Narrow funnel-like interaction energy distribution is an indicator of specific protein interaction partners

Juyoung Choi (rgdfs@sogang.ac.kr)
Sogang University https://orcid.org/0000-0002-4290-998X

Article

Keywords: protein–protein interaction, kinases, E3 ubiquitin ligases, protein docking, interaction energy

Posted Date: March 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1412786/v2

Narrow funnel-like interaction energy distribution is an indicator of specific protein interaction partners

Juyoung Choi*
Department of Life Science, Sogang University, Seoul, 04017, South Korea
* Correspondence: Juyoung Choi
Email: rgdfs@sogang.ac.kr

Author Contributions: This study was conducted by the sole author mentioned above.
Competing Interest Statement: The author declares no conflict of interest.
Keywords: protein–protein interaction, kinases, E3 ubiquitin ligases, protein docking, interaction energy.

This file includes the following:

Abstract

Biological mechanisms consist of protein interaction networks. However, most of protein interaction prediction are based on interspecies biological evidences (knowledge-based predictions) or exhibits low accuracy for weak interactions and requires high computational power (in silico docking method). In this study, I suggest a novel method to identify protein interaction using interaction energy distribution. Protein interactions with specific partners can be predicted by searching narrow funnel-like interaction energy distribution. Kinases and E3 ubiquitin ligase are important signaling components. However, as their interactions are weak and various, those interaction predictions are limited. I described that their interaction partners can be identified by analyzing interaction energy distribution with low computational cost. The accuracy of this prediction was similar or even higher than those of yeast two-hybrid screening, the most widely applied interaction screening method. Furthermore, I showed that other specific protein interactions have narrow funnel-like interaction energy distribution. In summary, this knowledge-free protein interaction prediction method would broaden our understanding of protein interaction networks.

Introduction

Most biological activities occur through complex biomolecular interaction networks. After the genome sequencing of several tens of thousands of organisms, understating biomolecule networks beyond the individual genes is becoming increasingly important. In particular, protein–protein interactions (PPI) are among the most diverse biomolecular networks and the center of system biology. Several experimental techniques, such as yeast two-hybrid screening, coimmunoprecipitation, affinity-chromatography, tandem affinity purification-mass spectroscopy, etc., have been developed to identify PPI [1]. However, with the rapid advancement of proteome identification, experimental results cannot keep up with the potential number of protein combinations. Recently, 200 million proteins in various organisms were predicted or identified [2]. However, international consortium curated only approximately 1 million PPIs from literatures and datasets [3,4]. Moreover, among the identified PPI, approximately 70% are *Homo sapiens*-related [4]. Most PPI might still be unidentified. Therefore, *in silico* methods, as well as high throughput experimental approaches, have been developed to broaden our PPI-related knowledge [5].

In silico PPI prediction methods use biological or structural and physical results. Biological pieces of evidence include gene or domain fusion [6-8], gene neighborhood [6,9], interolog [10,11], coexpression [6,12], coevolution [13,14], and phylogenetic similarity [6,15]. Although biological result-based PPI predictions have large-scale and powerful predictive ability, they are limited to well-known PPI. To predict PPI using biological results, omics scale interspecies data are required, or at least similar protein interactions in other species should be identified [5,16]. Structural and physical evidence-based PPI prediction uses protein–protein interface templates [17,18], interaction energy [16,19], and shape complementarity [19-21]. Template-based PPI predictions require identified PPIs with similar interface structures [17,18]. PPI predictions with interaction energy and shape complementarity are independent of the identified PPI data. However, shape complementarity is different for each PPI type. Obligate PPIs, which are stable only when interacting, have higher shape complementarity, although transient PPIs exhibit low

shape complementarity [5,22].Protein docking programs, which calculate the possible direction of interaction and interaction energy, have a high computational cost [5,19]. Transient PPIs with weak interactions are difficult to predict using interaction energy [19]. However, most PPIs are weak, and weak PPIs play pivotal roles in protein interaction networks [23]. Therefore, high computational cost and the inability to predict transient PPIs with weak interactions represent the major drawbacks of protein interaction partner prediction using *in silico* protein docking.

One of the most difficult PPI to predict is that of enzyme-specific interaction partners, such as substrates, inhibitors and regulators, or signaling pathway interactions. Their diversity of complicated signaling pathways [24] makes biological evidence- and structural template-based PPI prediction difficult. Furthermore, due to weak and transient interaction properties, interaction energy and shape complementarity are not useful for the prediction of these PPIs [19,22]. However, understanding PPIs in signaling pathways, such as the interactions of kinases or E3 ubiquitin ligase (EUL) with their interaction partners, is one of the most crucial parts of biology. In mammals, kinases and ubiquitination-related enzymes are the most and the fifth most abundant enzymes in signaling pathways, respectively [24]. Moreover, in plants, approximately 5% of the protein-coding genes in rice and *Arabidopsis* encode kinases and EULs [25-27]. Despite the importance of kinases and EULs, most interaction partner predictions of kinases and EULs are solely based on biological pieces of evidence, which require identified PPI with similar amino acid sequences or domains [28-34]. Moreover, because of the lack of identified PPIs, most prediction tools are limited to human kinases or EULs [28,30-34]. Therefore, in most cases, no high-accuracy PPI prediction tools are available for kinases and EULs.

In this study, I suggest a new, interaction energy distribution-based PPI prediction strategy. PPIs with specific interaction partner, such as PPIs of kinase and EULs, exhibit specificity to maintain sophisticated signaling pathways. For example, in humans, kinase phosphorylates have one to hundreds of specific phosphorylation sites among approximately 700,000 potential phosphorylation sites [35]. Moreover, multiple regulators and inhibitors function when they interact with kinases only in a specific orientation [36,37].

Although the interaction between kinases or EULs and their partners is weak, the specificity of these interactions to occur in a specific orientation distinguishes them from other PPIs [35,38]. Due to this specificity, I hypothesized that an interaction energy diagram PPIs with specific interaction partners would display a narrow funnel-like landscape, exhibiting a stable state in a specific orientation (Fig. 1a). In this study, using Rosetta energy, knowledge (or Boltzmann relation)-based macromolecular energy function [39], I calculated PPI interaction energy distributions. Finally, I suggest criteria for specific protein interaction partner predictions.

Results

Workflow for the interaction energy distribution of kinases/E3 ubiquitin ligases and their interaction partners

To analyze the interaction energy landscape of kinases/E3 ubiquitin ligases and their interaction partners, I retrieved 135 kinase-interaction partners (regulator, inhibitor, and substrate) and 189 E3 ubiquitin ligase-substrate pairs from known databases [3,40,41]. To construct negative controls, I randomly paired kinases and E3 ubiquitin ligases with different partners with no evidence of existence of functional interaction in these pairs. Randomly paired partners might later turn out to be real interaction partners, but most likely they would not be interaction partners. Therefore, as a negative control, 106 kinase- and 124 E3 ubiquitin ligase-random partner pairs were constructed (Fig. 1B, Supplementary Data 1).

Using Phyre2, a server for template-based structure modeling [42], I obtained the full predicted structures of kinases, E3 ubiquitin ligases, and their partners (Supplementary Data 2). Owing to the computational costs, the calculation of every possible protein docking structure is almost impossible. Therefore, I sampled 1,000 possible docking structures per pair using two docking programs (Table 1, Fig. 1b). RosettaDock uses the Monte–Carlo algorithm with a knowledge-based energy function [39], and includes structure refinement of the docking structures [43]. HDOCKlite uses the fast Fourier transform (FFT)

algorithm with a shape-based scoring function and does not include structure refinement [44].

To analyze the interaction energy landscape (Fig. 1a), I calculated the interaction energy with Rosetta energy and the following equation:

$$Interaction\ Energy\ = \{Rosetta\ energy\ of\ simulated\ docking\ structure\} - \sum_{Each\ component\ in\ pair} \{Rosetta\ energy\ of\ single\ components\}$$

As every theoretical protein energy exhibits differences compared with the experimental data, I ignored structural change-related energy and considered only the affinity to minimize error. To indicate interacting positions, I used Interface Root Mean Square deviation (iRMS) [45]. Originally, iRMS was developed to compare simulated docking and native structures [46]. In this study, I slightly changed iRMS. Instead, of the native structure, I set the strongest interacting structure as a reference structure (Fig. 1c). The redefined iRMS was as follows:

$$iRMS\ = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \parallel x_i - y_i \parallel^2}$$

N: the number of backbone atom of interface residues

x_i: the position of i th backbone atom of interface residues in docking structure

y_i: the position of i th backbone atom of interface residues in the strongest interacting structure

As Fig. 1c shows, in the strongest interacting structure, residues closely located within 10 Å were set as interface residues. In different docking structures, the same residues were in different positions. iRMS is the root mean square deviation of the backbone atoms of interface residues in two structures. For every docking structure generated from each enzyme-interaction and random partner pairs using two docking programs, interaction energy and iRMS were analyzed and plotted (Supplementary Data 3 and Supplementary Figs 1–4: generated using RosettaDock; Supplementary Data 4 and Supplementary Figs 5–8: generated using HDOCKlite). As HDOCKlite does not include structure refinements,

the interaction energy calculation indicates that most docking structures with HDOCKlite were unfavorable (Interaction energy > 0 kcal/mol). However, affinity comparison would be possible by comparing the interaction energies.

Narrow funnel-like interaction energy distribution of kinases/E3 ubiquitin ligases and their specific interaction partners

To illustrate the structural meaning of the narrow funnel-like interaction energy distribution, Fig. 2 shows the interactions between cyclin-dependent kinase 4 (CDK4) and its specific/random partners. It is well known that CDK4 interacts with G1/S-specific cyclin-D1 (CCND1) and mediates cell cycle progression (G1-to-S phase, [47]. Inhibitor of nuclear factor kappa-B kinase subunit alpha (IKKα) is part of the IKK complex and phosphorylates inhibitor of nuclear factor kappa-B [48]. Then, the activated nuclear factor kappa-B mediates immune response, inflammation, and apoptosis [49]. CDK4 randomly paired with IKKα, and there are no literature reports or datasets that identify them as interaction partners. Using RosettaDock, 1,000 docking structures of CDK4-CCND1 and CDK4-IKKα pairs were established. Then, by calculating the interaction energy, the top five strongest docking structures have been summarized in Fig. 2. The CDK4-CCND1 pair, an identified interaction pair, interacts in similar positions. However, the CDK4-IKKα pair, a random pair, shows scattered interacting positions on the surface of CDK4 and IKKα. Although the average interaction energies are similar and a little higher in the CDK4-IKKα pair, the average iRMSs are significantly different.

For the quantitative analysis of the interaction energy distribution, I devised a strategy to distinguish narrow funnel-like interaction energy distribution (Fig. 3a). Docking structures with some weak interactions are regarded as functionally noninteracting structures. Docking structures with strong interactions but large iRMSs indicate broad funnel-like interaction energy distribution. Therefore, I only had to consider docking structure ratios with strong interactions and small iRMSs. I defined a score (narrow funnel distribution) for the distinction of narrow funnel-like distribution (Fig. 3a). If the narrow funnel distribution was bigger than or equal to certain distribution criteria, this interaction energy distribution was determined as exhibiting narrow funnel-like distribution. Then,

only three criteria (Interaction energy, iRMS, and distribution criteria) had to be decided (Fig. 3a).

As per my hypothesis, if a protein pair with specific interacting partners showed narrow funnel-like interaction energy distribution, specific partners could be predicted with interaction energy distribution analysis by taking specific criteria. Therefore, I calculated the difference between the true-positive (identified pair with narrow funnel-like interaction energy distribution) and false-positive (random pair with narrow funnel-like interaction energy distribution) ratios for each criterion. As per the hypothesis, the narrow funnel distribution of the identified pairs was higher than that of random pairs in the case of most criteria.

For whole protein pairs, docking structures generated by RosettaDock, HDOCKlite showed up to 11.9%–20.3% difference of true- and false positives (Supplementary Fig. 9). Interaction energy, iRMS, and distribution criteria at the maximum points are described in Supplementary Table 1. As the interaction calculation depends on structure accuracy, I filtered out low-accuracy-predicted structures with less than 70% region with high confidence (>90% confidence) [42] and analyzed the remaining pairs with high-accuracy structure prediction (identified kinase-interaction pair: n = 56 for RosettaDock, n = 58 for HDOCKlite; random kinase pair: n = 52 for RosettaDock, n = 54 for HDOCKlite; identified E3 ubiquitin ligase pair: n = 59 for RosettaDock, n = 70 for HDOCKlite; random E3 ubiquitin ligase pair: n = 62 for RosettaDock and HDOCKlite). The docking structures, generated by RosettaDock and HDOCKlite, showed up to 17.9%–26.4% difference of true- and false positives (Supplementary Fig. 10). Interaction energy, iRMS, and distribution criteria at the maximum point are described in Supplementary Table 2.

As protein global docking works well for small proteins[50], I again filtered out pairs with large proteins (>700 residues) and analyzed the remaining pairs with high-accuracy structure prediction and relatively small proteins (identified kinase-interaction pair: n = 28 for RosettaDock and HDOCKlite; random kinase pair: n = 26 for RosettaDock and HDOCKlite; identified E3 ubiquitin ligase pair: n = 25 for RosettaDock, n = 36 for HDOCKlite; random E3 ubiquitin ligase pair: n = 28 for RosettaDock and HDOCKlite). For the kinase pairs, the docking structures generated by RosettaDock and HDOCKlite

showed up to 44.2% and 48.9% difference of true- and false positives, respectively. For the E3 ubiquitin ligase pairs, the docking structures generated by RosettaDock and HDOCKlite showed up to 35.9% and 27% difference of true- and false positives, respectively (Fig. 3b). Interaction energy, iRMS, and distribution criteria at the maximum points are described in Table 2. In order to investigate the effectiveness of finding specific interaction partners using interaction energy distribution, true- and false-positive rates were compared with those of the yeast two-hybrid screening approach. True- and false-positive rates from genome-scale yeast two-hybrid trials in three species were compared [51]. Whole pairs and pairs with high-accuracy structure predictions had similar accuracy to those of the yeast two-hybrid screening (Supplementary Figs 9 and 10). Pairs with high-accuracy structure predictions and relatively small proteins showed significantly better results than those of the yeast two-hybrid screening approach (Fig. 3c).

Interaction energy distribution of other protein complexes

As kinase and E3 ubiquitin ligase PPIs represent only a small part of the whole PPIs, I investigated the interaction energy distribution in other protein complexes. I retrieved 183 curated experimentally determined protein complex structures from the IntAct database [4], then analyzed the PPI between the interacting chains with the interface area in the protein structure using RosettaDock (Fig. 1b and Supplementary Data 5). The interaction energy and iRMS distribution of the protein complexes are plotted in Supplementary Fig. 11. I retrieved 110 experimentally determined protein structures, which were unlikely to be engaged in direct interactions, from the Negatome 2.0 database as negative controls [52], and analyzed the PPI using RosettaDock (Fig. 1b and Supplementary Data 5). The interaction energy and iRMS distribution of the protein complexes are plotted in Supplementary Fig. 12.

To confirm whether other protein complexes showed narrow funnel-like interaction energy distribution, I calculated the iRMS averages of possible docking structures. As per the hypothesis (Fig. 1a), the average iRMS of the interacting protein complexes were significantly smaller than the average iRMS of non-interacting proteins (95% confidence using Student's *t*-test, Fig. 4). It means interaction energy distribution of interacting protein

complexes were much narrower than those of non-interacting proteins. Furthermore, I functionally annotated these proteins according to level-2 gene ontology distribution from Generic GO slim [53], then divided the protein pairs according to the related biological processes and calculated the average iRMS for each part (Fig. 4). Almost every division showed smaller average iRMS of the interacting complexes than those of the noninteracting pairs, except for the negative regulation of the biological processes. In particular, protein pairs related to cellular, metabolic, and developmental processes, signaling, response to stimuli, and biological process regulation showed significantly smaller average iRMSs than noninteracting protein pairs (95% confidence using Student's *t*-test). These processes are well known to display molecular cascades with specific PPIs. Among those, cellular process-, developmental process-, and response to stimuli-related proteins pairs showed higher differences (99% confidence using Student's *t*-test). In particular, developmental process-related proteins pairs showed the biggest differences of iRMSs between the interacting complexes and noninteracting pairs (median of average iRMS of the interacting complexes: 22.37 Å, median of average iRMS of noninteracting pairs: 29.6 Å). This result might be related to developmental processes controlled by complex molecular cascades and networks with specific PPIs [54].

Discussion

Currently, most protein interaction prediction methods are biological evidence- and identified protein complex structure template-based [5-14,17,18]. However, these methods that require omics scale interspecies biological data or identified similar interactions display limited predictability. Knowledge-free protein interaction prediction methods would be necessary to describe the interactome in various species. Therefore, I focused on a protein docking program, which simulates the physical interactions of proteins. Protein docking programs have been developed for over 30 years [55]. With communitywide docking program assessment, CAPRI (Critical Assessment of PRediction of Interactions) [56], recent docking programs are showing good performance in

predicting protein complex structures [57]. However, docking programs are rarely used to find interacting partners for two reasons: low accuracy in weak and transient interactions and computational costs [5,19,55].

In 2011, using supercomputers, the protein docking structures of interacting and nonbinding protein pairs were generated for up to 100,000 structures per pair and were compared [19]. Although >50% of the protein complexes had better docking scores than 85% of the background, the approach showed lower performance for enzyme-inhibitor interactions. Most protease interactions had values similar to those of the background [19]. Three reasons explain why docking score, which comprises the interaction energy and the shape complementarity of the docking structure, was not adequate for high-accuracy prediction of interacting partners. First, investigation of proteomic quantity is missing. Even a interaction partner with weak interaction energy can be a dominant interaction partner if is present in the cell in large amounts [23]. However, the cellular quantities of most proteins are unknown. Second, every protein energy function has a small difference from reality. Although multiple protein energy functions have been developed, energy-based free protein structure modeling showed poorer performance than template-based modeling [58]. The Rosetta energy function, one of the successful and widely used energy functions, predicts that the $\Delta\Delta G$ of the HIV1 protease T193V mutation is −4.95 kcal/mol, but the experimental result was −1.11 kcal/mol [39,59]. In contrast, owing to the competition effect, slightly higher interaction energy is sufficient to be a dominant interaction partner [35]. Therefore, small errors in the energy function can hinder the prediction of interaction partners. Third, the low shape complementarity of transient interactions may not be distinguishable from those of others [5,22]. Certain docking scores use protein complex geometric shape complementarity [44,60], but transient interaction partners do not have dominant shape complementarity that is distinguishable from those of others.

In this study, I used specificity, as well as interaction energy, to distinguish protein interaction partners. Specificity is the result of a long evolution [61]. Because of specificity, complicated molecular cascades in biological processes and complex life activities can be maintained [35]. As specificity in multiple PPIs depends on the interaction in specific regions, such as Vander Waals and electrostatic interactions [35,61], such specific

PPIs are in a specific orientation. For example, p53, a well-known tumor suppressor, has more than 20 and 10 phosphorylation and ubiquitination sites, respectively. Each site is phosphorylated and ubiquitinated by specific kinases and E3 ubiquitin ligases [62]. These PPIs in specific orientation result in narrow funnel-like interaction energy distribution (Fig. 1a). In this study, I showed that specific interaction partner search with narrow funnel-like interaction energy distribution achieved similar accuracy as yeast two-hybrid screening (Supplementary Figs 9b and 10b). For protein pairs with high structure prediction rates and relatively small proteins, this method showed far better accuracy than yeast two-hybrid screening (Fig. 3c). Therefore, the narrow funnel-like interaction energy distribution is a key indicator of specific interaction partners.

High computational cost is another hurdle to using docking programs for interaction partner predictions. In this study, one of the surprising results was the performance of HDOCKlite, an FFT-based rigid body docking program [44], in specific interaction partner distinction (Fig. 3c). As this algorithm runs fast even on personal computers (Table 1), it can predict specific interaction partners without additional computational power. Moreover, protein interaction prediction on the proteomic level might be feasible with small computational power. RosettaDock includes docking structure refinement using the Monte–Carlo algorithm [43]. Therefore, it can calculate cooperative interactions, accompanied by structure changes [63]. Furthermore, I showed that 1,000 docking structures per each pair are enough for interaction partner prediction. Therefore, the narrow funnel-like interaction energy distribution-based approach adapted to small computational power. Due to the small computation amount, interaction predictions on the proteome scale are also possible.

However, this study also has clear limitations. First, protein structures should be predicted or determined and the prediction result affects prediction results (Fig. 3, Supplementary Figs 9 and 10). However, amazing progress has been recently achieved in protein structure prediction. Protein structure prediction tools using deep learning were developed [64,65]. In a recent blind protein structure prediction assessment, CASP14 (Critical Assessment of protein Structure Prediction 14) [58], AlphaFold predicted certain targets with higher accuracy than template-based predictions [66]. Protein structure

prediction error is decreasing close to the experimental error [65,66]. Moreover, they are accumulating predicted structures in databases and publishing such structures [67]. Therefore, anyone can use proteome scale-predicted protein structures in the near future. As I used template-based structure prediction, this study is not fully template-free. However, using deep learning prediction-based protein structures, knowledge-free protein interaction predictions could be achieved.

Second, protein size affects the prediction results (Fig. 3, Supplementary Figs 9 and 10). As global docking works better for small than large proteins [50], interaction energy distribution is affected by protein size. This bias can be improved through high resolution with more docking structures [19]. Furthermore, interaction partner prediction with interaction energy distribution can be used in parallel with other interaction partner predictions. Widely used biological evidence- and template structure-based protein interaction prediction methods are biased by well-studied interaction pairs [5]. However, the interaction energy distribution is completely irrelevant for the extent of how well-studied these interactions are. This method is biased by protein pairs with well-predicted structures and of small size (Fig. 3, Supplementary Figs 9 and 10). They can be used together to complement each other.

Two other studies could contribute to the further development of these aspects. One focuses on protein energy function improvement. There are two kinds of protein energy functions. One is the physics-based energy function, constructed based on classical mechanics and correction terms with perturbations in quantum mechanics, such as the CHARMM force field [68]. The other is knowledge-based energy function, using Boltzmann relation to retrieve energy terms from experimentally determined protein structures, such as Rosetta energy [39]. These energy functions are developed using a bottom-up approach consisting of each formula and parameter of physical properties like building blocks [39,68]. Terms might also be missing or contain big errors. Therefore, from a holistic point of view, protein energy might be retrieved using the Boltzmann relation from protein structures. To do so, significant amounts of structures are necessary, although the experimentally identified protein structures are limited. However, due to the recent progress in the protein structure prediction-related deep learning field, we can

generate unlimited protein structures. In AlphaFold, predicted distance and torsion distribution are generated by deep learning and the potentials are retrieved with predicted distributions using Boltzmann relation without specific physical energy formula [64]. Therefore, interaction energy can be generated with a top–down approach using deep learning. Moreover, the Boltzmann distribution is valid in the equilibrium state [69]. However, living organisms are dynamic, open, and are in a nonequilibrium state. To describe life phenomena more precisely, nonequilibrium statistical mechanics, such as generalized Boltzmann distribution, would be necessary [70,71].

The other further study is the multifactor analysis of the interaction energy distribution using deep learning. In this study, protein interaction energy distributions were different for each protein family (Fig. 4). Moreover, certain protein complexes display multifunnel-like distribution. Not all specific protein complex properties can be considered narrow funnel analysis-based in this study. Therefore, the deep learning-based multifactor analysis would be helpful to consider these properties. Interaction prediction using interaction energy distribution can be simplified to binary classification problems, which have been already successfully solved in multiple cases using deep learning [72]. Training data are interaction energy distribution and interaction results (interaction pair or not). Using deep learning algorithms for binary classification, the program would decide whether the given interaction energy distribution was from interacting pairs or not. In this study, I tried to use deep learning to predict interaction partners with interaction energy data but failed to achieve successful predictions (Optimizer: RMSprop [73] and adam [74], Loss function: binary cross-entropy [75], Epochs:1~300, Hidden layers:1~10, K-fold validation results: 0.45~0.6). This might be caused by the lack of data since I only had several hundreds of interaction energy distributions. Proteomic-scale interaction energy distribution analysis might thus be necessary to apply deep learning.

Another possible interaction energy distribution analysis application is that narrow funnel-like interaction energy distribution can be a specificity indicator. Finding specific interactions is important in drug discovery. Nonspecific interactions can result in side effects [76]. Although narrow funnel-like interaction energy distribution does not guarantee the exclusion of binding other molecules, it shows that proteins bind to a

specific site and can be used as a specific interaction indicator. In particular, several kinases and E3 ubiquitin ligases are considered as targets of multiple drugs, such as anticancer drugs [77,78]. Therefore, this study could be applied in drug discovery as a specific interaction indicator.

Materials and Methods

Protein structure preparation

To analyze kinase interactions, 135 kinase-interaction pairs were retrieved from kinase and protein interaction databases [3,40]. Using NumPy's random number generator [79], 106 random, nonidentified interacting pairs were generated (Supplementary Data 1). From the E3 ubiquitin ligase-substrate interaction database UbiBrowser [29], 189 E3 ubiquitin ligase-substrate interactions were retrieved. Using NumPy's random number generator [79], 124 random, nonidentified interacting pairs were generated (Supplementary Data 1). Using the Phyre2 server [42], protein structures were predicted using the intensive mode option (Supplementary Data 2).

To analyze interacting protein complexes, 183 protein complex structures were retrieved from the protein interactome database IntAct [4]. Interacting chains in the protein complexes were analyzed using the interfaceResidue script of PyMOL [80]. As negative controls, 110 experimentally determined protein structures, unlikely engaged in direct interactions, were retrieved from the Negatome 2.0 database [52].

Protein docking structure generation using RosettaDock

For RosettaDock, I used the Rosetta 2020.08 bundle. Each kinase and E3 ubiquitin ligase protein interaction pairs were merged using PyMOL [80]. Kinase or E3 ubiquitin ligase chain IDs were designated as "A" and those of interacting partners as "B" using PyMOL [80]. Before kinase and E3 ubiquitin ligase protein pair docking using RosettaDock [43], I optimized their side-chain conformations (prepacking) using the following command:

./bin/docking_protocol.static.linuxgccrelease -in:file:s (Merged protein structure file) -docking:partners A_B -dock_pert 3 8 -randomize1 -randomize2 -spin -out:path:all (Output directory)

Next, I generated 1,000 docking structures using the following command: ./bin/docking_protocol.static.linuxgccrelease -in:file:s (prepacked files) -docking:partners A_B -dock_pert 3 8 -randomize1 -randomize2 -spin -use_ellipsoidal_randomization true -nstruct 1000 -out:path:all (output directory)

To analyze interacting proteins complexes and noninteracting protein pairs, I optimized their side-chain conformations (prepacking) using the following command: ./bin/docking_protocol.static.linuxgccrelease -in:file:s (protein complex structure file) -docking:partners (chain ID of proteins which participating in interaction)_(chain ID of proteins in opposite side of interaction) -dock_pert 3 8 -randomize1 -randomize2 -spin -out:path:all (output directory)
Then, I generated 1,000 docking structures using the following command: ./bin/docking_protocol.static.linuxgccrelease -in:file:s (protein complex structure file) -docking:partners (chain ID of proteins which participating in interaction)_(chain ID of proteins in opposite side of interaction) -dock_pert 3 8 -randomize1 -randomize2 -spin -use_ellipsoidal_randomization true -nstruct 1000 -out:path:all (output directory)

To analyze kinase and E3 ubiquitin ligase interactions using HDOCKlite, I generated 1,000 docking structures using the following command:
./hdock (kinase or E3 ubiquitin ligase) (Interaction partner) -out (docking file)
./createpl (docking file) top1000.pdb -nmax 1000 -complex -models

All commands were written as a Linux shell file and executed. For further analysis, kinase or E3 ubiquitin ligase chain IDs were designated as “A” and those of interacting partners as “B” using PyMOL [80]. The graphical representations of the docking structure were exported using PyMOL [80].

Interaction energy and iRMS analysis

To calculate interaction energy, I calculated the Rosetta energy of each protein or complex using the following command:
./bin/score_jd2.linuxgccrelease -in:file:s (protein or complex) -out:file:scorefile (score file)

Then, I calculated the interaction energy using the following equation:

$$Interaction\ Energy\ = \{Rosetta\ energy\ of\ simulated\ docking\ structure\} - \sum_{Eac\quad component\ in\ pair} \{Rosetta\ energy\ of\ single\ components\}$$

Then, iRMS were calculated using DockQ [45] and the following command:
./DockQ.py (Docking structure) (Strongest interacting docking structure) -short >>(iRMS file)

For interacting protein complexes and noninteracting protein pairs, iRMS were calculated using the following command:
./DockQ.py (Docking structure) (Strongest interacting docking structure) -short -native_chain1 (chain ID of proteins which participating in interaction) -model_chain1 (chain ID of proteins which participating in interaction) -native_chain2 (chain ID of proteins in opposite side of interaction) -model_chain2 (chain ID of proteins in opposite side of interaction) -perm1 -perm2 >>(iRMS file)

All commands were written in a Linux shell file and executed.

Gene ontology analysis

The bioinformatic analysis was performed as described previously [81], with small modifications. Briefly, using OmicsBox v2.0.36 [82], protein sequences were BLASTed against the NCBI nr database [83]. Functional domains and motifs were identified using the

EMBL-EBI InterPro database [84]. Then, OmicsBox was used to annotate these results. Finally, I summarized gene ontology results according to the level-2 gene ontology distribution from Generic GO slim [53] using OmicsBox.

Acknowledgments

I appreciate Dr. Stefano Scopel at Sogang University for his advice on using deep learning as well as Dr. Doseok Kim at Sogang University, Soft-matter Optical Spectroscopy Laboratory in Sogang University, and Mr. Eunho Song at Seoul National University for their advice on data representation and identification of the missing explanation. Furthermore, I would like to thank Enago (www.enago.kr) for the English language review.

Funding

This study was conducted independently without any funding source.

References

1. Rao, V.S., Srinivas, K., Sujini, G. & Kumar, G. Protein-protein interaction detection: methods and analysis. International journal of proteomics 2014(2014).

2. Bateman, A. et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research (2021).

3. Orchard, S. et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nature methods 9, 345-350 (2012).

4. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic acids research 42, D358-D363 (2014).

5. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting protein–protein interactions from the molecular to the proteome level. Chemical reviews 116, 4884-4909 (2016).

6. Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research 49, D605-D612 (2021).

7. Reid, A.J., Ranea, J.A., Clegg, A.B. & Orengo, C.A. CODA: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion. PloS one 5, e10908 (2010).

8. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86-90 (1999).

9. Saha, S., Chatterjee, P., Basu, S., Kundu, M. & Nasipuri, M. FunPred-1: Protein function prediction from a protein interaction network using neighborhood analysis. Cellular and Molecular Biology Letters 19, 675-691 (2014).

10. Matthews, L.R. et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. Genome research 11, 2120-2126 (2001).

11. Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. & Oliva, B. BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference. Nucleic acids research 40, W147-W151 (2012).

12. Bhardwaj, N. & Lu, H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. Bioinformatics 21, 2730-2738 (2005).

13. Yin, C. & Yau, S.S.-T. A coevolution analysis for identifying protein-protein interactions by Fourier transform. PLoS One 12, e0174862 (2017).
14. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. Science 365, 185-189 (2019).
15. Pazos, F., Juan, D., Izarzugaza, J.M., Leon, E. & Valencia, A. Prediction of protein interaction based on similarity of phylogenetic trees. in Functional Proteomics 523-535 (Springer, 2008).
16. Dong, S. et al. Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. Plant physiology 179, 1893-1907 (2019).
17. Zhang, Q.C., Petrey, D., Garzón, J.I., Deng, L. & Honig, B. PrePPI: a structure-informed database of protein–protein interactions. Nucleic acids research 41, D828-D833 (2012).
18. Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. & Gursoy, A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. Nucleic acids research 42, W285-W289 (2014).
19. Wass, M.N., Fuentes, G., Pons, C., Pazos, F. & Valencia, A. Towards the prediction of protein interaction partners using physical docking. Molecular systems biology 7, 469 (2011).
20. Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T. & Akiyama, Y. MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data. Protein and peptide letters 21, 766-778 (2014).
21. Dai, B. & Bailey-Kellogg, C. Protein interaction interface region prediction by geometric deep learning. Bioinformatics (2021).
22. Kuroda, D. & Gray, J.J. Shape complementarity and hydrogen bond preferences in protein–protein interfaces: implications for antibody modeling and protein–protein docking. Bioinformatics 32, 2451-2456 (2016).
23. Hein, M.Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163, 712-723 (2015).

24. Ochsner, S.A. et al. The Signaling Pathways Project, an integrated 'omics knowledgebase for mammalian cellular signaling pathways. Scientific data 6, 1-21 (2019).
25. Wang, H., Chevalier, D., Larue, C., Cho, S.K. & Walker, J.C. The protein phosphatases and protein kinases of Arabidopsis thaliana. The Arabidopsis Book/American Society of Plant Biologists 5(2007).
26. Dardick, C., Chen, J., Richter, T., Ouyang, S. & Ronald, P. The rice kinase database. A phylogenomic database for the rice kinome. Plant physiology 143, 579-586 (2007).
27. Choi, J., Lee, W., An, G. & Kim, S.-R. OsCBE1, a Substrate Receptor of Cullin4-Based E3 Ubiquitin Ligase, Functions as a Regulator of Abiotic Stress Response and Productivity in Rice. International journal of molecular sciences 22, 2487 (2021).
28. Huang, K.-Y., Weng, J.T.-Y., Lee, T.-Y. & Weng, S.-L. A new scheme to discover functional associations and regulatory networks of E3 ubiquitin ligases. in BMC systems biology Vol. 10 27-36 (Springer, 2016).
29. Wang, X. et al. UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase–substrate interactions in eukaryotic species. Nucleic acids research (2021).
30. Li, H., Wang, M. & Xu, X. Prediction of kinase–substrate relations based on heterogeneous networks. Journal of bioinformatics and computational biology 13, 1542003 (2015).
31. Qin, G.-M., Li, R.-Y. & Zhao, X.-M. PhosD: inferring kinase–substrate interactions based on protein domains. Bioinformatics 33, 1197-1204 (2017).
32. Wang, H. et al. HKPocket: human kinase pocket database for drug design. BMC bioinformatics 20, 1-11 (2019).
33. Gan, J. et al. Ksimc: Predicting kinase–substrate interactions based on matrix completion. International journal of molecular sciences 20, 302 (2019).
34. Xue, B., Jordan, B., Rizvi, S. & Naegle, K.M. KinPred: A unified and sustainable approach for harnessing proteome-level human kinase-substrate predictions. PLoS computational biology 17, e1008681 (2021).

35. Ubersax, J.A. & Ferrell Jr, J.E. Mechanisms of specificity in protein phosphorylation. Nature reviews Molecular cell biology 8, 530-541 (2007).
36. Jenardhanan, P., Panneerselvam, M. & Mathur, P.P. Targeting kinase interaction networks: a new paradigm in PPI based design of kinase inhibitors. Current topics in medicinal chemistry 19, 467-485 (2019).
37. Huse, M. & Kuriyan, J. The conformational plasticity of protein kinases. Cell 109, 275-282 (2002).
38. Berndsen, C.E. & Wolberger, C. New insights into ubiquitin E3 ligase mechanism. Nature structural & molecular biology 21, 301-307 (2014).
39. Alford, R.F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation 13, 3031-3048 (2017).
40. Huang, K.-Y. et al. RegPhos 2.0: an updated resource to explore protein kinase–substrate phosphorylation networks in mammals. Database 2014(2014).
41. Li, Y. et al. An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. Nature communications 8, 1-9 (2017).
42. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols 10, 845-858 (2015).
43. Chaudhury, S. et al. Benchmarking and analysis of protein docking performance in Rosetta v3. 2. PloS one 6, e22477 (2011).
44. Yan, Y., Tao, H., He, J. & Huang, S.-Y. The HDOCK server for integrated protein–protein docking. Nature protocols 15, 1829-1852 (2020).
45. Basu, S. & Wallner, B. DockQ: a quality measure for protein-protein docking models. PloS one 11, e0161879 (2016).
46. Lensink, M.F., Méndez, R. & Wodak, S.J. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins: Structure, Function, and Bioinformatics 69, 704-718 (2007).
47. Baker, S.J. & Reddy, E.P. CDK4: a key player in the cell cycle, development, and cancer. Genes & cancer 3, 658-669 (2012).

48. Mercurio, F. et al. IKK-1 and IKK-2: cytokine-activated IκB kinases essential for NF-κB activation. Science 278, 860-866 (1997).
49. Mitchell, S., Vargas, J. & Hoffmann, A. Signaling via the NFκB system. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 8, 227-241 (2016).
50. Daily, M.D., Masica, D., Sivasubramanian, A., Somarouthu, S. & Gray, J.J. CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock. PROTEINS: Structure, Function, and Bioinformatics 60, 181-186 (2005).
51. Huang, H., Jedynak, B.M. & Bader, J.S. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. PLoS computational biology 3, e214 (2007).
52. Blohm, P. et al. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic acids research 42, D396-D400 (2014).
53. Consortium, G.O. The Gene Ontology (GO) database and informatics resource. Nucleic acids research 32, D258-D261 (2004).
54. Davidson, E.H. et al. A genomic regulatory network for development. science 295, 1669-1678 (2002).
55. Vakser, I.A. Protein-protein docking: From interaction to interactome. Biophysical journal 107, 1785-1793 (2014).
56. Janin, J. et al. CAPRI: a critical assessment of predicted interactions. Proteins: Structure, Function, and Bioinformatics 52, 2-9 (2003).
57. Lensink, M.F. et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. Proteins: Structure, Function, and Bioinformatics 89, 1800-1823 (2021).
58. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Proteins: Structure, Function, and Bioinformatics 89, 1607-1617 (2021).
59. Altman, M.D., Nalivaika, E.A., Prabu-Jeyabalan, M., Schiffer, C.A. & Tidor, B. Computational design and experimental study of tighter binding peptides to an

inactivated mutant of HIV-1 protease. Proteins: Structure, Function, and Bioinformatics 70, 678-694 (2008).
60. Pierce, B.G. et al. ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics 30, 1771-1773 (2014).
61. Siddiq, M.A., Hochberg, G.K. & Thornton, J.W. Evolution of protein specificity: insights from ancestral protein reconstruction. Current opinion in structural biology 47, 113-122 (2017).
62. Boehme, K.A. & Blattner, C. Regulation of p53 Activity. Current Chemical Biology 4, 1-12 (2010).
63. Marze, N.A., Roy Burman, S.S., Sheffler, W. & Gray, J.J. Efficient flexible backbone protein–protein docking for challenging targets. Bioinformatics 34, 3461-3469 (2018).
64. Senior, A.W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710 (2020).
65. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871-876 (2021).
66. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021).
67. Varadi, M. et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research 50, D439-D444 (2022).
68. Vanommeslaeghe, K. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. Journal of computational chemistry 31, 671-690 (2010).
69. McDowell, S.A. A simple derivation of the Boltzmann distribution. Journal of chemical education 76, 1393 (1999).
70. Shear, D.B. The generalized Boltzmann distribution. Journal of theoretical biology 39, 165-169 (1973).
71. Lin, M.M. Generalized Boltzmann distribution for systems out of equilibrium. arXiv preprint arXiv:1610.02612 (2016).

72. Kim, K., Lee, B. & Kim, J.W. Feasibility of Deep Learning Algorithms for Binary Classification Problems. Journal of intelligence and information systems 23, 95-108 (2017).
73. Kurbiel, T. & Khaleghian, S. Training of deep neural networks based on distance measures using RMSProp. arXiv preprint arXiv:1708.01911 (2017).
74. Zhang, Z. Improved adam optimizer for deep neural networks. in 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) 1-2 (IEEE, 2018).
75. Ruby, U. & Yendapalli, V. Binary cross entropy with deep learning technique for image classification. International Journal of Advanced Trends in Computer Science and Engineering 9(2020).
76. Lounkine, E. et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature 486, 361-367 (2012).
77. Jia, L. & Sun, Y. SCF E3 ubiquitin ligases as anticancer targets. Current cancer drug targets 11, 347-356 (2011).
78. Attwood, M.M., Fabbro, D., Sokolov, A.V., Knapp, S. & Schiöth, H.B. Trends in kinase drug discovery: Targets, indications and inhibitor design. Nature Reviews Drug Discovery 20, 839-861 (2021).
79. Charbonneau, P. C. RANDOM NUMBERS AND WALKS. in Natural Complexity 321-337 (Princeton University Press, 2017).
80. DeLano, W.L. Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography 40, 82-92 (2002).
81. Choi, J., Shin, J.-H., An, H.J., Oh, M.J. & Kim, S.-R. Analysis of secretome and N-glycosylation of Chlorella species. Algal Research 59, 102466 (2021).
82. Gutierrez, J. Metagenomic analysis of two soda lakes, with and without cyanobacterial bloom, with OmicsBox.
83. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 35, D61-D65 (2007).

84. Mitchell, A. et al. The InterPro protein families database: the classification resource after 15 years. Nucleic acids research 43, D213-D221 (2015).

Tables

Table 1. Docking programs for possible docking structure generation.

	RosettaDOCK	HDOCKlite
Algorithm	Monte-Carlo	Fast-Fourier Transform
Scoring	Knowledge-based Energy Function	Shape-based Scoring Function
Structure Refinement	O	X (rigid body)
Speed	Slow*	Fast**
Reference	(43)	(44)

*Several hours~several days per interaction pair in personal computer (Intel® Core i7-10700K 3.80GHz with 8 cores and 32GB of RAM)
**Several minutes~few hours per interaction pair in personal computer (Intel® Core i7-10700K 3.80GHz with 8 cores and 32GB of RAM)

Table 2. Criteria to distinguish narrow funnel-like interaction energy distribution of interaction pairs with high-accuracy structure prediction results and relatively small proteins.

	RosettaDock		HDOCKlite	
	Kinase pair	E3 ubiquitin ligase pair	Kinase pair	E3 ubiquitin ligase pair
Interaction energy criteria (kcal/mol)	-3.6	-23.4	26~400	150~151
iRMS criteria (Å)	44.6	39.6~40	44.2	29.2~29.6
Distribution criteria (%)	92.2	0.8	0.2	0.8

Figures and legends

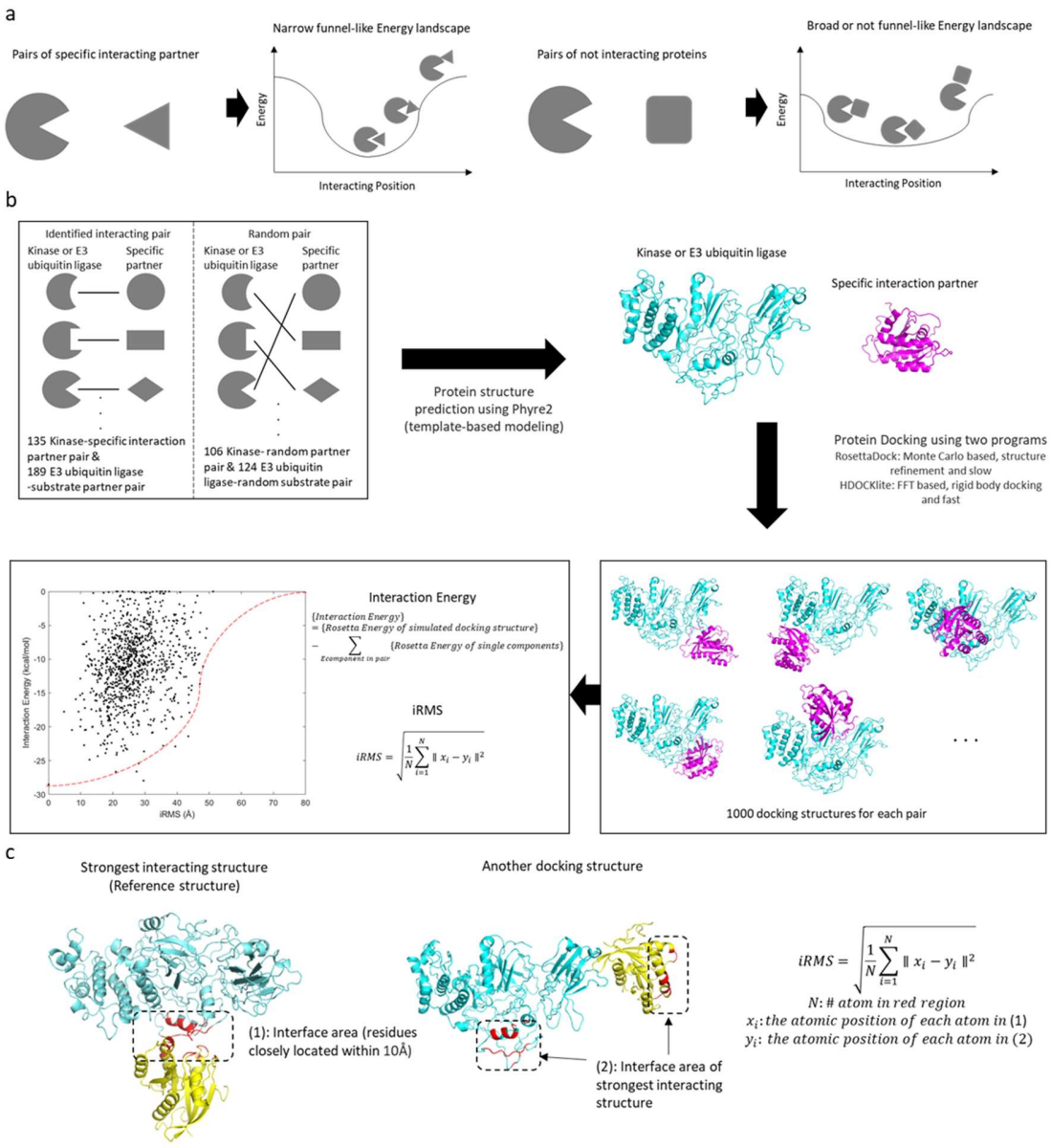

Fig. 1: Narrow funnel-like interaction energy distribution hypothesis and workflow for interaction energy distribution analysis.

a. Schematic diagram of narrow funnel hypothesis. Specific interactions might have narrow funnel-like interaction energy distribution. **b.** Schematic diagram of interaction energy distribution analysis workflow. **c.** Introduction of redefined iRMS. Residues of the interface area in the strongest interacting structure are colored in red.

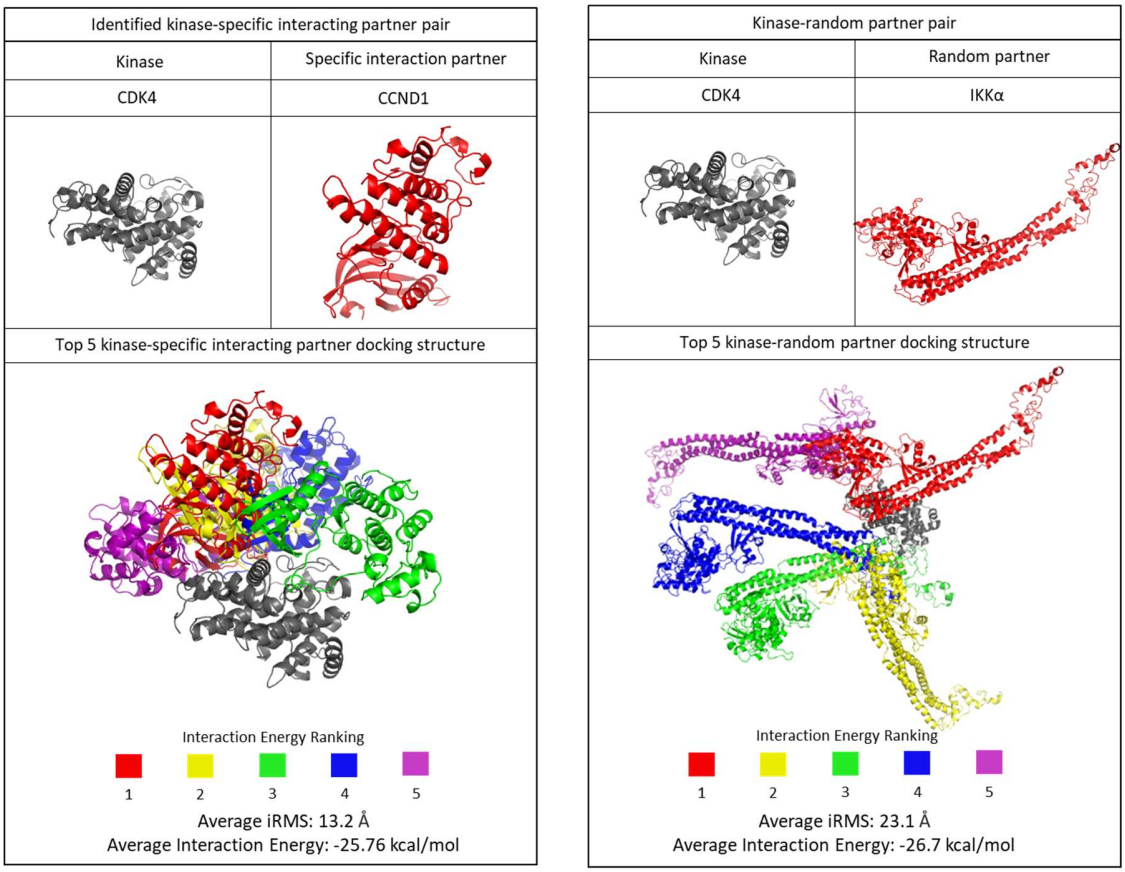

Fig. 2: Structural meaning of narrow funnel-like interaction energy distributions. The five strongest interacting structures of CDK4-CCND1 (identified specific interaction pair) and CDK4-IKKα (random pair). To avoid confusion, CDK4 in each pair is set in the same position and CDK4 in the strongest interacting structure is represented in gray. Each partner is colored in rainbow color in the order of interaction energy.

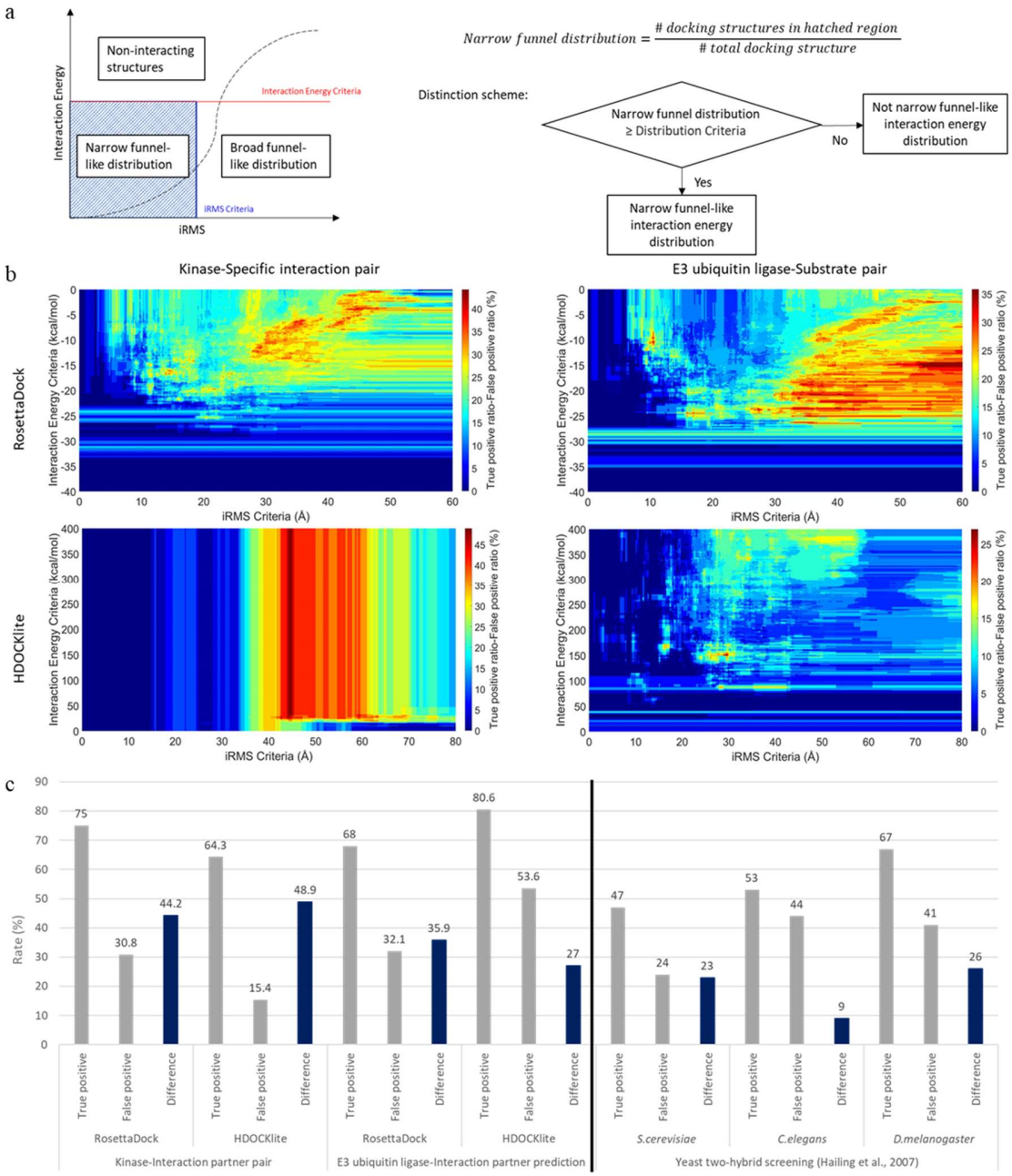

Fig. 3: Quantitative analysis of a narrow funnel-like interaction energy distribution and accuracy of the prediction of protein interactions by searching a narrow funnel-like interaction distribution

a. Scheme of the strategy to distinguish narrow funnel-like interaction energy distributions. **b.** How well narrow funnel interaction energy distribution can distinguish

specific interaction partners for each criteria. Among whole interaction pairs, pairs with high-accuracy prediction results (>70% region with high confidence) and relatively small proteins (≤700 residues) were analyzed. Difference ratios between true-positive (identified pair with narrow funnel-like interaction energy distribution) and false-positive (random pair with narrow funnel-like interaction energy distribution) for each criteria. Due to graphical representation limitations, only the maximum ratios are shown for each distribution criteria. **c.** Rate comparisons for finding specific interaction partners using the narrow funnel and yeast two-hybrid screening methods. Comparison of genome-scale yeast two-hybrid analysis for three species. The criteria for finding narrow funnel-like energy distribution are described in Table 2.

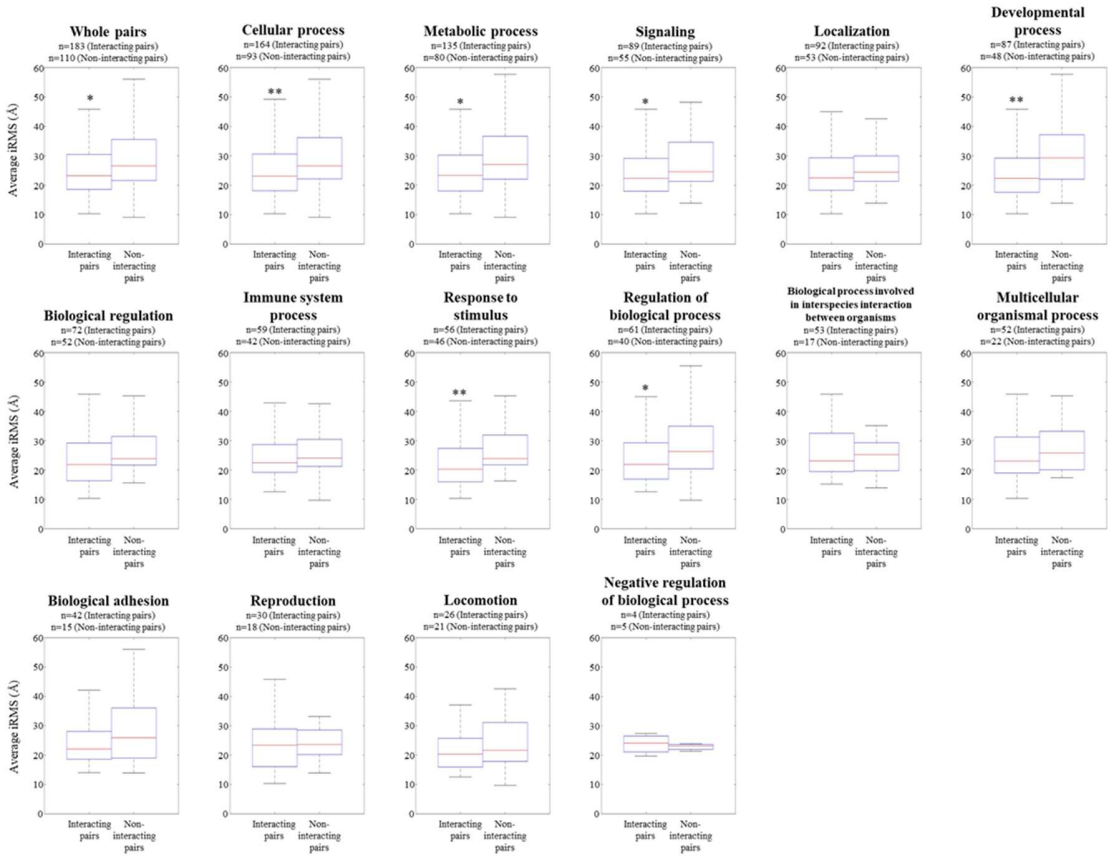

Fig. 4: Distribution of average iRMS of interacting protein complexes and noninteracting protein pairs.

Distribution of average iRMS of interacting protein complexes and noninteracting protein pairs are shown using box-and-whisker plot. Protein pairs divided according to the functional annotation with level-2 gene ontology distribution from Generic GO slim. Each related biological process in level-2 gene ontology from Generic GO slim and the number of protein pairs are notated. Significant differences of average iRMS with 95% confidence using Student's *t*-test are marked with asterisks and 99% confidence marked with double asterisks.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure.pdf
- SupplementaryTable.docx
- SupplementaryData1.xlsx
- SupplementaryData2.xlsx
- SupplementaryData3.xlsx
- SupplementaryData4.xlsx
- SupplementaryData5.xlsx
- SupplementaryData6.xlsx