousness of identifying functional NT mutations by detecting spatial clusters of NT mutations on 3D protein structures (*9–15*, *18*, *19*). However, the majority of single proteins and PPIs in humans do not have experimentally-determined structures. Therefore, it remains largely unknown how NT mutations are spatially organized on those proteins and PPIs. Recent breakthroughs in deep-learning-based strategies for resolving protein structural information are filling in the missing 3D structural details (*24–26*). For instance, AlphaFold 2 generates single protein structures with atomic accuracy for all proteins in humans as well as in some other species (*25*). PIONEER, our latest deep-learning framework built upon our previous machine-learning framework ECLAIR (*27*) with much improved performance, generates the binding interfaces for all known human PPIs.

In order to achieve full-coverage spatial mapping of NT mutations at the protein level, we compiled a complete repository that contains the structures of all human proteins as well as the binding interfaces of all known human PPIs and available multi-protein complex structures, which we named “the Human Protein Structurome” (Fig. 1B; materials and methods). Specifically, we collected (i) 155,902 structures of 20,813 single proteins in humans from both

the Protein Data Bank (PDB) (*28*, *29*) and the AlphaFold Protein Structure Database (AlphaFold DB) (*25*, *30*), (ii) 28,737 structures of multi-chain protein complexes from PDB (*28*, *29*) which cover 9,252 known human PPIs, and (iii) the binding interfaces of 146,137 known human PPIs consisting of 528,672 non-redundant amino acid residues from PIONEER. Overall, 64% of single proteins and 94% of known PPIs in humans completely rely on recent deep-learning algorithms to resolve their structures and binding interfaces, respectively (Fig. 1B).

The Human Protein Structurome we compiled contains a wealth of 3D structural information obtained from cutting-edge deep learning techniques, which were not available to previous 3D clustering algorithms. This resource enables us to identify spatial clusters of NT mutations in an unprecedentedly thorough manner, and thus provides us with a unique view of how the submolecular spatial clusters are arranged in the PPI network. Unlike traditional PPI network analyses, our unified computational framework, NetFlow3D, focuses on mutations as the smallest unit of interest rather than proteins. Our key concept is to systematically interpret the functional effects of individual NT mutations across multiple scales by integrating the local spatial organization of numerous NT mutations on protein structures and their global topological organization in the PPI network. This approach enables us to characterize the diverse functional effects caused by different mutations on the same protein, greatly improving the resolution of delineating cancer-driving events.

The first part of NetFlow3D is a 3D clustering algorithm that identifies spatial clusters of NT mutations throughout the entire Human Protein Structurome (Fig. 1C; materials and methods). Our algorithm looks for both 3D clusters within single proteins (intra-protein 3D clusters) and 3D clusters spanning interacting proteins (inter-protein 3D clusters). Unlike most existing algorithms for similar purposes, our method (i) determines the physical contact between every pair of amino acid residues by accounting for their varying 3D distances across all available structures instead of solely based on a single snapshot represented by one structure, and (ii) considers the diverse background mutability of different nucleotide contexts when determining the significance of 3D clusters instead of assuming uniform mutability as used by most 3D clustering algorithms (see materials and methods). Since our method makes better use of the dynamic nature of proteins and better accounts for the heterogeneous background mutability along the genome, we manage to thoroughly identify potential driver NT mutations with unprecedented sensitivity and specificity, and characterize functionally important regions in proteins.

The second part of NetFlow3D is a network propagation model that characterizes the pathway-scale effects of the potential driver NT mutations identified in the first part (Fig. 1D; materials and methods). Inspired by HotNet2 (*16*), we use a heat diffusion model to model the effects of mutations in one protein towards other proteins, which are spread through the PPI network. Specifically, we place heat sources at the proteins with mutations and let the heat flow through the PPI network until equilibrium state, and then identify the subnetworks with strong internal heat exchanges. Importantly, we made two major improvements compared to HotNet2 (*16*) and other PPI network analyses. First, we define the heat sources as the proteins involved in 3D clusters instead of any proteins with NT mutations. In this way, we strictly focus on a small set of potential driver mutations instead of considering all NT mutations equivalent to each other. Second, we force the heat to diffuse unevenly from a given protein to its neighbors by designing a 5:1 heat delivery ratio towards the neighbors when 3D clusters are on vs. not on the corresponding PPI interfaces. This design is inspired by recent studies (*31*, *32*) that report a

5-fold difference in PPI-perturbation rates when known disease mutations are on vs. not on the binding interfaces. Taken together, our network propagation model incorporates the local spatial organization of mutations on the 3D structures of proteins and PPIs, and therefore achieves a higher resolution than the conventional node-and-edge view of the PPI network. In our model, the identified subnetworks act as the converging points of different 3D clusters, and can be interpreted as potential pathways dysregulated by these mutations. As a result, perturbation of these subnetworks is likely to play functional roles in cancer development, and therefore can be used to characterize the higher-level functional effects of the spatially clustered NT mutations identified in the first part.

As a complement to the first and second parts that focus on the NT mutations, NetFlow3D also accounts for the functional effects of loss-of-function (LOF) mutations. Specifically, NetFlow3D identifies the proteins that are significantly enriched for LOF mutations scattered across their entire sequences, and integrates these enrichment signals as heat sources into the heat diffusion model in the second part (Fig. 1A and 1D; materials and methods).

Overall, NetFlow3D maps the functional effects of somatic mutations across multiple scales (Fig. 1E): from (i) the point mutations in cancer genomes to (ii) the protein residues with potential functional importance (suggested by 3D clusters), to (iii) the perturbation of key proteins/PPIs, to (iv) the dysregulation of relevant cellular subnetworks. This multiscale functional map provides rich insights for understanding the tumorigenic mechanisms driven by the subtle genomic mutations in somatic tissues.

Significant intra- and inter-protein 3D clusters throughout the Human Protein Structurome

We applied the 3D clustering algorithm in NetFlow3D to the 356,683 somatic NT mutations in our preprocessed TCGA pan-cancer dataset. As a result, we identified 1,656 significant intra-protein 3D clusters and 3,343 significant inter-protein 3D clusters throughout the Human Protein Structurome, among which 724 (44%) and 1,709 (51%) clusters were identified based on the deep learning-generated single protein structures and PPI interfaces, respectively (Fig. 2A; table S2). NT mutations incorporated by these significant 3D clusters were considered potentially functional (table S3). 98.5% of non-redundant NT mutations ("non-redundant" refers to the amino acid residue changes) in significant 3D clusters are rare (i.e. occur in no more than three different tumor samples), and thus can hardly be identified by recurrence-based approaches. Identifying these likely functional rare mutations highly relies on their spatial organization on protein structures. The increased number of single protein structures and PPI interfaces resolved by deep learning algorithms play key roles in promoting the identification of novel rare driver mutations. For example, a significant 3D cluster within the NRP1 protein is identified based on the AlphaFold 2-generated structure, which consists entirely of rare NT mutations (Fig. 2B).

Importantly, single protein structures alone (even though covering the entire human proteome) are still not enough for the comprehensive identification of all 3D clusters. This is because many driver mutations accumulate at the binding interfaces of cancer-related PPIs (*22*, *32*, *33*). Only looking at individual proteins will split inter-protein 3D clusters into smaller fragments on individual proteins, making them harder to identify. This is demonstrated by the fact that, among the non-redundant NT mutations incorporated by the significant inter-protein clusters, 75.2%

would not have been identified if we only searched for significant clusters within single protein structures. Such situations are exemplified by a significant inter-protein cluster on the PIONEER-generated PPI interface between FNTA and FNTB proteins (Fig. 2C). These results highlight the importance of knowing PPI interfaces, which are mostly generated by our deep learning framework PIONEER, in identifying potential driver mutations. Overall, 4,976 tumor samples (83.6%) have at least one potential driver NT mutation incorporated by our significant 3D clusters, demonstrating the thoroughness of our 3D cluster identification.

We then evaluated the clustering performance of NetFlow3D and compared it with four state-of-the-art 3D clustering algorithms (*9–11*, *13*) which represent major sources of 3D cluster identification (Fig. 2D and 2E; materials and methods). We applied each algorithm to the same TCGA pan-cancer dataset, and compared the 3D clusters identified by different algorithms. Considering that (i) some algorithms only focus on intra-protein clusters (*10*, *11*) while some others identify both (*9*, *13*), and (ii) some algorithms only use experimentally-determined structures (*9*, *11*) while some others also include comparative protein structure models (*10*, *13*), we therefore make coherent comparisons by (i) assessing the intra- and inter-protein clusters separately, and (ii) limiting the comparisons to the 3D clusters identified on experimentally-resolved structures. We compared the top genes with the most significant 3D clusters on their protein products ranked by each algorithm. As a result, among the same number of top genes ranked by each algorithm, NetFlow3D-ranked genes consistently include more known cancer genes (table S4) (*34*, *35*) as well as less non-cancer-associated genes (table S5) (*36–38*), demonstrating our advanced sensitivity and specificity. Therefore, NetFlow3D significantly outperforms other algorithms for the identification of both intra- and inter-protein 3D clusters.

Furthermore, we evaluated the NT mutations incorporated by the significant 3D clusters identified by NetFlow3D, and compared these mutations with known driver and passenger mutations (*39–41*) (Fig. 2F; materials and methods). Specifically, we used PolyPhen-2 (*42*) to score the effects of NT mutations on protein structures and functions, and calculated the fraction of mutations scored as "probably damaging" for five categories of mutations: (i) Known driver mutations (table S6) (*39*, *40*); (ii) Mutations incorporated by our significant 3D clusters on experimentally-determined structures; (iii) Mutations incorporated by our significant 3D clusters on deep learning-generated single protein structures and PPI interfaces; (iv) Mutations not incorporated by any of our significant 3D clusters; (v) Known passenger mutations (table S7) (*41*). As a result, the fraction of probably damaging mutations in category ii significantly surpasses that in all other categories. The fraction in category iii is significantly higher than that in category iv and v, although lower than that in category i. These results suggest that the molecular effects of NT mutations incorporated by our significant 3D clusters are substantially greater than that of non-clustered mutations as well as known passenger mutations. Our significant 3D clusters on experimentally-resolved structures even incorporate more impactful mutations than known driver mutations. In contrast, non-clustered NT mutations are characterized by a much smaller fraction of probably damaging mutations. This fraction is comparable to that for known passenger mutations, indicating that non-clustered NT mutations are mostly non-functional. Taken together, these results further corroborate the functional roles of the NT mutations incorporated by the NetFlow3D-identified significant 3D clusters, and thus demonstrate the accuracy of our 3D clustering algorithm.

NetFlow3D uncovers significantly dysregulated subworks previously blurred by overwhelming passenger mutations

The 3D clustering part of NetFlow3D has enabled us to extract a small set of potential driver mutations from the vast pool of somatic NT mutations that are often overshadowed by non-functional passenger mutations, and better understand the molecular effects of nominated driver mutations on specific proteins and PPIs. Next, we designed a network propagation model to interpret the pathway-scale functional effects of these nominated driver mutations. The key idea is to characterize the long-range effects of nominated driver mutations on other proteins in the PPI network, and thus understand the converging effects of these mutations on common cellular subnetworks. Compared to HotNet2 (*16*) and other canonical PPI network analyses, the major improvement of NetFlow3D is the incorporation of the local spatial mapping of mutations on the single protein structures and PPI interfaces. Thus, NetFlow3D defines heat source and heat transfer rate in a 3D structurally-informed way (Fig. 3A). First, since the vast majority of somatic mutations are passengers (*21*) and 3D clusters suggest the accumulation of driver mutations (*9–15*, *18*, *19*), strictly focusing on 3D clusters can highlight cancer pathways which are otherwise obscured by the overwhelming passenger mutations. In NetFlow3D, the heat sources are defined as the proteins involved in 3D clusters (or LOF enrichment signals), while in standard PPI network analyses, proteins with any mutations will be considered as heat sources. Second, previous studies have shown a 70.6% PPI-perturbation rate for known disease mutations on the interfaces, as well as a 13.3% rate for non-interface mutations (*31*, *32*). Driven by this ~5-fold difference, NetFlow3D uses a 5:1 ratio of heat transfer rate through PPIs when 3D clusters are on vs. not on the interfaces. In contrast, since standard PPI network analyses lack the knowledge of which mutation is on the interface with which interaction partner, they generally consider the effects of mutations on every interaction partner equivalent.

As a result, although NetFlow3D starts from only one fifth of heat sources compared to standard PPI network analyses, it ends up with 5.5-fold discoveries of significantly dysregulated subnetworks (Fig. 3B and 3D; materials and methods). These results highlight the boosted statistical power when combining the local spatial organization of mutations on protein structures and their global topological organization in the PPI network. To estimate the contributions of NetFlow3D-defined heat source and heat transfer rate to the improvement of statistical power, we built a mixed model where the definition of heat source is the same as that in NetFlow3D while the heat transfer rate is defined in the same way as standard PPI network analyses (Fig. 3C; materials and methods). As a result, the mixed model identifies 3.1-fold more significantly dysregulated subnetworks compared to standard PPI network analyses, exhibiting the statistical power gained by performing 3D clustering analysis (Fig. 3C and 3D). While compared to NetFlow3D, the number of significantly dysregulated subnetworks identified by the mixed model reduces by almost half, which further exhibits the statistical power gained from our carefully-designed heat transfer rate (Fig. 3B and 3C).

We then evaluated the significantly dysregulated subnetworks identified by NetFlow3D (Fig. 3E; materials and methods). As benchmarks, we collected 32 well-known cancer signaling pathways (table S8) from NetSlim (*43*) as positive controls and 2,055 general biological processes (table S9) from Gene Ontology (GO) (*44–48*) as negative controls. In each of these pathways, we calculated two types of mutation enrichment: one for the mutations in 3D clusters, and the other for all mutations. As a result, in well-known cancer pathways (*43*), there is a median of 4.2-fold increase in the enrichment when switching from all mutations to the mutations in 3D clusters

(Fig. 3E). While in general biological processes (*44–48*) , there is no significant difference between the two groups of mutations (Fig. 3E). These results further demonstrate the boosted sensitivity of identifying cancer pathways by strictly focusing on 3D clusters. Similar to the well-known cancer pathways, the NetFlow3D-identified significantly dysregulated subnetworks also show striking increases in the enrichment when contrasting the two groups of mutations, with a median of 3.9-fold increase among the subnetworks containing established cancer genes documented by the Cancer Gene Census (CGC) (*34*, *35*), as well as a median of 3.7-fold increase among the subnetworks without CGC genes, which likely reflect novel cellular processes or functions involved in cancer development (Fig. 3E).

Taken together, NetFlow3D improves the statistical power of identifying significantly dysregulated subnetworks by strictly focusing on a small set of spatially clustered mutations and emphasizing their varying effects on different interaction partners. This improvement would not have been achieved without knowing the comprehensive 3D structural details of the entire human proteome and interactome. Consequently, NetFlow3D facilitates the discovery of many previously unknown pathways and cellular functions involved in cancer development, which can hardly be identified using canonical PPI network analyses.

A pan-cancer functional map of somatic mutations across scales

By applying NetFlow3D to the TCGA pan-cancer dataset across 19 cancer types, we generated a multiscale functional map of somatic mutations in cancer (Fig. 4). This map starts from 44,521 potential driver mutations in 5,312 tumor samples, including 32,393 NT mutations and 12,128 LOF mutations (table S3). These mutations are mapped to 4,999 significant 3D clusters and 469 LOF enrichment signals (table S2). Furthermore, 2,139 significant 3D clusters and 131 LOF enrichment signals converge on 377 cellular subnetworks (table S10).

Among these subnetworks, 44 are significantly larger than expected ("large" refers to the number of proteins in the subnetwork). Subnetworks of such sizes can hardly be observed if we randomly shuffle heat sources and heat transfer rates (see materials and methods). These significantly dysregulated subnetworks cover a broad spectrum of cellular processes and functions (Fig. 4; table S11-S54). Some of them are portions of well-known cancer pathways such as mTOR signaling (table S15) (*49*, *50*) and BCL2-regulated apoptosis (table S16) (*51*). Some others are supported by more recent evidence, such as the disorganization of PRC2 (table S20) (*52*, *53*) and non-canonical PRC1 (table S47) (*53*, *54*). Many of these significantly dysregulated subnetworks have been largely overlooked by existing PPI network analyses since their signals are severely diluted by the overwhelming passenger mutations throughout the PPI network. Such subnetworks include portions of the mediator complex (table S24 and S25), MCM2-7 complex (table S13), neddylation of cullins (table S14), complement system (table S28), TRiC (table S45), mitochondrial ribosome (table S50), spliceosome (table S22 and S23), etc. These subnetworks can facilitate the discovery of potentially novel pathways and cellular functions involved in cancer. Moreover, NetFlow3D also characterizes the relationships between different subnetworks (see materials and methods). Specifically, some dysregulated subnetworks have one-way effects on other subnetworks. Among these 44 significantly dysregulated subnetworks, such one-way effects exist between 13 pairs of subnetworks (Fig. 4; table S55). The proteins mediating the spread of effects from one dysregulated subnetwork to the other are mainly involved in signaling, transcription regulation and protein ubiquitination. Overall, 87.4% of

genes in the NetFlow3D-identified subnetworks have not yet been documented by CGC. Therefore, our multiscale functional map of somatic mutations can help uncover many previously unknown cancer mechanisms.

Moreover, among all subnetworks identified by NetFlow3D, 35 subnetworks contain proteins targeted by FDA-approved drugs (Fig.4; table S56; materials and methods). Considering that (i) NetFlow3D-identified subnetworks are converging points of different potential driver mutations, and (ii) these potential driver mutations usually come from different cancer patients, NetFlow3D results suggest that many of these cancer patients may have similar tumorigenic mechanisms because their potential driver mutations are likely to dysregulate the same cellular pathways. Therefore, NetFlow3D provides useful information for exploring new potential uses of existing drugs that are effective for specific mechanisms but can only act on very few proteins. Nevertheless, further research is still needed to fully realize the potential benefits of the information provided by NetFlow3D, since the specific applications are still dependent on a more detailed characterization of the pathways and a thorough understanding of how specific drugs act.

NetFlow3D enables more accurate prediction of the converging mechanisms across patients and cancers

The main focus of canonical PPI network analyses is to identify cancer proteins and pathways rather than accurately pinpointing driver mutations. However, not all cancer patients with mutations in cancer pathways have those pathways dysregulated, as many mutations don't have functional effects. Since NetFlow3D pinpoints a small set of spatially clustered mutations for each subnetwork, we manage to prioritize a group of most relevant cancer patients for each potential pathway. As a result, NetFlow3D provides useful information for the accurate prediction of which therapeutic strategies targeting a specific mechanism will be effective for which group of patients.

As an example, we present one of the significantly dysregulated subnetworks identified by NetFlow3D (Fig. 5A; table S47). Most proteins in this subnetwork are subunits of non-canonical polycomb repressive complex 1 (ncPRC1), a histone ubiquitin ligase generally associated with transcription repression by establishing H2AK119ub (*53*). Target genes of ncPRC1 are involved in cell metabolism, cell cycle, regulation of gene expression, and signaling pathways (*55*, *56*). Recent experimental evidence has shown that knocking out ncPRC1 subunits led to de-repression of target genes and promoted tumorigenesis (*57–62*). NetFlow3D identified this subnetwork because it is the converging point of 16 significant 3D clusters and 4 LOF enrichment signals (Fig. 5B; table S2).

In our TCGA pan-cancer dataset, there are 454 cancer patients with somatic NT mutations in this subnetwork, among which NetFlow3D carefully prioritizes 68 patients (15%) with mutations incorporated by significant 3D clusters (Fig. 5B). As suggested by NetFlow3D, these prioritized patients are more likely to have dysregulated ncPRC1 compared to the rest of patients. Although the prioritized patients have 16 types of cancer and their potential driver mutations in this subnetwork are diverse, NetFlow3D suggests that they might share similar tumorigenic mechanisms (Fig. 5C). The potential convergence of mechanisms across patients and cancers suggests that similar therapeutic strategies targeting ncPRC1 dysregulation may be effective in treating these patients.

It is the first time that the significant dysregulation of ncPRC1 is reported by a computational approach. There are several reasons why this subnetwork cannot be identified by using existing recurrence-based, 3D structure-based, or PPI network-based approaches individually: (i) Every amino acid residue in this subnetwork is rarely mutated (except for BCL6 R594 that is mutated in four tumor samples); (ii) Only 21.8% of amino acid residues in this subnetwork are covered by experimentally-resolved structures; (iii) The total number of spatially clustered mutations on both experimentally- and computationally-resolved single protein structures and PPI interfaces only account for 12.5% of all somatic NT mutations in this subnetwork, indicating that the signals can easily be obscured by the overwhelming passenger mutations. Therefore, the Human Protein Structurome and our NetFlow3D tool are powerful for uncovering potentially novel mechanisms underlying tumorigenesis.

Discussion

We present a unified computational framework, NetFlow3D, to systematically interpret the multiscale functional effects of somatic mutations in cancer. NetFlow3D helps better understand the mechanisms underlying tumorigenesis, and thus is complementary to the approaches that assign pathogenicity scores to given mutations. NetFlow3D maps both the molecular- and pathway-scale effects of somatic mutations by integrating their local spatial organization on protein structures and their global topological arrangement in the PPI network. This integration hinges upon the Human Protein Structurome, which covers the 3D structural details of every single protein as well as all known PPIs in humans, representing the most comprehensive resource of 3D structural information to date. By applying NetFlow3D to the TCGA pan-cancer dataset across 19 cancer types, NetFlow3D identified almost twice as many significant 3D clusters throughout the Human Protein Structurome as those found on experimentally-resolved structures. These significant 3D clusters incorporate a large number of potential driver mutations, many of which have never been reported before. NetFlow3D further interprets the pathway-scale effects of the potential driver mutations using a network propagation model. Unlike canonical PPI network analyses, NetFlow3D strictly focuses on the significant 3D clusters and emphasizes their varying PPI-perturbing effects on different interaction partners. Thus, NetFlow3D achieves a 5.5-fold higher statistical power for identifying significantly dysregulated subnetworks. The boosted power leads to the discovery of many potentially novel mechanisms underlying cancer, including the dysregulation of non-canonical PRC1, mediator complex, MCM2-7 complex, neddylation of cullins, complement system, TRiC, mitochondrial ribosome, spliceosome, etc. In addition, since NetFlow3D zooms into mutations instead of only focusing on identifying cancer-related proteins or pathways, it helps better distinguish the cancer patients with potential driver mutations from those with likely passenger mutations in the same subnetwork. Thus, NetFlow3D provides useful information for accurately predicting which set of cancer patients may have better responses to the treatments specific to the dysregulation of which subnetwork. NetFlow3D can also help explore new potential uses of existing drugs which target specific mechanisms but can only act on very few proteins. Our web server enables the broader biomedical community to access NetFlow3D and the findings of our pan-cancer study with greater ease and accessibility.

The application of NetFlow3D is not limited to somatic mutations in cancer. Instead, NetFlow3D can be generalized to interpret the genomically-driven mechanisms underlying other human diseases. In addition to cancer, spatial clustering of germline variants on protein structures has

also been observed in neurodevelopmental disorders (*63*, *64*) as well as in large collections of human diseases (*19*). Therefore, it is a generalizable idea to integrate (i) the genomic data of specific diseases, (ii) the 3D structural information of proteins and PPIs existed in the corresponding tissues/cell types, and (iii) the topology of tissue/cell type-specific PPI networks, to understand the multiscale functional effects of genomic variants in specific diseases. The growing numbers of genomically sequenced disease samples are constantly providing increasing statistical power for such implementations. The context-specific PPI networks can be inferred by integrating the curated compilation of high-quality human PPIs such as HINT (*65*) and the tissue/cell type-specific gene expression data. To make NetFlow3D adaptable to user-specified contexts, both our interactive web server and command line tool allow users to customize the genomic data as well as the context-specific gene expression and interactome data. For rare cancer types and diseases, the insufficient number of genomically sequenced samples may still limit the power of NetFlow3D. Considering that functional NT mutations only affect particular regions in a protein although this protein might be involved in many diseases, a future direction to understand the mechanisms of rare diseases could be to: (i) use the genomic data of large collections of diseases to identify 3D clusters; (ii) use these 3D clusters to prioritize potential functional mutations in the disease of interest; (iii) characterize the prioritized mutations in the corresponding tissue/cell type-specific PPI network.

As with any computational approaches, the performance of NetFlow3D is limited by the quality of available resources, especially the deep-learning-generated single protein structures and PPI interfaces. Further improvement can be anticipated by using more accurately resolved protein structures, particularly the 3D structures of multi-chain protein complexes. Although state-of-the-art tools such as AlphaFold-Multimer (*24*) have been able to predict protein complexes with atomic accuracy, a comprehensive resource that covers the 3D structures of all known human PPIs is still not available. The bottleneck is that the prediction of protein complexes is computationally intensive. For that reason, we used PIONEER in this study, which generates interactome-wide protein-protein interface predictions at the highest resolution possible. To ensure that the 3D clustering algorithm in NetFlow3D is independent of the noise in the Human Protein Structurome, we only used experimentally-determined structures to evaluate the 3D clustering performance. It is foreseeable that NetFlow3D will achieve better performance if higher-quality resources of 3D structural information become available in the near future.

Materials and Methods

Data collection and preprocessing

Somatic mutation data were obtained from Chang et al. (*1*) (Download link: https://github.com/taylor-lab/hotspots). We restricted our focus to the TCGA tumor samples of 19 cancer types, because for these 19 cancer types, the RNA-seq data of matched normal tissues could be found on the GDC data portal (*66*) (https://portal.gdc.cancer.gov/). The RNA expression data is required to filter out likely unexpressed mutations. A single canonical effect per mutation was reported using Variant Effect Predictor (VEP) version 107 (*67*) (https://grch37.ensembl.org/Homo_sapiens/Tools/VEP). According to the VEP annotations, we only retained the protein-altering mutations, including loss-of-function (LOF) mutations (Consequence: frameshift_variant, stop_gained, stop_lost, start_lost, splice_acceptor_variant, splice_donor_variant, splice_donor_5th_base_variant) and non-truncating (NT) mutations (Consequence: missense_variant, inframe_deletion, inframe_insertion). Mutations whose

consequences included both LOF and NT terms were considered LOF mutations. We then mapped the mutations to UniProt entries based on the ENSP IDs, ENST IDs, ENSG IDs, and gene symbols annotated by VEP. The mapping was implemented using the ID mapping tool on the UniProt website (https://www.uniprot.org/id-mapping) (*68*). Each mutation was initially mapped to UniProt entries using its corresponding ENSP ID. In cases where multiple UniProt entries were matched, the first entry in alphabetical order was selected. If no matches were found, the ENST ID, ENSG ID, and gene symbol were sequentially used to map the mutation to UniProt entries. We then excluded the mutations that could not be mapped to any UniProt entries. To remove likely germline variants misattributed as somatic mutations, we excluded the mutations with non-zero frequency in any populations in NHLBI-Exome Sequencing Project (ESP) or Genome Aggregation Database (gnomAD) according to the VEP annotations. Considering that mutations arising in genes not expressed in the corresponding tissues are less likely to exert biological impact, we only retained the mutations in the genes with RNA expression levels >=1 FPKM in >=80% of normal tissue samples near the tumors of the corresponding cancer types. As used by Leiserson et al. (*16*), mutations in 18 well-known cancer genes (AR, CDH4, EGFR, EPHA3, ERBB4, FGFR2, FLT3, FOXA1, FOXA2, MECOM, MIR142, MSH4, PDGFRA, SOX1, SOX9, SOX17, TBX3, WT1) that have low transcript detection levels were exempted from the expression-based filtering. The resulting dataset was used as the input to NetFlow3D and for downstream analysis.

Construction of the Human Protein Structurome

Experimentally-resolved structures. Experimentally-resolved structures (asymmetric units) were obtained from the Protein Data Bank (PDB, http://www.rcsb.org/) (*28*, *29*). Residue-level mapping between UniProt and PDB entries were obtained from the Structure Integration with Function, Taxonomy and Sequences (SIFTS, https://www.ebi.ac.uk/pdbe/docs/sifts/index.html) (*69*, *70*). Based on the PDB structures, we constructed two undirected graphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$. G_1 describes the physical contacts between residues in the same polypeptide chains, while G_2 describes the physical contacts between residues in different polypeptide chains. V_1 includes the UniProt residues covered by at least one PDB structure. E_1 is the set of residue pairs in the same polypeptide chains whose minimal three-dimensional (3D) distances among all relevant PDB structures are no larger than 6Å. The 3D distance between two residues in a given PDB structure is defined as the euclidean distance between their closest atoms in that structure. E_2 is the set of inter-chain residue pairs whose minimal 3D distances are no larger than 9Å. V_2 is the set of residues involved in E_2. G_1 and G_2 were added to the Human Protein Structurome.

3D structural information resolved by deep learning algorithms. Predicted single protein structures were obtained from the AlphaFold Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk/) (*25*, *30*), from which we constructed $G_3=(V_3,E_3)$ following the same procedures used for constructing G_1. Residues with all levels of model confidence in these structures were taken into account. G_3 was added to the Human Protein Structurome. Moreover, the predicted interface residues of 146,137 known human protein-protein interactions (PPIs) were obtained from the Protein-protein InteractiON IntErfacE pRediction (PIONEER). We used "very high" confidence predictions from PIONEER. This dataset was added to the Human Protein Structurome.

NetFlow3D: Identifying 3D clusters of mutated residues

Intra-protein clusters. NT mutations were mapped to G_1 and G_3 respectively. Vertices affected by these mutations and the edges between them were extracted as subgraph g_1 and g_3. C_1 and C_3 are the sets of connected components in g_1 and g_3, which were considered intra-protein clusters identified based on PDB and AlphaFold DB structures respectively.

Inter-protein clusters. We obtained 119,526 high-quality binary PPIs for Homo Sapiens from HINT (http://hint.yulab.org/) (*65*). This dataset was released in August 2021. We restricted our focus to these PPIs (denoted as H) for the inter-protein 3D cluster identification. (i) Identification based on PDB structures: e_2 is a subset of edges in G_2 whose endpoints are both affected by NT mutations. For a PPI between protein A and B, $g_{1A}=(v_{1A},e_{1A})$ and $g_{1B}=(v_{1B},e_{1B})$ are subgraphs extracted from g_1 incorporating the vertices and edges in A and B, respectively. e_{2AB} is a subset of e_2 where each edge connects one vertex in v_{1A} and one vertex in v_{1B}. In the merged graph $g_{2AB}=(v_{1A} \cup v_{1B}, e_{1A} \cup e_{1B} \cup e_{2AB})$, C_{2AB} is the set of connected components having at least one edge in e_{2AB}. These connected components were considered inter-protein clusters between A and B. $C_2 = \bigcup_{\{A,B\} \in H} C_{2AB}$ represents all inter-protein clusters identified based on PDB structures. (ii) Identification based on PIONEER-predicted PPI interface residues: For a PPI between protein A and B, NT mutations were mapped to the interface residues predicted by PIONEER. If the set of mutated interface residues included residues in both A and B, it was considered an inter-protein cluster, denoted as C_{4AB}. $C_4 = \left\{C_{4AB} | \{A, B\} \in H\right\}$ represents all inter-protein clusters identified based on PIONEER-predicted PPI interface residues.

NetFlow3D: Determination of cluster significance

We accounted for the nucleotide contexts and gene-specific mutation rates when estimating residue-wise background mutation rates. To address the varying mutability of different nucleotide contexts, we obtained the estimated mutability of 32 possible trinucleotides in the coding regions of cancer genomes from Chang et al. (table S57) (*1*). To account for the gene-specific mutation rates, we calculated the expected trinucleotide mutability in a given gene g using the equation below, as used by Chang et al. (*1*):

$$m_g = \sum_t \frac{N_{t,g} m_t}{L_g} \quad (1)$$

m_t is the mutability of trinucleotide t. $N_{t,g}$ is the number of occurrences of trinucleotide t in gene g. L_g is the length of gene g in nucleotides. The m_g's of genes encoding the same UniProt entry u were concatenated by:

$$m_u = \frac{\sum_{g \in U} m_g}{|U|} \quad (2)$$

U is the set of genes encoding the UniProt entry u. The relative mutational landscape within the UniProt entry u is then described by:

$$r_{t,u} = \frac{m_t}{m_u} \quad (3)$$

We then calculated the average mutation rate of a residue in the UniProt entry u as:

$$M_u = \frac{n_u}{s l_u |U|} \quad (4)$$

n_u is the number of NT mutations affecting the UniProt entry u across s tumor samples. l_u is the length of the UniProt entry u in amino acids. To avoid overestimating M_u for a given UniProt entry u with one or a few dominant hotspot residues, we calculated a truncated form of M_u by removing these residues from the UniProt entry u. A dominant hotspot residue j in a UniProt entry u is defined by the mutation count n_j across all tumor samples:

$$n_j > max(99th\ percentile\ of\ \{n_j | j \in D_u\}, 20|U|) \quad (5)$$

D_u is the set of residues affected by NT mutations in the UniProt entry u. We estimated the background mutability B_j of residue j in the UniProt entry u encoded by trinucleotide t as:

$$B_j(t, u) = r_{t,u} M_u \quad (6)$$

Considering that B_j's of the residues in UniProt entries encoded by rarely mutated genes are exceedingly small, we controlled false positives by allowing B_j to get no smaller than the 20th percentile of all B_j's calculated for the residues affected by NT mutations. Equation (3)-(6) were adapted from Chang et al. (*1*). The expected mutability of a 3D cluster C was then calculated as:

$$B_C = 1 - \prod_{j \in C} (1 - B_j) \quad (7)$$

The significance of C was determined by the one-sided p value of a binomial test:

$$p_C = \sum_{i=k}^{s} \binom{s}{i} B_C^{\,i} \left(1 - B_C\right)^{s-i} \quad (8)$$

k is the number of tumor samples where the 3D cluster C is affected by at least one NT mutation. The binomial test was applied to all 3D clusters and Bonferroni correction was applied to the one-sided p values of 3D clusters in C_1, C_2, C_3, and C_4 respectively. 3D clusters with $p_C < 0.05$ after multiple testing correction were considered significant, and thus considered indicators of positive selection in the evolution of cancer. Before outputting significant 3D clusters, we removed redundancy from them. Specifically, for a significant intra-protein cluster C from C_3 within the UniProt entry u, if at least one significant intra-protein cluster from C_1 is also found within u, and at least one residue in C is covered by at least one PDB structure, C will be removed. However, for a PPI between protein A and B, the set of significant inter-protein cluster(s) identified based on PDB structures C_{2AB} and the significant inter-protein cluster identified based on PIONEER-defined PPI interface residues C_{4AB} were not considered redundant. This is because C_{2AB} and C_{4AB} aggregate positive selection signals in different ways. On one hand, C_{2AB} accounts for the mutated residues that are not directly involved in the PPI interface but modulate the PPI through allosteric effects. On the other hand, C_{4AB} combines the mutated residues that may not be in physical contact with each other but are all on the interface between the same PPI. Therefore, C_{2AB} and C_{4AB} can jointly capture positive selection signals from complementary perspectives.

<u>NetFlow3D: Identifying proteins enriched for LOF mutations</u>

For a UniProt entry u with X_u LOF mutations across all tumor samples, the enrichment of LOF mutations in u was calculated as:

$$Enr_u = \frac{X_u / \sum_{w \in P} X_w}{l_u / \sum_{w \in P} l_w} \quad (9)$$

P is the set of UniProt entries that have at least one LOF mutation. The logarithm of Enr_u has a standard error calculated by the delta method:

$$SE\left(ln\left(Enr_u\right)\right) = \sqrt{\frac{1}{X_u} - \frac{1}{\sum_{w \in P} X_w} + \frac{1}{l_u} - \frac{1}{\sum_{w \in P} l_w}} \quad (10)$$

A Z-score was then calculated as:

$$Z_u = \frac{ln\left(Enr_u\right)}{SE\left(ln\left(Enr_u\right)\right)} \quad (11)$$

A two-tailed P-value was generated based on Z_u. Benjamini-Hochberg correction was applied to the P-values of all UniProt entries in P. The UniProt entries with Q-values<0.05 and Enr_u>1 were considered significantly enriched for LOF mutations.

<u>NetFlow3D: Network propagation model</u>

Construction of the PPI network. The initial PPI network was built out of the aforementioned high-quality binary human PPIs from HINT. The unexpressed proteins were then removed from the PPI network. Proteins encoded by the genes with RNA expression levels >=1 FPKM in >=80% of normal tissue samples for at least one cancer type were considered expressed. Proteins encoded by the aforementioned 18 well-known cancer genes with low transcript detection levels were retained. The resulting PPI network was represented by an undirected graph $G_{PPI}=(V_{PPI},E_{PPI})$.

Heat definition. The initial amount of heat placed on a protein was determined by how many types of positive selection signals were identified. If either significant 3D cluster(s) or LOF enrichment signal was identified, the protein was assigned one unit of heat. If both types of signals were identified, two units of heat were placed on it. Otherwise no heat was placed on it.

Anisotropic heat diffusion. Our heat diffusion model was adapted from HotNet2 (*16*) but had major improvements. In our model, proteins in the PPI network pass to and receive heat from their neighbors at each time step, while retaining a fraction β of their heat. Notably, when a protein transfers its remaining 1-β fraction of heat to its neighbors, the heat is unevenly distributed. To quantify this effect, we defined a ratio α of heat delivered along PPIs with and without significant 3D clusters on the binding interfaces. The model is run until equilibrium is reached. If a unit of heat is placed on protein j, the net heat transferred from protein j to protein i is given by the (i, j) entry of the diffusion matrix F defined by:

$$F = \beta(I - (1 - \beta)W)^{-1} \quad (12)$$

where

$$W_{i,j} = \frac{w_{i,j}}{\sum_{k \in Z_j} w_{k,j}} \quad (13)$$

$w_{i,j} =$ α if significant 3D cluster(s) is on the binding interface between protein i and j,

$$w_{i,j} = 1 \text{ otherwise. } (14)$$

Z_j refers to the neighbors of protein j. The initial heat distribution is described by a diagonal matrix D_h where the (i, i) entry is the amount of heat placed on protein i. The exchanged heat matrix E is then defined by:

$$E = FD_h \ (15)$$

$E(i, j)$ is the net heat transferred from protein j to protein i, which is determined by the initial amount of heat placed on protein j, the topology of PPI network, the global arrangement of significant inter-protein 3D clusters throughout the PPI network, the ratio α, and the insulating parameter β. We interpreted $E(i, j)$ as the overall effect of potential driver mutations in protein j on protein i.

Identification of subnetworks with strong internal heat exchanges. We constructed a weighted directed graph based on E, as was done by HotNet2 (*16*). If $E(i, j) > \delta$, there was a directed edge from protein j to protein i of weight $E(i, j)$. We then identified strongly connected components in this graph. A strongly connected component c in a directed graph is a set of vertices such that for every pair u,v of vertices in c there is a path from u to v. The strongly connected components containing at least two proteins were considered subnetworks with strong internal heat exchanges. The dysregulation of such a subnetwork might be the converging outcome of the potential driver mutations across different proteins in this subnetwork. The subnetworks containing >5 proteins were considered significant (see *Parameter determination*). In the main text, these subnetworks were referred to as "significantly dysregulated subnetworks". If there is a directed edge that connects a protein in subnetwork 1 to a protein in another subnetwork 2, it implies a one-way effect by which the dysregulation of subnetwork 1 influences subnetwork 2.

Parameter determination. (i) Insulating parameter β: We used $\beta = 0.5$, as used by HotNet2 (*16*). (ii) Heat delivery ratio α: we used $\alpha = 5$, which was determined based on experimental evidence. Studies of known disease mutations have shown a 70.6% rate of PPI-perturbing effects for missense mutations at the interfaces, as well as a 13.3% rate for non-interface mutations (*31*, *32*). These findings suggest that functional missense mutations on the partner-specific PPI interfaces are ~5 times more likely to perturb the corresponding interactions than the interactions with other partners. (iii) Edge weight parameter δ: The rationale behind selecting a delta is based on the fact that randomized data will typically not yield large subnetworks. Therefore, choosing an appropriate value of delta can help identify subnetworks that are statistically significant and likely to be biologically relevant. To generate a random undirected graph G_{random}, we randomly swapped $|E_{PPI}|$ edge pairs in G_{PPI} while keeping the initial amount of heat on each protein fixed. The weights of all edges in G_{random} were reset to 1. Edge swapping was used to maintain the degree of each protein constant during the randomization process. We then applied the aforementioned heat diffusion model to G_{random} and identified the minimum δ such that all subnetworks with strong internal heat exchanges had size ≤ 5. We generated 20 such random directed graphs and identified a δ for each of them. We used the smallest value among these δ's as the final value of δ.

In order to assess the robustness of our heat diffusion model, we intentionally introduced perturbations to each of the model's parameters (namely α, β, and δ) by plus or minus 5% and 10%, while keeping the values of the remaining parameters fixed at their previously determined values. We then assessed the effects of these perturbations by examining the percentage of

subnetworks that exhibited drastic changes. To determine whether a subnetwork reported by the original model was unchanged, we compared it to subnetworks reported by the perturbed models using the Jaccard index, which measured the degree of overlap between two sets of proteins. Specifically, we defined a subnetwork reported by the original model as unchanged if another subnetwork reported by the perturbed model had a Jaccard index greater than 0.5 with the original subnetwork. Our results demonstrate that, for all three parameters tested, our model exhibits strong robustness (Fig. S1).

Standard PPI network analyses

The initial amount of heat placed on each protein in G_{PPI} was determined by the number of tumor samples where the protein had mutation(s). Both NT and LOF mutations were accounted for. The heat delivery ratio α was set to 1. The remaining settings used for the standard PPI network analyses were identical to those used in the heat diffusion model described earlier.

Compiling benchmark gene sets

Known cancer genes. A list of 729 known cancer genes were obtained from The Cancer Gene Census (CGC) (https://cancer.sanger.ac.uk/census, 02/27/2022 release) (*34*, *35*). This list combines tier 1 and tier 2 genes in CGC.

Non-cancer-associated genes. Non-cancer-associated genes were compiled from three sources: (i) 1,297 genes from Reva et al. (*36*) in their category iv, i.e., genes with no functional mutations and no available associations with cancer; (ii) 129 genes annotated as "nonfunctional" by Saito et al. (*37*), including genes frequently affected by passenger hotspot mutations and olfactory genes; (iii) 194 genes confidently under neutral selection in human cancers identified by Hess et al. (*38*). By combining these three datasets we got a total of 1,574 unique genes, from which we removed 47 genes in CGC. The remaining 1,527 genes were considered non-cancer-associated.

Comparison with other 3D clustering algorithms

We applied four state-of-the-art 3D clustering algorithms to our pre-processed TCGA dataset, namely HotSpot3D (*9*), 3D hotspots (*11*), HOTMAPS (*13*), and PSCAN (*10*). Given that these approaches compiled protein structures from different resources but they all used PDB (*28*, *29*), we restricted the focus to the 3D clusters identified based on PDB structures to make fair comparisons. For all four approaches, default parameters were used if not specified. We applied HotSpot3D (*9*) to our mutation data through the HotSpot3D web server (*71*). For PSCAN (*10*), we tested both the mean and variance of the genetic effects within each scan window using the PSCAN R package. PSCAN input files were generated from SCORE-Seq (*72*) as suggested. For the mutations in each gene, SCORE-Seq was applied to the corresponding genotypes in the affected tumor samples and their matched normal samples. When specifying SCORE-Seq parameters, we set the minor allele frequency (MAF) upper bound to 1 and minor allele count (MAC) lower bound to 0 to include all mutations.

Compiling benchmark mutation sets

Known driver mutations. We compiled a list of 3,569 known driver mutations from the Integrative Onco Genomics (IntOGen) (https://www.intogen.org/) (*39*). Specifically, we used the catalog of driver mutations released in May 2016 (*40*). We selected the NT mutations whose

driver label is "known", i.e., validated tumorigenic mutations according to the state-of-the-art clinical and experimental data.

Known passenger mutations. A list of 927 known passenger mutations were obtained from The database of Cancer Passenger Mutations (dbCPM) (http://www.xialab.info:8080/dbCPM/) (*41*). We combined level 1 and level 2 mutations released on June 8, 2018 and only used the NT mutations.

Compiling benchmark pathways

Cancer signaling pathways. 32 manually curated cancer signaling pathways were obtained from NetSlim (http://www.netpath.org/netslim) (*43*). Specifically, we extracted DataNode from the GPML file of each pathway, from which we excluded DataNode without type or whose type is "Metabolite" or "Complex".

General biological processes. 7,530 Gene Ontology (GO) biological processes were obtained from the Molecular Signatures Database (MSigDB) C5 collection (http://www.gsea-msigdb.org/gsea/msigdb) (*44–48*). To ensure that the selected biological processes were not specifically associated with cancer, we removed GO biological processes containing genes already documented as cancer genes by CGC. Additionally, we excluded GO biological processes that did not contain any protein-coding genes. After this filtering, we were left with a final set of 2,055 GO biological processes for downstream analysis.

Mutation enrichment patterns in the significantly dysregulated subnetworks

We calculated the enrichment of two categories of mutations for each significantly dysregulated subnetwork: (i) enrichment of NT mutations incorporated by the significant 3D clusters; (ii) enrichment of all NT mutations. Enrichment is calculated as the ratio of the observed fraction of mutations in a given subnetwork over the relative length of proteins in this given subnetwork. Relative length is defined as the total sequence length of the proteins in a subnetwork divided by the total sequence length of all proteins with NT mutations. The fold change of enrichment for each subnetwork was characterized by the ratio between (i) and (ii). This analysis was also applied to the cancer signaling pathways and GO biological processes compiled before.

Web server

NetFlow3D tool and the pan-cancer results generated in this study can be accessed from http://netflow3d.yulab.org.

References and Notes

1. M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, C. Kandoth, J. Gao, N. D. Socci, D. B. Solit, A. B. Olshen, N. Schultz, B. S. Taylor, Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).

2. H. Zhang, M. S. Xu, X. Fan, W. K. Chung, Y. Shen, Predicting functional effect of missense variants using graph attention neural networks. *Nature Machine Intelligence*. **4**, 1017–1028 (2022).

3. N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, W. Sieh, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

4. K. E. Samocha, J. A. Kosmicki, K. J. Karczewski, A. H. O'Donnell-Luria, E. Pierce-Hoffman, D. G. MacArthur, B. M. Neale, M. J. Daly, Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017), p. 148353.

5. L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, K. K.-H. Farh, Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

6. H. Qi, H. Zhang, Y. Zhao, C. Chen, J. J. Long, W. K. Chung, Y. Guan, Y. Shen, MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).

7. K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, G. Bejerano, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).

8. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

9. B. Niu, A. D. Scott, S. Sengupta, M. H. Bailey, P. Batra, J. Ning, M. A. Wyczalkowski, W.-W. Liang, Q. Zhang, M. D. McLellan, S. Q. Sun, P. Tripathi, C. Lou, K. Ye, R. J. Mashl, J. Wallis, M. C. Wendl, F. Chen, L. Ding, Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).

10. Z.-Z. Tang, G. R. Sliwoski, G. Chen, B. Jin, W. S. Bush, B. Li, J. A. Capra, PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal

variant detection. *Genome Biol.* **21**, 217 (2020).

11. J. Gao, M. T. Chang, H. C. Johnsen, S. P. Gao, B. E. Sylvester, S. O. Sumer, H. Zhang, D. B. Solit, B. S. Taylor, N. Schultz, C. Sander, 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017).

12. S. Kumar, D. Clarke, M. B. Gerstein, Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 18962–18970 (2019).

13. C. Tokheim, R. Bhattacharya, N. Niknafs, D. M. Gygax, R. Kim, M. Ryan, D. L. Masica, R. Karchin, Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res.* **76**, 3719–3731 (2016).

14. M. J. Meyer, R. Lapcevic, A. E. Romero, M. Yoon, J. Das, J. F. Beltrán, M. Mort, P. D. Stenson, D. N. Cooper, A. Paccanaro, H. Yu, mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum. Mutat.* **37**, 447–456 (2016).

15. A. Kamburov, M. S. Lawrence, P. Polak, I. Leshchiner, K. Lage, T. R. Golub, E. S. Lander, G. Getz, Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5486–95 (2015).

16. M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, B. J. Raphael, Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).

17. F. Vandin, E. Upfal, B. J. Raphael, Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).

18. C. Dincer, T. Kaya, O. Keskin, A. Gursoy, N. Tuncbag, 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients. *PLoS Comput. Biol.* **15**, e1006789 (2019).

19. R. M. Sivley, X. Dou, J. Meiler, W. S. Bush, J. A. Capra, Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am. J. Hum. Genet.* **102**, 415–426 (2018).

20. F. Zheng, M. R. Kelly, D. J. Ramms, M. L. Heintschel, K. Tao, B. Tutuncuoglu, J. J. Lee, K. Ono, H. Foussard, M. Chen, K. A. Herrington, E. Silva, S. N. Liu, J. Chen, C. Churas, N. Wilson, A. Kratz, R. T. Pillich, D. N. Patel, J. Park, B. Kuenzi, M. K. Yu, K. Licon, D. Pratt, J. F. Kreisberg, M. Kim, D. L. Swaney, X. Nan, S. I. Fraley, J. S. Gutkind, N. J. Krogan, T. Ideker, Interpretation of cancer mutations using a multiscale map of protein systems. *Science*. **374**, eabf3067 (2021).

21. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr, K. W. Kinzler,

Cancer genome landscapes. *Science*. **339**, 1546–1558 (2013).

22. F. Cheng, J. Zhao, Y. Wang, W. Lu, Z. Liu, Y. Zhou, W. R. Martin, R. Wang, J. Huang, T. Hao, H. Yue, J. Ma, Y. Hou, J. A. Castrillon, J. Fang, J. D. Lathia, R. A. Keri, F. C. Lightstone, E. M. Antman, R. Rabadan, D. E. Hill, C. Eng, M. Vidal, J. Loscalzo, Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353 (2021).

23. S. Chen, R. Fragoza, L. Klei, Y. Liu, J. Wang, K. Roeder, B. Devlin, H. Yu, An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat. Genet.* **50**, 1032–1040 (2018).

24. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021), p. 2021.10.04.463034.

25. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

26. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **373**, 871–876 (2021).

27. M. J. Meyer, J. F. Beltrán, S. Liang, R. Fragoza, A. Rumack, J. Liang, X. Wei, H. Yu, Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods*. **15**, 107–114 (2018).

28. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

29. S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. Feng, S. Ganesan, D. S. Goodsell, S. Ghosh, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, C. L. Lawson, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Persikova, C. Randle, A. Rose, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, Y.-P. Tao, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, M. Zhuravleva, RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in

fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).

30. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

31. N. Sahni, S. Yi, M. Taipale, J. I. Fuxman Bass, J. Coulombe-Huntington, F. Yang, J. Peng, J. Weile, G. I. Karras, Y. Wang, I. A. Kovács, A. Kamburov, I. Krykbaeva, M. H. Lam, G. Tucker, V. Khurana, A. Sharma, Y.-Y. Liu, N. Yachie, Q. Zhong, Y. Shen, A. Palagi, A. San-Miguel, C. Fan, D. Balcha, A. Dricot, D. M. Jordan, J. M. Walsh, A. A. Shah, X. Yang, A. K. Stoyanova, A. Leighton, M. A. Calderwood, Y. Jacob, M. E. Cusick, K. Salehi-Ashtiani, L. J. Whitesell, S. Sunyaev, B. Berger, A.-L. Barabási, B. Charloteaux, D. E. Hill, T. Hao, F. P. Roth, Y. Xia, A. J. M. Walhout, S. Lindquist, M. Vidal, Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*. **161**, 647–660 (2015).

32. S. Chen, Y. Liu, Y. Zhang, S. D. Wierbowski, S. M. Lipkin, X. Wei, H. Yu, A full-proteome, interaction-specific characterization of mutational hotspots across human cancers. *bioRxiv* (2019), p. 2019.12.20.885293.

33. E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, A. Godzik, A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* **11**, e1004518 (2015).

34. Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, S. A. Forbes, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*. **18**, 696–705 (2018).

35. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

36. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).

37. Y. Saito, J. Koya, M. Araki, Y. Kogure, S. Shingaki, M. Tabata, M. B. McClure, K. Yoshifuji, S. Matsumoto, Y. Isaka, H. Tanaka, T. Kanai, S. Miyano, Y. Shiraishi, Y. Okuno, K. Kataoka, Landscape and function of multiple mutations within individual oncogenes. *Nature*. **582**, 95–99 (2020).

38. J. M. Hess, A. Bernards, J. Kim, M. Miller, A. Taylor-Weiner, N. J. Haradhvala, M. S. Lawrence, G. Getz, Passenger Hotspot Mutations in Cancer. *Cancer Cell*. **36**, 288–301.e14 (2019).

39. F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, N. Lopez-Bigas, A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*. **20**, 555–572 (2020).

40. D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, N. Lopez-Bigas, Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).

41. Z. Yue, L. Zhao, J. Xia, dbCPM: a manually curated database for exploring the cancer passenger mutations. *Brief. Bioinform.* (2018), doi:10.1093/bib/bby105.

42. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**, 248–249 (2010).

43. R. Raju, V. Nanjappa, L. Balakrishnan, A. Radhakrishnan, J. K. Thomas, J. Sharma, M. Tian, S. M. Palapetta, T. Subbannayya, N. R. Sekhar, B. Muthusamy, R. Goel, Y. Subbannayya, D. Telikicherla, M. Bhattacharjee, S. M. Pinto, N. Syed, M. S. Srikanth, G. J. Sathe, S. Ahmad, S. N. Chavan, G. S. S. Kumar, A. Marimuthu, T. S. K. Prasad, H. C. Harsha, B. A. Rahiman, O. Ohara, G. D. Bader, S. Sujatha Mohan, W. P. Schiemann, A. Pandey, NetSlim: high-confidence curated signaling maps. *Database* . **2011**, bar032 (2011).

44. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

45. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. **1**, 417–425 (2015).

46. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. **27**, 1739–1740 (2011).

47. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

48. Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

49. E. Dazert, M. N. Hall, mTOR signaling in disease. *Curr. Opin. Cell Biol.* **23**, 744–755 (2011).

50. A. Villanueva, D. Y. Chiang, P. Newell, J. Peix, S. Thung, C. Alsinet, V. Tovar, S. Roayaie,

B. Minguez, M. Sole, C. Battiston, S. Van Laarhoven, M. I. Fiel, A. Di Feo, Y. Hoshida, S. Yea, S. Toffanin, A. Ramos, J. A. Martignetti, V. Mazzaferro, J. Bruix, S. Waxman, M. Schwartz, M. Meyerson, S. L. Friedman, J. M. Llovet, Pivotal role of mTOR signaling in hepatocellular carcinoma. *Gastroenterology*. **135**, 1972–83, 1983.e1–11 (2008).

51. B. Tessoulin, A. Papin, P. Gomez-Bougie, C. Bellanger, M. Amiot, C. Pellat-Deceunynck, D. Chiron, BCL2-Family Dysregulation in B-Cell Malignancies: From Gene Expression Regulation to a Targeted Therapy Biomarker. *Front. Oncol.* **8**, 645 (2018).

52. I. Comet, E. M. Riising, B. Leblanc, K. Helin, Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat. Rev. Cancer*. **16**, 803–810 (2016).

53. A. Piunti, A. Shilatifard, The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.* **22**, 326–345 (2021).

54. M. F. Bakhoum, J. H. Francis, A. Agustinus, E. M. Earlie, M. Di Bona, D. H. Abramson, M. Duran, I. Masilionis, E. Molina, A. N. Shoushtari, M. H. Goldbaum, P. S. Mischel, S. F. Bakhoum, A. M. Laughney, Loss of polycomb repressive complex 1 activity and chromosomal instability drive uveal melanoma progression. *Nat. Commun.* **12**, 5402 (2021).

55. V. van den Boom, H. Maat, M. Geugien, A. Rodríguez López, A. M. Sotoca, J. Jaques, A. Z. Brouwers-Vos, F. Fusetti, R. W. J. Groen, H. Yuan, A. C. M. Martens, H. G. Stunnenberg, E. Vellenga, J. H. A. Martens, J. J. Schuringa, Non-canonical PRC1.1 Targets Active Genes Independent of H3K27me3 and Is Essential for Leukemogenesis. *Cell Rep.* **14**, 332–346 (2016).

56. V. Parreno, A.-M. Martinez, G. Cavalli, Mechanisms of Polycomb group protein function in cancer. *Cell Res.* **32**, 231–253 (2022).

57. L. M. Kutscher, K. Okonechnikov, N. V. Batora, J. Clark, P. B. G. Silva, M. Vouri, S. van Rijn, L. Sieber, B. Statz, M. D. Gearhart, R. Shiraishi, N. Mack, B. A. Orr, A. Korshunov, B. L. Gudenas, K. S. Smith, A. L. Mercier, O. Ayrault, M. Hoshino, M. Kool, K. von Hoff, N. Graf, G. Fleischhack, V. J. Bardwell, S. M. Pfister, P. A. Northcott, D. Kawauchi, Functional loss of a noncanonical BCOR-PRC1.1 complex accelerates SHH-driven medulloblastoma formation. *Genes Dev.* **34**, 1161–1176 (2020).

58. D. Shinoda, Y. Nakajima-Takagi, M. Oshima, S. Koide, K. Aoyama, A. Saraya, H. Harada, B. Rahmutulla, A. Kaneda, K. Yamaguchi, Y. Furukawa, H. Koseki, K. Shimoda, T. Tanaka, G. Sashida, A. Iwama, Insufficiency of non-canonical PRC1 synergizes with JAK2V617F in the development of myelofibrosis. *Leukemia*. **36**, 452–463 (2022).

59. Y. Isshiki, Y. Nakajima-Takagi, M. Oshima, K. Aoyama, M. Rizk, S. Kurosawa, A. Saraya, T. Kondo, E. Sakaida, C. Nakaseko, K. Yokote, H. Koseki, A. Iwama, KDM2B in polycomb repressive complex 1.1 functions as a tumor suppressor in the initiation of T-cell leukemogenesis. *Blood Adv*. **3**, 2537–2549 (2019).

60. X. Zhu, M. Yan, W. Luo, W. Liu, Y. Ren, C. Bei, G. Tang, R. Chen, S. Tan, Expression and clinical significance of PcG-associated protein RYBP in hepatocellular carcinoma. *Oncol.*

Lett. **13**, 141–150 (2017).

61. H. Zhou, J. Li, Z. Zhang, R. Ye, N. Shao, T. Cheang, S. Wang, RING1 and YY1 binding protein suppresses breast cancer growth and metastasis. *Int. J. Oncol.* **49**, 2442–2452 (2016).

62. A.-H. Tong, J. Tan, J.-H. Zhang, F.-J. Xu, F.-Y. Li, C.-Y. Cao, Overexpression of RYBP inhibits proliferation, invasion, and chemoresistance to cisplatin in anaplastic thyroid cancer cells via the EGFR pathway. *J. Biochem. Mol. Toxicol.* **33**, e22241 (2019).

63. S. H. Lelieveld, L. Wiel, H. Venselaar, R. Pfundt, G. Vriend, J. A. Veltman, H. G. Brunner, L. E. L. M. Vissers, C. Gilissen, Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. *Am. J. Hum. Genet.* **101**, 478–484 (2017).

64. D. Lessel, D. M. Zeitler, M. R. F. Reijnders, A. Kazantsev, F. Hassani Nia, A. Bartholomäus, V. Martens, A. Bruckmann, V. Graus, A. McConkie-Rosell, M. McDonald, B. Lozic, E.-S. Tan, E. Gerkes, J. Johannsen, J. Denecke, A. Telegrafi, E. Zonneveld-Huijssoon, H. H. Lemmink, B. W. M. Cham, T. Kovacevic, L. Ramsdell, K. Foss, D. Le Duc, D. Mitter, S. Syrbe, A. Merkenschlager, M. Sinnema, B. Panis, J. Lazier, M. Osmond, T. Hartley, J. Mortreux, T. Busa, C. Missirian, P. Prasun, S. Lüttgen, I. Mannucci, I. Lessel, C. Schob, S. Kindler, J. Pappas, R. Rabin, M. Willemsen, T. Gardeitchik, K. Löhner, P. Rump, K.-R. Dias, C.-A. Evans, P. I. Andrews, T. Roscioli, H. G. Brunner, C. Chijiwa, M. E. S. Lewis, R. A. Jamra, D. A. Dyment, K. M. Boycott, A. P. A. Stegmann, C. Kubisch, E.-C. Tan, G. M. Mirzaa, K. McWalter, T. Kleefstra, R. Pfundt, Z. Ignatova, G. Meister, H.-J. Kreienkamp, Germline AGO2 mutations impair RNA interference and human neurological development. *Nat. Commun.* **11**, 5797 (2020).

65. J. Das, H. Yu, HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).

66. R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, L. M. Staudt, Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

67. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

68. UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

69. J. M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O'Donovan, M. Martin, S. Velankar, SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).

70. S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M.-J. Martin, G. J. Kleywegt, SIFTS: Structure Integration with Function,

Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483–9 (2013).

71. S. Chen, X. He, R. Li, X. Duan, B. Niu, HotSpot3D web server: an integrated resource for mutation analysis in protein 3D structures. *Bioinformatics*. **36**, 3944–3946 (2020).

72. D.-Y. Lin, Z.-Z. Tang, A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).

Acknowledgments: The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We thank members of the Yu laboratory for helpful discussions and guidance; and Zui Tao for her suggestions on the conceptualization.

Funding:

National Institutes of Health grant R01GM124559 (HY)

National Institutes of Health grant R01GM125639 (HY)

National Institutes of Health grant R01DK115398 (HY)

Simons Foundation grant 575547 (HY)

Simons Foundation grant 893926 (HY)

Author contributions:

Conceptualization: YZ, HY

Methodology: YZ, JB, AKL

Visualization: AKL, YZ, TQ

Formal analysis: YZ, AKL, LL

Data curation: YZ, AKL, JZ, SW

Funding acquisition: HY

Supervision: HY

Writing – original draft: YZ

Writing – review & editing: HY, LL, JZ, SW, AKL

Competing interests: The authors declare no competing interests.

Data and materials availability: Data that support the findings of this study are available in the main text or the supplementary materials. NetFlow3D tool and the pan-cancer results are publicly accessible via our web server (http://netflow3d.yulab.org). The source code of NetFlow3D is publicly available at https://github.com/hyulab/NetFlow3D.

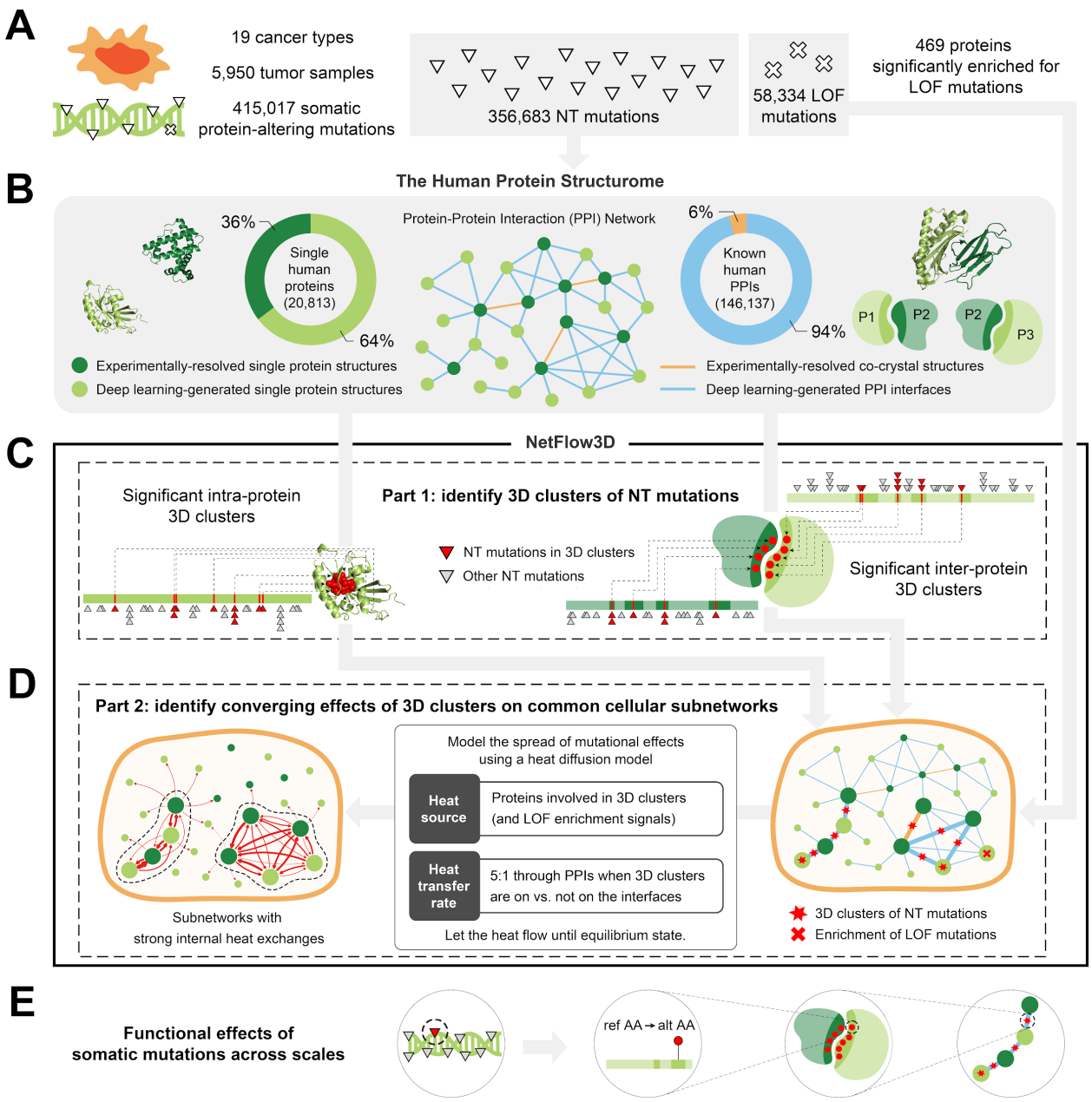

Fig. 1. Framework of mapping the multiscale functional effects of somatic mutations. (**A**) Preprocessed TCGA pan-cancer mutation dataset. (**B**) Detailed information about the Human Protein Structurome, a complete repository that incorporates the 3D structures of every single protein as well as the binding interfaces of all known PPIs in humans. (**C**) The first part of NetFlow3D, a 3D clustering algorithm for identifying both intra- and inter-protein 3D clusters of non-truncating (NT) mutations throughout the Human Protein Structurome. (**D**) The second part of NetFlow3D, a network propagation model for interpreting the pathway-scale effects of mutations in 3D clusters and identifying the converging effects on common cellular subnetworks. (**E**) The multiscale mechanistic effects of potential driver mutations mapped by NetFlow3D.

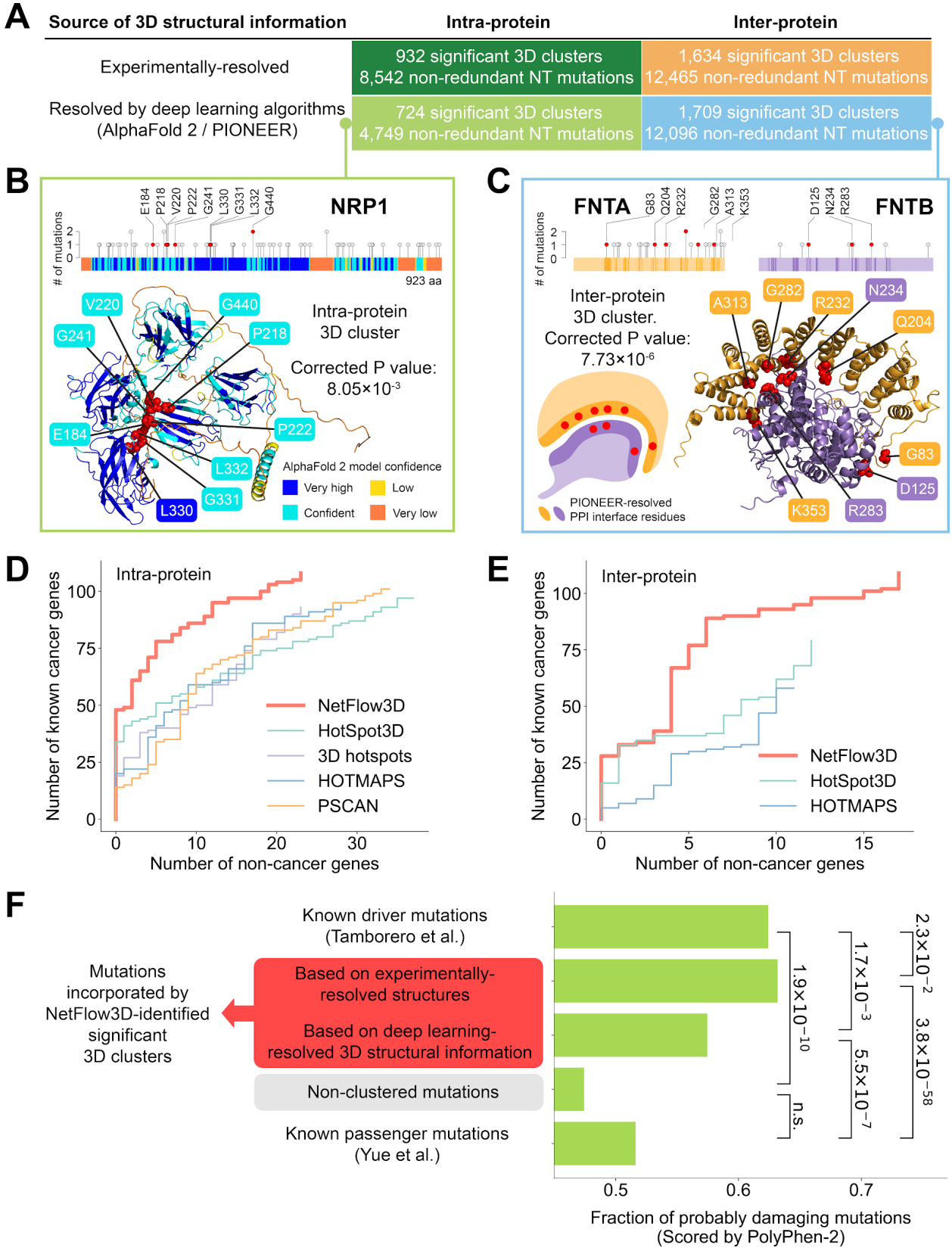
A
Source of 3D structural information
Intra-protein
Inter-protein
Experimentally-resolved
932 significant 3D clusters
8,542 non-redundant NT mutations
1,634 significant 3D clusters
12,465 non-redundant NT mutations
Resolved by deep learning algorithms
(AlphaFold 2 / PIONEER)
724 significant 3D clusters
4,749 non-redundant NT mutations
1,709 significant 3D clusters
12,096 non-redundant NT mutations
B
NRP1
# of mutations
E184
P218
V220
P222
G241
L330
G331
L332
G440
923 aa
Intra-protein
3D cluster
Corrected P value:
8.05×10-3
AlphaFold 2 model confidence
Very high
Low
Confident
Very low
C
FNTA
FNTB
G83
Q204
R232
G282
A313
K353
D125
N234
R283
Inter-protein
3D cluster.
Corrected P value:
7.73×10-6
PIONEER-resolved
PPI interface residues
D
Intra-protein
Number of known cancer genes
Number of non-cancer genes
NetFlow3D
HotSpot3D
3D hotspots
HOTMAPS
PSCAN
E
Inter-protein
NetFlow3D
HotSpot3D
HOTMAPS
F
Known driver mutations
(Tamborero et al.)
Mutations
incorporated by
NetFlow3D-identified
significant
3D clusters
Based on experimentally-
resolved structures
Based on deep learning-
resolved 3D structural information
Non-clustered mutations
Known passenger mutations
(Yue et al.)
1.9×10−10
n.s.
1.7×10−3
5.5×10−7
2.3×10−2
3.8×10−58
Fraction of probably damaging mutations
(Scored by PolyPhen-2)

Fig. 2. Description of 3D clusters identified by NetFlow3D and performance evaluation. (**A**) Summary of the significant intra- and inter-protein 3D clusters identified by NetFlow3D. (**B**) An example of a significant intra-protein 3D cluster identified on the AlphaFold 2-generated structure of NRP1 protein. All mutations incorporated by this cluster are on the residues with "very high" or "confident" model confidence. These potential driver mutations can hardly be identified by recurrence-based approaches because none of them occurs in more than two tumor samples. (**C**) An example of a significant inter-protein 3D cluster on the binding interface between FNTA and FNTB proteins. The PPI interface is resolved by PIONEER. For visualization purposes, a 3D structure of this protein complex is generated using AlphaFold Multimer (*24*) to highlight the significant inter-protein 3D cluster. The potential driver mutations in this cluster cannot be identified by looking for 3D clusters on individual proteins. (**D**) Performance comparison between NetFlow3D and state-of-the-art 3D clustering algorithms. For each curve, top 1-500 genes are plotted from bottom left to upper right. (**E**) Evaluation of molecular effects for five categories of mutations.

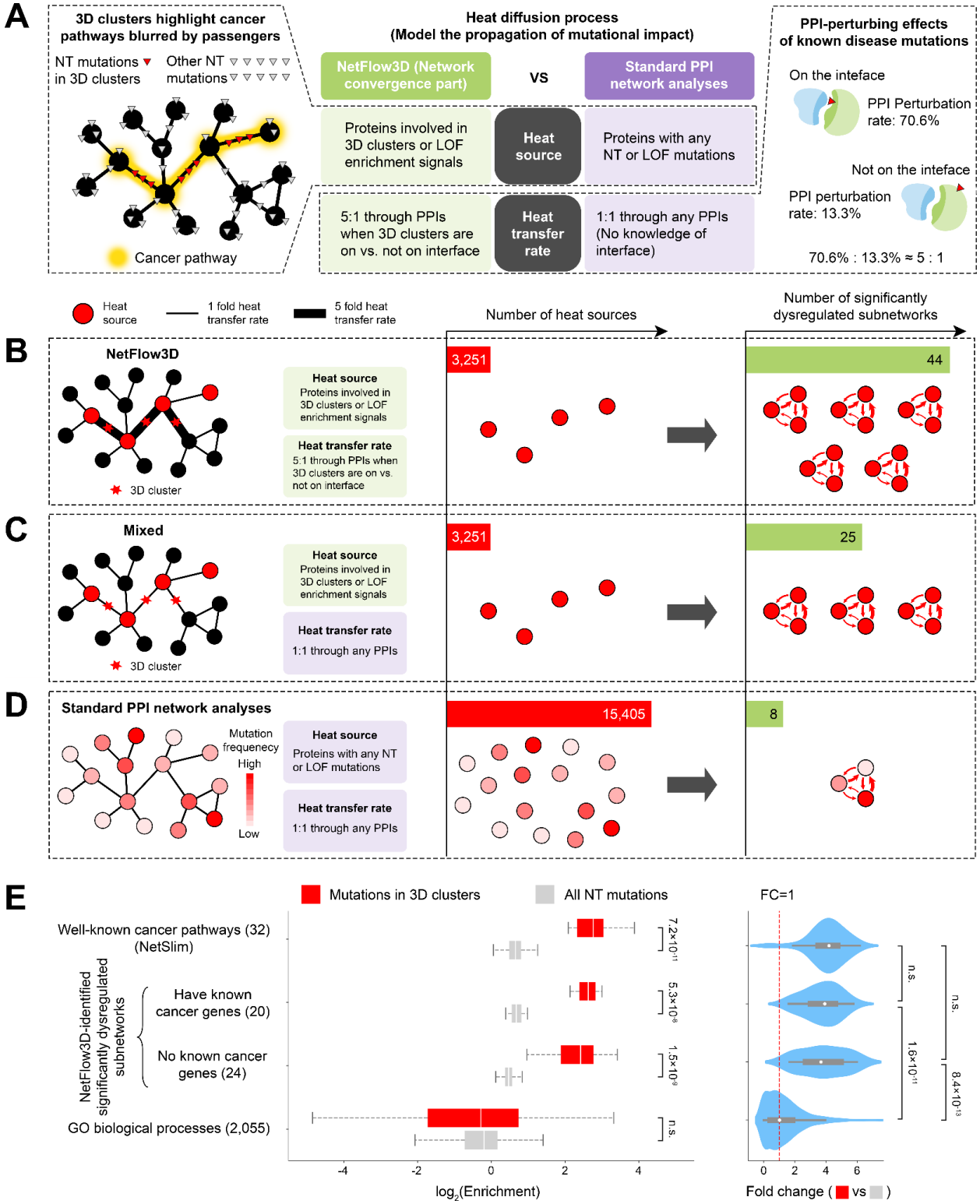

Fig. 3. Comparison between NetFlow3D and standard PPI network analyses and performance evaluation. (**A**) Key differences between NetFlow3D and standard PPI network analyses. (**B-D**) Evaluation of different PPI network-based approaches regarding their statistical

power of identifying significantly dysregulated subnetworks. (**B**) Statistical power evaluation for NetFlow3D. (**C**) Statistical power evaluation for a mixed model that is used to understand how much power of NetFlow3D is gained from its unique definition of heat source and heat transfer rate, respectively. (**D**) Statistical power evaluation for standard PPI network analyses. (**E**) Evaluation of significantly dysregulated subnetworks identified by NetFlow3D. The patterns of mutation enrichment in these subnetworks are compared with those in well-known cancer pathways and general biological processes.

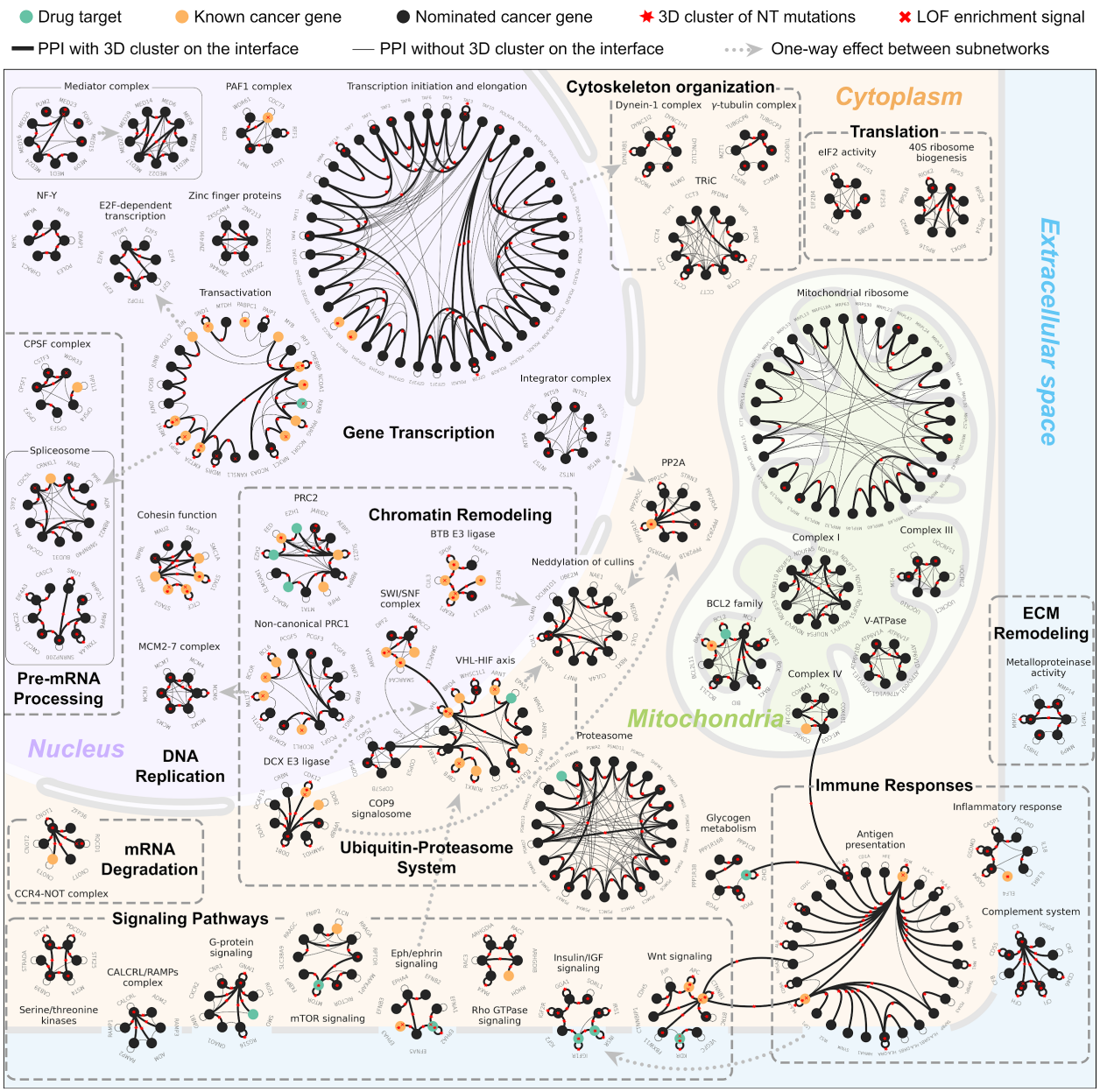

Fig. 4. A multiscale functional map of somatic mutations in cancer. The map displayed in this figure includes 44 significantly dysregulated subnetworks. A complete map that incorporates all subnetworks identified by NetFlow3D is detailed in table S2, S3, and S10. Genes whose protein products are targeted by the FDA-approved drugs are colored in green. Genes documented by the Cancer Gene Census are colored in gold.

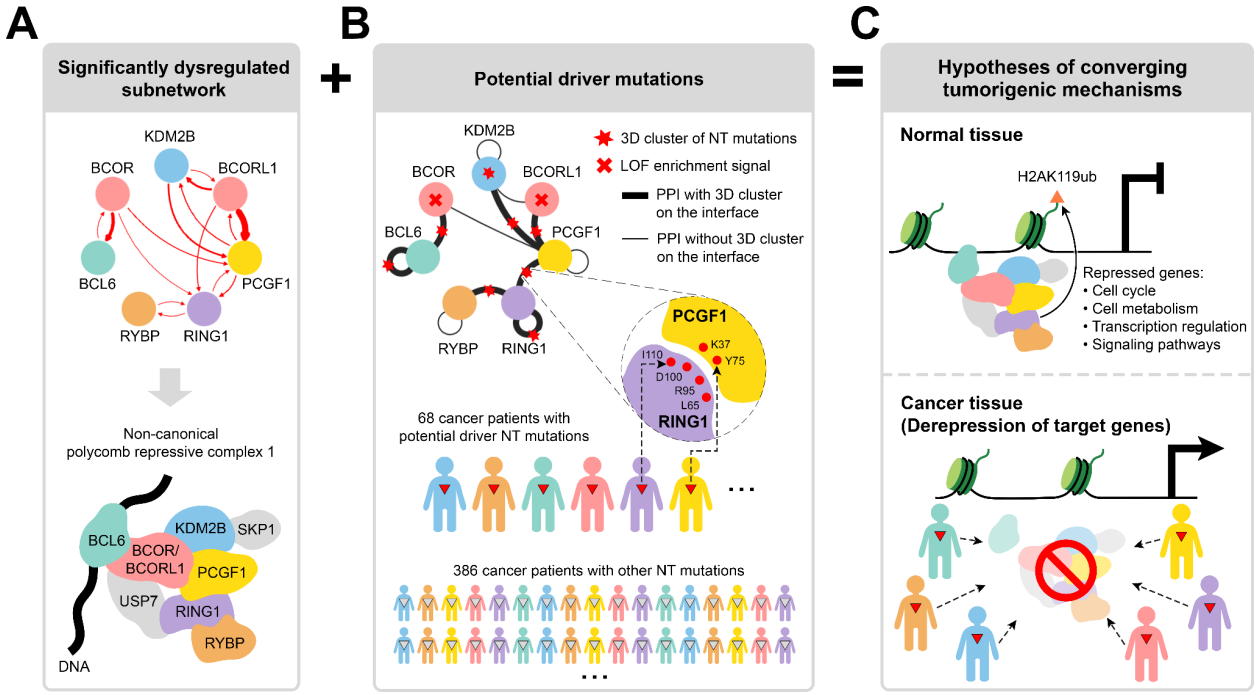

Fig. 5. Converging mechanisms across a subset of cancer patients with potential driver mutations dysregulating ncPRC1. (**A**) The biological entity corresponding to part of a significantly dysregulated subnetwork identified by NetFlow3D. Red arrows indicate the overall effects from potential driver mutations on one protein towards the other protein. Arrow thickness indicates the strength of effects. The unshown proteins in this subnetwork are mostly paralogs of the displayed proteins, including three paralogs of PCGF1 (PCGF3, PCGF5, PCGF6), and a paralog of RING1 (RNF2). Complete information about this subnetwork can be found in table S47. (**B**) Information about the somatic mutations in this subnetwork. Among the 454 cancer patients with somatic NT mutations in this subnetwork, 68 patients have potential driver mutations incorporated by significant 3D clusters. As suggested by NetFlow3D, these potential driver mutations are more likely to dysregulate ncPRC1. (**C**) Potential mechanisms underlying cancer driven by the potential driver mutations in this subnetwork. Cancer patients with diverse potential driver mutations may share similar mechanisms.

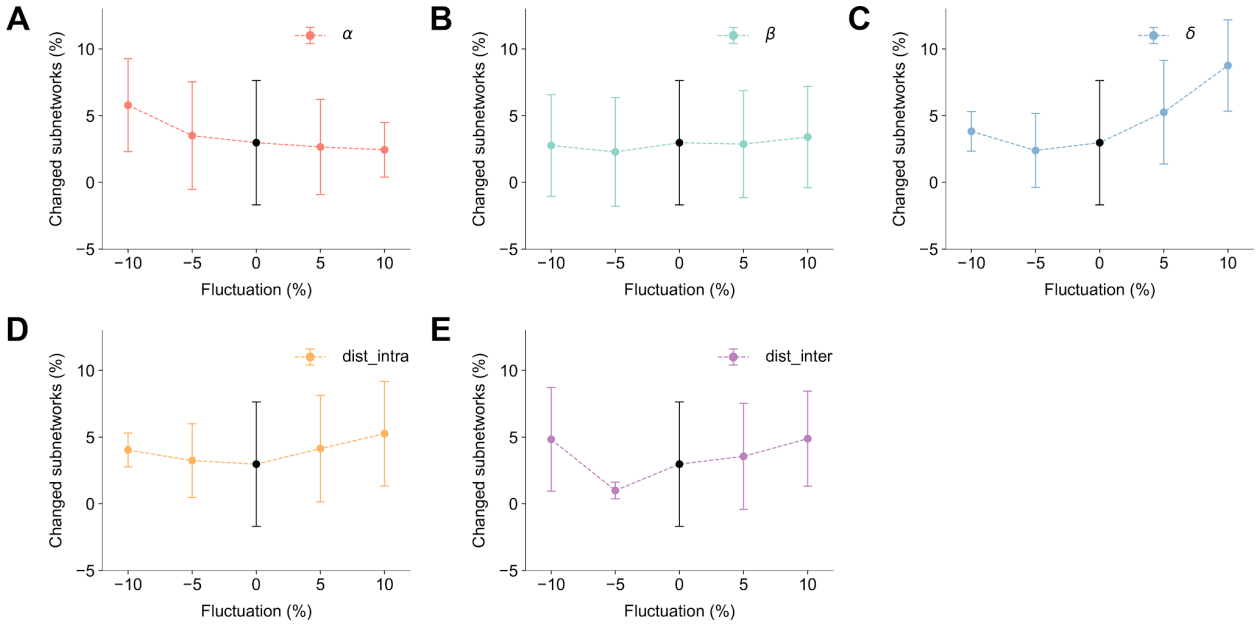

Fig. S1. Evaluation of the robustness of the heat diffusion model in NetFlow3D. (**A**) Robustness of the model in response to perturbations of the heat delivery ratio α. (**B**) Robustness of the model in response to perturbations of the insulating parameter β. (**C**) Robustness of the model in response to perturbations of the edge weight parameter δ. (**D**) Robustness of the model in response to perturbations of the distance cutoff used for defining the physical contact between a pair of residues in the same polypeptide chain. (**E**) Robustness of the model in response to perturbations of the distance cutoff used for defining the physical contact between two residues belonging to different polypeptide chains.