Deciphering Peptide-Protein Interactions via Composition-Based Prediction: A Case Study with Survivin/BIRC5

Atsarina Larasati Anindya[a], Torbjörn Nur Olsson[a], Maja Jensen[a], Maria-Jose Garcia-Bonete[a], Sally P. Wheatley[b], Maria Bokarewa[c,d], Stefano A. Mezzasalma[e,f], and Gergely Katona[a,*]

[a] Department of Chemistry and Molecular Biology, Faculty of Science, University of Gothenburg, Sweden.

[b] School of Life Sciences, Faculty of Medicine & Health Sciences, University of Nottingham, UK. NG7 2UH

[c] Department of Rheumatology and Inflammation Research, Institute of Medicine, University of Gothenburg, Box 480, 40530 Gothenburg, Sweden.

[d] Rheumatology Clinic, Sahlgrenska University Hospital, Gröna stråket 16, 41346 Gothenburg, Sweden

[e] Division of Materials Physics, Laboratory for Optics and Optical Thin Films, Ruđer Bošković Institute, Bijenička cesta 54 – 10000, Zagreb, Croatia

[f] Institute for advanced Neutron and X-ray Science (LINXS), Lund University, IDEON Building: Delta 5, Scheelevägen 19, 223 70 Lund, Sweden

[*]Corresponding author Gergely Katona, Department of Chemistry and Molecular Biology, Faculty of Science, University of Gothenburg, Box 462, 40530 Gothenburg, Sweden, phone +46-709-123309, E-mail: gergely.katona@gu.se

Abstract

In the realm of atomic physics and chemistry, composition emerges as the most powerful means of describing matter. Mendeleev's periodic table and chemical formulas, while not entirely free from ambiguities, provide robust approximations for comprehending the properties of atoms, chemicals, and their collective behaviours, which stem from the dynamic interplay of their constituents.

Our study illustrates that protein-protein interactions follow a similar paradigm, wherein the composition of peptides plays a pivotal role in predicting their interactions with the protein survivin, using an elegantly simple model. An analysis of these predictions within the context of the human proteome not only illuminates the known cellular locations of survivin and its interaction partners, but also introduces novel insights into biological functionality. It becomes evident that an electrostatic- and primary structure-based description falls short in

predictive power, leading us to speculate that protein interactions are orchestrated by the collective dynamics of the functional groups of the peptides involved.

INTRODUCTION

Most biological processes are directed by protein-protein interactions (PPI). These interactions are associated with a network of non-covalent bonds in a way that are not completely understood, and their deficiencies are often related with disease progression, i.e. aggregation and amyloid formation.

Early studies readily account interactions between proteins with opposite charges, and that increasing the solution ionic strength reverses the interaction[1,2]. Short-range forces such as hydrogen bonding and van der Waals' were also acknowledged to contribute[3]. After insulin was successfully sequenced in 1951[4], the growing number of sequenced proteins brought forth a paradigm to treat the numerous possible combinations of protein primary amino acid sequence as the sole key to determine protein structure, function, and, by extension, how it recognize interaction partners, often disregarding the physicochemical properties of the amino acids embodying the biological message.

Each amino acid indeed has unique properties that ultimately contribute to the overall electrostatic and polar/apolar nature of the protein (charge density and distribution, net charge, and hydrophobicity). However, the physical description based on complementary charge of two partners and similarity in hydrophobicity are not regarded to be sufficiently detailed to direct specific interactions.

At the same time, research on identifying protein binding partners and determining binding affinity keeps rising with the help of high-throughput techniques like two hybrid yeast[5,6], phage display[7], bimolecular fluorescence complementation[8], and peptide microarrays[9,10]. Our understanding of these interactions is not happening with the same pace and generally explained by the involvement of complex physical interactions. Black box machine learning runs the risk of merely transforming existing data into yet another (overfit) model representation.

Peptide microarrays, which has its roots in antigen-antibody reaction determination method[11], is an effective tool in studying PPI due to its high-throughput nature. Simultaneous determination of PPI is performed by exposing multiple antibody spots, arranged on a solid surface, to antigens. In peptide microarrays, the target protein is often divided into short peptide segments and orderly arranged on a solid surface, before being exposed fluorescence-labelled ligand proteins or antibody in solution. The detected fluorescence intensity at each peptide spot correlates to the binding strength of that definite peptide to the ligand protein[12], making mapping binding and non-binding regions in a protein possible, as we have previously shown with the mapping of Polycomb Repressor Complex 2 (PRC2) binding regions to survivin[13].

Rapid accumulation of PPI experimental data from peptide microarray becomes a valuable factor to formulate patterns to identify and understand the driving forces behind. A major challenge to derive prediction tools comes from inconsistencies on defining the cut-off distance of non-covalent interactions[14], which are modelled based on amino acid physicochemical and biochemical properties; AAindex is a database example containing various indices representing these properties[15].

Protein association is thought to start with encounter complex formation mediated by long-range electrostatic forces resulting in a favourable (exothermic) enthalpy change[16], which can likely be predicted by statistical coupling analysis[17]. Encounter complex formation implies an entropic loss from reduced conformational, rotational, and translational degrees of freedom before definite hydrogen and ionic bonds are formed, leading to structural rearrangements and eventual desolvation of the interacting surfaces[18-20].

Certain characteristics are connected to protein interfaces: stable interactions have larger buried surface areas and hydrophobic surfaces than transient ones[21]. Stable interactions also have more conserved and tryptophane, tyrosine, and arginine residues, where most of the binding energy is concentrated[22,23]. The interface centre of stable interactions is usually composed of hydrophobic and aromatic residues, while peripheral regions are often crowded with polar and charged residues. Interfaces with stronger affinities tend to have high modularity involving large, complex interfacial sub-regions, suggesting that individual pairwise interacting residues have additive effects[24,25]. Side-chain interactions are also prominent in PPIs, whereas they only contribute to 36% of the internal bonds stabilizing a protein[26]. There is also a significant compositional difference between transient and permanent interactions (i.e. in quaternary structures or complexes)[27].

Despite knowing conserved residues forming PPIs, binding affinity prediction based solely on amino acid sequence have not been unambiguous, as diverse sequences can have similar binding affinities[28]. However, amino acid sequence data remains the most common feature for inferring structure, and by extension, protein activity. Multiple Sequence Alignment combined with structural template-trained model is the current state-of-the-art strategy employed by well-known prediction tools, such as AlphaFold Multimer[29], AlphaPullDown[30], and RoseTTAFold[31]. These tools have limitations on proteins with no solved homologous structure, a common problem with inherently disordered proteins (IDPs) and transient/weak binding contact interfaces.

Methods using feature extraction from sequence alignment is often inspired from natural language processing methods, such as masked language[32], attention network, and long short-term memory (LSTM) models[31,33-35]. The last two are currently developed to counter the problem of long-range communications between residues, which is found in classical recurrent neural network. Cadet et al recently developed a model relying on the Fourier transform of numerical-encoded sequences, creating a unique pattern from residue frequency for each protein [36]. This study used a training dataset of single point mutations of epoxide hydrolase from *Aspergillus niger* and achieved 81% accuracy on validation dataset to predict epistatic interactions[36]. On the other hand, binding site predictions from the joint analysis of solvation potentials, amino acid composition, conservation, electrostatics, and hydrophobicity seems to yield more promising results[37-42].

Chemical reactions in dynamic equilibrium, including PPIs, are assumed to depend on the law of mass action, where the product to reactant ratio is constant for given temperature. The amount of protein complex depends on interacting partner concentrations and their binding affinity. However, PPI dissociation constant rarely surpasses 1, implying higher proportion of protein complexes at lower interacting partner concentrations compared to higher ones. The scarcity of such observations points to a prevalence of attractive interactions among proteins or to an observational bias focusing on attractive interactions.

Considering the large size and complexity of individual proteins, and the diversity among proteins, it is of practical interest to shift the focus from the concentration of polypeptide chains and protein complexes to whether the concentration of simpler entities composing proteins plays a pivotal role in guiding PPIs. Here, we use survivin, a small protein (16.5 kDa) involved in apoptosis, cell division, and epigenetic regulation for transcription[13], as the target protein model and its binding partners to illustrate how protein composition is a predictive factor for biological interactions. We validate our assertion by referring to extensive experimental data available in the literature about survivin interaction partners and the role of survivin in various biological functions. Finally, we discuss the effectiveness of composition in determining protein function versus the collective dynamics of proteins.

RESULTS AND DISCUSSION

PPI classification based on atom type composition and sequence.

This study used a peptide microarray dataset containing fluorescence intensity from survivin interactions to 36 known and potential binding partners[13], such as Polycomb Repressive Complex 2 (PRC2) subunits[43] and Chromosomal Passenger Complex (CPC). The peptides were then represented by counting various chemical groups they contain as features, encompassing the standard main chain atoms (MC), methylene group from glycine Cα atom (CA-Gly), side chains (carboxyl, amide, phenolic hydroxyl, and sulfhydryl groups and imidazole and indole rings) among others. A comprehensive list and explanation on feature development can be found in the Methods section, with cues to appropriate Supplementary Figures and Tables. Briefly, C-Coord1 model as a starting step contains 17 features (Table S1), C-Coord2 with 15 features (Table S2), and C-Coord3 with 13 features (Table S3). C-Pos model characterizes the distance of side chain carbon atoms from the main chain (Table S4), AA-Letter model (Table S5) simply substitutes amino acids with equal weight values, and Chemical Elements model counts hydrogen, carbon, nitrogen, oxygen, and sulfur atoms (Table S6). Electrostatic and hydrophobic properties are represented in Pos/Neg/HyPho (Table S7) and NetChrg/SumHyPho models (Table S8)[44]. We also compared our approach with Fourier Transform patterns from AAindex numerical value representation of amino acids[36].(Seq, Table S9) The real part of the complex term yields a peptide spectrum (FT-Seq). We also used equal weight value representation of amino acids in alphabetical order (from A=10 to W=200) (UniSeq).

Figure 1 presents the performance indicators of various approaches, using a very simple multilayer perceptron with two intermediate layer perceptrons. As previously reported[13], binders are more readily identified compared to non-binders. Models with more limited features exhibit suboptimal performance, with accuracy and F1 scores for binders falling within the range of 65—74 % and 0—62%, respectively. The FT-Seq, FT-UniSeq feature sets, and the simplest physicochemical descriptor (NetChrg/SumHyPho) similarly exhibit underperformance. It is intriguing to observe that the elemental composition provides a more effective description compared to features based on electrostatic and hydrophobicity characteristics of the peptides. Conversely, symbolic AA-Letter description exhibits lower performance when employed in a simple training model. We speculate that AA-Letter lacks a direct alignment with fundamental chemical principles, i.e. it fails to consider the varying numbers of atoms among different amino acids and does not reflect any similarities between them. More sophisticated models may learn these aspects with sufficient data (as demonstrated in the embedding example later), yet the C-Pos, C-Coord1, and C-Coord2 notably align better with intuitive understanding of amino acid features with limited data.

Simplification of features is exhibited by similar performances between 15 feature (C-Coord2) model with 17 (C-Coord1) or 19 (C-Pos) feature ones. However, overall performance declined with reduction to 13 (C-Coord3) features, suggesting the crucial role of differentiating oxygen atoms in carboxyl and carbonyl groups and nitrogen atoms in amino and amide groups in prediction. As C-Pos and C-Coord1 only differ in the labeling of aliphatic carbon atoms, the contrast between the two models implies the significance of aliphatic carbon atoms in biomolecular recognition, an area where current awareness is limited.

The C-Coord2 model demonstrated an average accuracy of 79.8% (Table 1) compared to the FT-Seq with 63.2% accuracy by only predicting non-interacting peptides. Specifically, the F1 scores for detecting non-binders and binders were 85% and 71% for the C-Coord2 model, whereas the FT-Seq approach failed completely to detect binders. The performance indicators show that the model is biased towards non-binding peptides, which might be explained by the slightly imbalanced dataset, comprised of 60% non-binders.

The impact of network complexity on the prediction accuracy and applicability

Next, we investigated how the complexity of the multilayer perceptron impacts prediction performance with different feature sets. To maintain simplicity, we employed a single intermediate layer of perceptrons and explored changing the number of perceptrons, ranging from the default 100 down to as few as 1. We also evaluated the performance without relying on a distinct test set to assess the model susceptibility in making predictions based on insignificant details and noise within the training set (overfitting).

By decreasing model complexity, composition-based models not only increased overall accuracy but also significantly enhanced the more challenging detection of binders, as demonstrated by the F1 Positive score (Figures 2A and 2B). The optimal configuration, with peak accuracy and F1 scores, emerges with a two-perceptron intermediate layer. Remarkably, there is minimal disparity between the performance of the model with and without a test set, indicating a well-balanced accuracy and generalization. Notably, there exists a substantial non-monotonic variation in F1 performance, even with minor alterations in the number of perceptrons, suggesting that specific logical configurations are more advantageous than others when evaluating the connectivity between perceptrons. The subpar performance of the single perceptron model refutes the argument to dismiss amino acid and atom type diversity for precise binder detection.

Sequence-based methods evidently require a more complex network to enhance binder detection, as their F1 Positive score reaches plateaus around 0.20 (Figures 2C and 2D). The widening gap between the performance with and without cross-validation as model complexity increases suggests that the best-performing model may lack sufficient generalizability.

Intuitive interpretation of the peptide preferences of survivin

The optimized network architecture (Figure 3) highlights two distinct pathways for making survivin binding decisions: 1. the presence of carboxyl groups which inhibit perceptron 2 (binding inhibitor) while being moderated by amino and guanidino groups and 2. for amino-rich peptides, which directly promotes activating perceptron 1. Interestingly, carboxyl group paired with phenyl rings as a significant supportive factor effectively inhibits neuron 2, further enhancing the prediction of survivin binding. All atom types contribute significantly to the prediction, although the weighted edges of features below 0.5 or exceed –0.5 are not displayed in Figure 3.

The occurrences of amino acids in peptide sequence are inherently uncorrelated since their limited placements can be approximated with multinomial distribution. Each amino acid comprises various functional groups, some promote survivin binding while others inhibit it, allowing closely related amino acids, i.e. aspartate and glutamate, to be discerned based on their counts of methylene groups.

The contribution of the peptide primary structure

To improve the sequence-based predictions, we employed bidirectional LSTM[45] networks and embedding techniques. The most effective architecture has two LSTM layers, with 90 neurons in the first layer and 70 neurons in the second. The model comprised 243221 trainable parameters, demonstrated an accuracy of 85.7±0.1 % (cross-validated) and 89.7±0.2 % (non-validated).

Upon comparing the best composition-based and sequence-based predictions, we can infer that the composition alone cannot account for approximately 5% of the accuracy observed in the sequence-based method. However, we must still account for position-related artifacts. Overfitting was likely due to the low data to parameter ratio, as shown by the performance gap between the cross-validated and non-validated trainings. Given the Poisson counting errors in fluorescence detection, which can mislabel binders as non-binders, the high performance shown in non-validated training is unrealistic. Test set contamination[46] may also inflate the cross-validation performance, which can force the model to detect exact sequences instead of real physicochemical patterns. However, when using only every 3rd peptide, we unavoidably worsened the data-to-parameter ratio, with the accuracy of 79.2±0.2% (cross-validated) and 89.7±0.2% (non-validated) making the performance interpretation more challenging.

The performance gap between cross-validated and non-validated sets, along with the diminished performance compared to the complete dataset, underscores the challenge of test set contamination and overfitting within the more intricate LSTM architecture. Mitigating contamination increased the network generalizability while reducing prediction accuracy to the level comparable with the composition-based prediction. The contrast between the composition-based and sequence-based networks is more pronounced when we consider the number of trained parameters for optimization, with 40 parameters (C-Pos) in composition-based network and 243221 in sequence-based LSTM model. Following Occam's razor principle, the model which has 6000 times less parameter should be preferred when performance is similar. While larger datasets would enhance the LSTM model performance, obtaining extensive experimental biochemical data often remains a formidable challenge.

We also assessed the impact of amino acid position in the peptide sequence, as shown in Figure 4, where the weights are extracted as a function of position. Despite having only 35 parameters, the accuracy is 79.7±0.0% (cross-validated) and 80.1±0.2% (non-validated), comparable to composition-based methods. Figure 4A shows that tyrosines, aspartates, tryptophanes and arginines contribute the most to the prediction, in agreement with the composition-based model (Figure 3).

The gradual increase in weights from N- to C-terminus (Figure 4B) is likely linked to how the C-terminus of the peptide is tethered to the substrate through a linker, while the N-terminus extends freely into the solvent[47], making it more likely for the N-terminus to have higher weights due to its closeness to survivin. Such a position specific pattern is unlikely to represent a universal mechanism guiding survivin binding, but these inherent experimental biases need to be addressed in the composition-based analysis.

Visualizing the structure of peptide composition and sequence space and its connection to function

To gain a more intuitive insight to the advantage of the composition-based prediction, we visualized the peptide distribution according to the similarities between atom type composition using T-distributed stochastic neighbour embedding (t-SNE) [48], a useful tool to map the multiple predictive features into a more intuitive 2-dimensional figure. Representing peptides based on their composition results in well-separated clusters (Figure S1A) with distinct separation between binders and non-binders, while only two clusters observed using the FT-Seq representation (Figure S1B) with binders and non-binders mostly overlap with each other. Varying the perplexity value between 5.0-50.0 did not have a significant impact on the data point distribution[49]. While one might assume that peptides from a single protein share similar compositions, our clustering analysis challenges this perspective. Peptides from a single protein can be present in more than five recognizable clusters, exemplified by protein EZH2 in Figure S2, suggesting that a protein can contribute to multiple functions, as regions with different compositions are held together covalently in the polypeptide chain. We also discussed potential pitfalls of assay related errors in the accompanying text under 'Statistical and experimental artifacts', the quantitative connection between fluorescence intensity and atom type composition using t-SNE in Figure S3 and the accompanying text under 'Quantitative binding representation and its relation to composition'.

Analysis of a known survivin complex and structurally characterized interaction partners

Using the composition-based model, we investigated the impact of point mutation on the surrounding residues from known survivin binding partners by creating a metric $M_{bind}(n)$, where 'n' represents a residue position along the sequence. Higher $M_{bind}(n)$ corresponds to a higher number of point mutations required to change the residue's survivin binding properties, making it more functionally conserved. The choice of an amino acid at a given position is often arbitrary, so the native and mutated amino acids have equal weights, seen by the frequent functionally neutral point mutations during evolution[50]. In contrast, the composition-based information is robust and evolutionarily conserved, since any significant change requires clearly concerted multiple point mutations.

It is evident that $M_{bind}(n)$ there is a clear boundary between survivin binding and non-binding regions along the sequence (Figure 5) and discontinuous jumps are also prevalent. Amino acids in the binding regions seem to have significantly lower $M_{bind}(n)$ than its surroundings, suggesting that native residue at position n is crucial for maintaining the affinity of the surrounding region, and one might anticipate a loss of function (binding) associated with point mutations at that specific position. Conversely, a higher $M_{bind}(m)$ compared to surroundings for an amino acid in a non-binding region indicates that the native residue at position m is unsuitable for survivin binding within the given context, but other residues at m can potentially convert peptides into survivin-binders. Consequently, position m is likely associated with gain-of-function mutations resulting in survivin binding.

Out of all identified survivin interaction partners, only the ternary complex of CPC involving survivin, borealin, and INCENP[51] has been structurally characterized (Figure 5A). This complex is connected via N-terminal segments of borealin and INCENP, together with coiled-coil of survivin. While the helices of survivin and INCENP align in parallel, the helix of borelin lies anti-parallel to them. Survivin is predicted to bind to colocalized regions of borealin and INCENP (residues 27-46 in INCENP and 15-29 in borealin, in opposite directions). Borealin exhibits another predicted high-affinity region (59-76), which is observable in the crystal structure. Notably, INCENP and, to some extent, borealin, possess robust binding sequences compatible with survivin at other locations, which may come into short-range contact with survivin during their interactions.

Aurora kinase B plays an important role in phosphorylating CPC subunits, including survivin (Figure 5B). Survivin may also equally target Aurora kinase B active site, rendering it a favourable substrate. Aurora kinase B and A share 70% identity in the catalytic domain, yet the point mutation G198N transforms Aurora kinase A into B in terms of their specificity to interact with INCENP and survivin[52]. The high $M_{bind}(198)$ of Aurora kinase A (Figure S4), flanked by lower $M_{bind}(197)$ and $M_{bind}(199)$ from tyrosine residues (ILRL**Y**G**Y**FHDA), suggests that replacing glycine with any other residue enhances survivin binding in that region, while removing even one tyrosine likely diminishes survivin affinity.

STAT3 is another survivin interacting partner[53], which exhibits no extensive regions with high survivin binding, seeing almost no $M_{bind}(n)$ value reaching 1.0 (Figure 5D). However, both the N- and C-terminal regions generally show elevated $M_{bind}(n)$, with the latter encompass the transactivation (responsible for survivin interaction[53]) and SH2 domains near the DNA[54], which is localized by $M_{bind}(n)$ with precision. The first helix of the coiled coil domain also displays heightened $M_{bind}(n)$[54].

Proteome-wide survivin binding prediction and overview of potential binding functions

We applied our survivin binding prediction model on the entire human proteome, leveraging the UniProt database. A higher binding ratio parameter and higher number of predicted binder was used as a parameter to judge a higher likelihood for affinity to survivin. The proteins with the highest binding ratios were used in enrichment analysis of biological annotations using Metascape[55] and the network representation was mapped (Figure 6). We also examined the background of T-cell specific proteins instead of the entire proteome (Figure S5) Nuclear, cytosolic, and mitochondrial proteins, including members of the respiratory chain, are present among survivin binders, which is in accordance with the fact that survivin is firmly established to localize within these compartments. As survivin is anti-apoptotic, it is encouraging to see the group of enriched apoptotic proteins including caspase 6 and 8, Death-associated protein kinase 3 (DAPK3), High mobility group box 1 and 2 (HMGB1/2), phorbol-12-myristate-13-acetate-induced protein 1 (BCL2 related protein PMAIP1), Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta (YWHAZ), SEM1 26S Proteasome Subunit, Myeloid Differentiation factor 2 (MD-2), Apoptosis And Caspase Activation Inhibitor (AVEN) and Occludin (OCLN). These apoptotic proteins are non-homologous, catalyzing different reactions, and belonging to different compartments, yet their composition renders them attractive to survivin. The mitochondrial Smac/DIABLO protein, with its proapoptotic activity, has an average binding ratio of 0.31.

Within the mitochondria, survivin influences oxidative phosphorylation[56,57] and represented in Figure 6 by ATP synthase (subunits ε and K), NADH:ubiquinone oxidoreductase (subunits chain 6, A1, B4, B5, B7, B9, B10, C2, iron sulfur proteins S5 and S8, CIA30, assembly factors NDUFAF1, NDUF8 and LYRM2), succinate dehydrogenase complex (assembly factor 2 and 3), cytochrome b/c1 complex (subunits VI and VII), cytochrome c oxidase (PET100, VIb and VIc, and assembly factors 8, COX14, COX19, COA5).

Figure 6 also shows a significant number of translation-related proteins and proteins involved in RNA metabolism, transcription, and splicing, revealing putative functions of survivin. Recent experimental insights on survivin in micro-RNA biogenesis[58] highlights the potential confirmation of binding predictions.

A particular limitation of enhancement analysis stems from arbitrarily selecting the number of proteins enriched in survivin binding regions. Opting too few may overlook important survivin functions, while choosing too many may reduce the enrichment analysis sensitivity. Estimating the number of relevant survivin interaction partners is challenging with the information scarcity, although we can still potentially establish a lower limit. How survivin molecules, with its small solvent-accessible surface, engage in specific pairwise interactions also pose major question, which may be caused by our conceptual limitation by understanding PPIs as pairwise short-range interactions organized in pathways. The alternative in which survivin interacts with potentially all cellular proteins via long-range interactions is theoretically possible, but unattainable by statistical enrichment analysis. This radically different view on biological systems necessitates a discussion that probes whether there are physicochemical mechanisms in the background justifying the observed agreement. As we will discover, the answer is positive.

Exploring the effectiveness of composition-based prediction

This study suggests exploring new approaches to understand and categorize interactions between proteins or biopolymers. While it is uncertain whether some fundamental concept may be missing, it is evident that variations in amino-acid compositions effectively capture structural and interacting protein features. Sequence and patterning information is certainly significant[59], but it may not always offer a convenient representation for certain protein-related challenges. To explore new avenues, classical and quantum theories of collective excitations may be fruitful, as well as the physics of polymers and colloids with phase transition models. Current literature indicates the necessity for this cross-fertilization, as it is exemplified by the existence of intrinsically disordered proteins (IDP) or unstructured domains under physiological conditions[60]. Recent hypotheses propose that liquid-liquid phase separations might be inherent to proteins/polypeptides, irrespective of their sequence and structure (but with distinct interactions, electrostatic, hydrophobic, and H-bonding[61]). Understanding protein interactions at a compositional level can impact, vice versa, the study of phase transitions, such as the formation of nuclei via interactions with parent particles or existing nuclei in the bulk (secondary nucleation).

Initially, our study, departing from structural information, focused on atom displacement analysis in folded proteins. The observation that atoms with similar chemical properties exhibit shared displacement patterns[62,63] led to a dynamic model in second-quantization, revealing coherent motion of phononic and dipolar collective excitations within the protein crystal[64]. This coherence extends between different proteins over intermolecular distances, involving identical functional groups. Testing this unconventional idea yielded a surprisingly simple and effective initial approximation[13].

Motivated by the chemical principle "like dissolves like," our view extends the concept of miscibility between molecules, emphasizing atom counting and categorizing functional groups. The success of atom counting, particularly in predicting water miscibility, relies on identifying functional groups—collections of identical or different elements with a localized electronic structure. The C/O ratio, for instance, serves as a primary descriptor[65] for inferring water solubility in organic molecules. However, exceptions to this trend exist[66], and secondary selection criteria include molecular size and structure.

The miscibility concept does not lead to a clear binary choice between hydrophobic and hydrophilic compounds. This is illustrated by per- and poly-fluoroalkyl substances (PFAS), which don't mix well with typical hydrophobic or hydrophilic substances. Our results suggest that the miscibility of simple molecules should parallel the interaction (or 'affinity') of proteins with different functional groups. Applying this logic to biology, protein interactions should be re-evaluated considering mechanisms leading to phase transitions or separations in condensed matter. Non-interacting proteins may not be neutral as well; their incompatibility can drive active structuring/segregation into separate phases. Analogies with polymer, colloid and interface sciences, such as "irreversible aggregation" or "reversible self-assembly" were proposed, with connections to diseases like amyloid fibrils formation[67].

Despite several advancements, quantitative challenges persist. Understanding elementary mechanisms behind protein transitions remains tough due to compensating energy–entropy effects at various length scales[68]. The realistic free energy behaviour of macromolecular systems with heterogeneous composition is a non-local functional of monomer density distributions[69], challenging the derivation of chemical potentials from phase diagrams. This complexity highlights the potential impact of more synoptic tools, such as the one based on atomic composition. However, the underlying explanation for empirical models may be more related to the dynamic behaviour of phases than their static properties alone. Common phase transitions can be associated with bifurcation theory, predicting significant perturbations in atomic dynamics with small alterations in control parameters[70].

It has been suggested that proteins may operate near a critical point where distinct phases merge, given their exceptional functional sensitivity and mutational robustness[71]. This delicate balance would allow proteins to work at the edge of instability, demonstrating both high "plasticity" to small environmental fluctuations and high structural stability to maintain integrity. Formally, the critical temperature should depend on protein concentration, reaching a tricritical state when the folding temperature approaches the Θ value[72], akin to a continuous coil-globule transition in synthetic polymers[73]. Acknowledging features from phase transitions of both first-order (e.g. mass density) and second-order (entropy) seems to be essential in current theoretical interpretations. A first-order transition then can become continuous in the presence of a surface[74].

A relevant example is the melting of a material, transitioning from an ordered solid to a liquid with disordered fluctuations. At the critical temperature, lattice vibrations significantly decrease, and the atomic/molecular constituents adopt a liquid-like diffusive regime. For a one-dimensional lattice of atoms, the equilibrium fluctuation of the mean squared displacement follows a spatial random walk[75], akin to the law governing the end-to-end distance in a polymer as a function of molecular weight. This relation depends on factors like equilibrium spacing, temperature, and force constants connecting atomic pairs and discussed in the accompanying text under the heading ‘Fluctuations in a Monoatomic Lattice’. For a linear diatomic lattice with long-range oscillations (acoustic branch), the overall fluctuations turn out to be affected by the atomic composition by the mass proportions of their species and this model is discussed in the accompanying text under the heading ‘Fluctuations in a Diatomic Lattice’. At the melting point, deviations from a perfectly periodic structure can be quantified while maintaining compositional-dependent long-range order. In reality, melting is a surface-initiated unstable process[76], with a characteristic temperature approximately (20-30)% lower than the bulk value. Nevertheless, this remark does not invalidate the previous conclusions.

A representative mean field theory for phase separation in polymeric mixtures is Flory-Huggins’ (FH) and its variants like Overbeek-Voorn’s (OV) or Edmond-Ogston’s (EO). OV improves FH with long-range electrostatic interactions, relevant for complex coacervation [77], while EO is a (truncated) virial expansion[78]. FH assumes short-ranged, pointwise force fields without distinguishing between charged and neutral residues. OV does not consider the sign of charges. Random phase approximations (RPAs) may be used for a patterning description of charged units, provided the transition is not driven by strong critical fluctuations[79]. The key quantity remains the FH parameter (χ), representing the energy cost for having lattice sites adjacent to the polymer occupied by solvent units, which we highlight in the accompanying text under the heading ‘Notes on Flory & Huggins’ (FH) solution theory’. Phase separation/aggregation occurs when the solvent is poor for the macromolecules (e.g. proteins rich in polar amino acids in water)[80]. When the segregation degree surpasses a critical value, the enthalpy contribution to separation prevails, resulting in a solution of (almost) pure phases in equilibrium. Discussing the magnitude of χN (N = polymer repeat unit number) aids in defining the character of segregation and the incompatibility unit at a microphase scale (e.g. the “oligo-nucleosomal clutch” in heterochromatin-like domains subjected to order-disorder transitions[81]).

In FH-like models, residues are treated as independent repeat units, lacking consideration for chain connectivity or molecular details, limiting the ability to address sequence- and structure-dependent interactions. Despite these constraints, this framework successfully balances energies contributing to phase separation[79,82]. Even a standard FH approach with average χ accurately predicts the critical temperature for the germ granule protein Ddx4[83]. The phase transition rate is governed by the extent of concentration and conformational fluctuation, reaching the largest amplitudes at the critical point, inducing spontaneous demixing. Enhancing protein-protein interactions results in multiple energy minima, defining the composition of separated phases along the volume fraction coordinate[82]. In a mixture of neutral molecules in a macromolecular blend, the inflection point on the FH free energy curve determines a compositional-dependent critical point.

From the viewpoint of associative polymer physics, a biological molecule may be modelled as a sequence of stickers and spacers, carriers of attractive forces and non-attractive macromolecular segments, respectively[84]. This connection to associative polymer physics is discussed in the accompanying online text under the heading ‘Notes on Associative Polymer Physics and Percolation’. Stickers are expected to be short linear motifs (1–10 residues, SLiMs) in intrinsically disordered domains, or even single nucleotides in unfolded RNA molecules. This perspective arises from the idea that many membraneless biological condensates result from phase transformations and percolation phenomena[85]. Percolation occurs when, at a certain sticker concentration threshold, macromolecules form a system-spanning network. For a mixture of different polymers with stickers, the percolation threshold is composition-dependent, representing a measure of the attractive volume related to the sticker pair[84].

Considering the number of identical atoms or functional groups as an order parameter reveals substantial changes when reaching a certain threshold in a given macromolecular domain. For instance in a peptide with 15 amino acids and a volume of 2000 Å^3, having from 2 to 15 identical functional groups sets a concentration $c_{eff} \approx n_F/(N_A l_P \; 135 \; \text{Å}^3) = (1.7 - 12.5)\; M$, where n_F is the number of functional groups or atoms and l_P is the polypeptide length.[86]. This range, on the average, remains valid even though there is only one peptide molecule (i.e. an infinitely diluted solution). Adopting this approximation, if the number of functional groups is taken for simplicity as proportional to the polypeptide length, the effective concentration remains constant, regardless of the peptide or protein size. The concept of scale-freeness aligns with the use of atomic ratios as predictive tools, and a phase transition can occur in rather small protein segments of heavily biased compositions.

The behaviour of peptide or protein solutions at the local composition level differs significantly from low molecular weight liquids where all units are uniformly dispersed. Due to covalent bonding, residues are locally constrained to lattice-like structures influenced by force fields such as residue-residue and solvent-residue interactions (solvation, hydrophobic forces, weak interactions, etc.) as well as the amino-acid sequence. This implies that a peptide segment may exhibit a substantial compositional bias, potentially being as impactful as a similarly biased composition in a larger protein. However, when an equivalent number of identical amino acids are uniformly dispersed within a larger protein, it might not necessarily induce a phase transition. In compounds highly soluble and freely diffusing in a given environment, the solubility limit often falls within the former range (1.7-12.5 M), leading to precipitation and the emergence of highly ordered coupled motions reminiscent of those detected in soft matter and materials[87]. This process, when associated with a bifurcation point, could facilitate the folding of these structures into a 3D configuration, akin to the principles governing crystal nucleation & growth, highlighting a dynamic attractor resistant to perturbations. Alternatively, it may give rise to distinct periodic coordinated dynamics rather than incoherent motions. Dynamic transitions can also be initiated by control parameters like temperature changes or the application of a pulling force, as observed in unfolding[88] or protein unfolding through pulling[39].

Over billions of years, organisms have leveraged fundamental physical principles, exemplified by signal peptides like nuclear localization signals with significant compositional biases[89]. Such peptides, despite low information content, hold substantial abstract signal value due to their unique enrichment in specific functional groups, as seen in the PKKKRKV segment found in the SV40 Large T-antigen. Amino acid distributions, primarily determined by genetic material, are further influenced by covalent post-translational modifications like phosphorylation, acetylation, methylation, ubiquitination, nitrosylation, hydroxylation, sulfation and deamidation. These modifications, observed individually or in patterns like hyperphosphorylation, enable protein composition to adapt, triggering new collective dynamics based on introduced groups and the organism's physiological state.

In summary, this work has focused on the predictive influence of elemental composition in protein-protein interaction systems, establishing its conceptual alignment with biological processes and molecular functions. Although finer mechanistic details await further elucidation and formalization, our identification of key questions signals the need for exploration to unravel the intricate organization of biological organisms.

METHODS

Peptide microarray experiments

The peptide microarray experiment was described previously[13] and corresponding data is the basis of this analysis[90]. Briefly, peptide microarray was designed with a total of 36 proteins using PEPperCHIP Peptide Microarrays (PEPperPRINT Gmbh). A complete list of proteins and more detailed description of the method can be found in previous research[13]. Each protein sequence was divided into 15 amino acid peptide units, with 10 amino acid overlap. Background interactions was examined by pre-staining one microarray with the secondary 6X His Tag Antibody DyLight680 antibody (1:1000) and monoclonal anti-HA (12CA5)-DyLight800 control antibody (1:1000). Another peptide microarray was incubated with survivin at a concentration of 1 µg/ml and stained with the secondary 6X His Tag Antibody DyLight680 antibody (Rockland Immunochemicals, Pottstown, PA, USA) and the monoclonal anti-HA (12CA5)-DyLight800 control antibody (Rockland Immunochemicals, Pottstown, PA, USA). The read-out was performed using LI-COR Odyssey Imaging System with scanning intensities of 7/7 (red/green). HA and His-tag peptides were also stained simultaneously in the assay as internal quality control. PepSlide Analyzer was used for quantification of spot intensities and peptide annotation. The resulting data was stored as a table with information on protein identifier, peptide sequence, and fluorescence intensity.

Development of feature scoring set for the representation of peptides

The features considered encompassed the standard four non-hydrogen main chain atoms (MC), except glycine (3 non-hydrogen atoms) and proline (2 non-hydrogen atoms). Additionally, specific features included the glycine Cα atom (CA-Gly), recognizing its methylene group nature instead of a methanetriyl-group. Proline-specific main chain atoms (Pro-MC) were considered due to the circular proline side chain linking the amide nitrogen to Cα, resulting in distinctive dynamics. The content of side chains were deconstructed to carboxyl groups of aspartate and glutamate (carboxyl), the amide groups of asparagine and glutamine (amide), the imidazole ring of histidine (His), the indole ring of tryptophan (Trp), the phenyl group of phenylalanine and tyrosine (Phe-Tyr), the phenolic hydroxyl group of tyrosine (OH-Tyr), the hydroxyl group of serine and threonine amino acids (OH), the sulfhydryl group or oxidized variants of cysteine (SH), the thioether of methionine (S), the amino group of lysine (NH3), and the guanidino group of arginine (Arg). Carbon atoms not part of the main chain, carboxyl, amide and aromatic functional groups were assigned to methyl- (-CH3), methylene- (-CH2-), and methanetriyl- (>CH-) groups. These categories resulted in 17 features, each corresponding to the number of non-hydrogen atoms included. (Table S1)

We also embarked on a stepwise simplification of the C-Coord1 model. Initially, we consolidated the features of phenolic and alkyl hydroxyl groups into a single feature. Additionally, we described glycine as comprising three standard main chain atoms and a methylene group, thus eliminating the necessity for a separate atom type for the Cα atom of glycine. This refinement resulted in the C-Coord2 model (Table S2), which incorporates 15 features. We further adjusted C-Coord2 model by restricting oxygen atoms in carboxyl and carbonyl side chains to be indistinguishable. Nitrogen atoms in non-heterocyclic side chains were compelled to be identical, and carbon atoms with no bonded hydrogens (C0) in amide, carboxyl, and guanidino side chain groups were also forced to be identical. These modifications led to the development of the C-Coord3 model, which encompasses 13 features. (Table S3)

Another method used here to characterize side chain carbon atoms involved considering their distance from the main chain (C-Pos, Table S4). The rationale is exemplified here by the varying dynamics of methylene side chain groups. For example, the Cβ atom of lysine forms a quasi-rigid group with the Cα, the main chain amide nitrogen, and carbonyl carbon, which in this (β) configuration mirrors the fluctuations of the main chain. On the other hand, the Cδ atom has a high degree of freedom, and its position is influenced by multiple χ torsion angles in the side chain. Suggesting that these two methylene groups exhibit similar behavior in models C-Coord1-3 could be an overly simplified abstraction when taking the dynamics of these atoms into account. The drawback of the C-Pos model is that number of features is 19 only one less than the number of naturally occurring amino acids and it does not distinguish between methyl- (-CH3), methylene- (-CH2-), and methanetriyl- (>CH-) groups. For comparison, the AA-Letter model (Table S5) was included to represent amino acid composition, assuming each amino acid is unique and possesses equal weight, resulting in a model with 20 features. We compared these four models with elemental composition models, where the features included the counts of hydrogen, carbon, nitrogen, oxygen, and sulfur atoms (5 features in "Chemical Formula", Table S6).

For the characterization of the electrostatic and hydrophobic properties of peptides, we used two different approaches. In the Pos/Neg/HyPho model (Table S7), we approximated these properties by considering the number of basic and acidic amino acids and whether they were classified as hydrophobic, resulting in 3 features. In the NetChrg/SumHyPho model (Table S8), we determined the net charge by subtracting the count of acidic amino acids from basic ones and calculated the sum of hydrophobicity according to the Kyte/Doolittle scale[44], yielding two features.

Applying the feature representation to peptide microarray dataset

Different sets of feature scorings were developed based on atom type composition and applied on each amino acid in the dataset. Briefly, a feature scoring table was created by listing 20 amino acids, with each row representing each distinct amino acid.

For feature representation based on atom type composition, each amino acid is represented by a set number of scores, according to the number of categories present in the scoring set, as shown in Figure 1. The scoring set was then applied to each peptide sequence from the microarray dataset, summing the scores of each atom type category from all amino acids in a particular peptide.

For feature representation based on AAindex numerical values, each amino acid is represented by 1 value based on RACS820104 index[91] from AAindex database[15]. Fourier transformation was applied on this one dimensional array of peptide representation using Fast Fourier Transform from scipy 1.6.2 package and only the real part of the Fourier Transform was considered.

t-SNE dimensionality reduction analysis of peptide microarray dataset

To calculate the spatial distance between peptides, t-SNE module from scikit-learn 1.0.1 package was applied to both atom type composition and AAindex representations. In the latter case, t-SNE was performed on the representation before (Seq, as in Figure S1C and S1D) and after (FT-Seq, as in Figure S1B) Fast Fourier Transform step. The following parameters were used: n_components=2, perplexity=30.0, verbose=1, and random_state=123.

Prediction of survivin binding based on the peptide microarray dataset

Peptides from the microarray dataset with non-zero fluorescence intensities were marked as binders, and the ones with zero intensities were marked as non-binders. 5388 peptides from the dataset were divided into equally large training and test sets. The model was built using multi-layer perceptron classifier from scikit-learn 1.0.1 on both sets with the following settings: solver='adam', learning_rate='constant', random_state=1, max_iter=10000, and activation='relu'. Hyperparameter tuning was done on solver, learning_rate, max_iter, hidden_layer_sizes, activation parameters. Fine tuning of the number of nodes is shown in Figure 2. Prediction statistics were obtained from the averages of 100 prediction runs.

Improving the sequence-based prediction of survivin binding using bidirectional LSTM

In bidirectional LSTM refinement, we made no alterations on the configuration similar to the one used for optimizing the dense layer, except by the addition or omission of one dense layer after the final bidirectional LSTM layer. We tested either the Keras Embedding Layer[92] or ProtT5 encoding[93], as well as testing up to two dense layers and two bidirectional LSTM layers. Additionally, the model incorporated dropout between layers to aid regularization.

To assess the impact of position, we utilized a straightforward model consisting of a one-dimensional embedding layer connected to a dense layer featuring a single intermediate layer neuron. This setup approximates a position-specific summation function, like our AA-Letter model, which has explicit summation of amino acids ignoring their positions.

Highlighting predicted survivin binding regions and measuring the impact of point mutation in survivin binding proteins

The metric for point mutation impact on the surrounding residues $M_{bind}(n)$, where 'n' represents a residue position along the 15 amino acid peptide sequence, is based on the number of point mutations required to make said residue being predicted to bind to survivin. Due to 10-residue overlap between 15 residue-long peptides, a residue position is typically found in three unique peptides, and with 20 natural amino acid variants, we ended up with 60 peptide permutations, each with a binding prediction. The total number of mutations predicted to be binders divided by the total number of peptides with variants at 'n' equals $M_{bind}(n)$.

Application to human proteome and gene ontology enrichment in predicted survivin binder proteins

Human proteome database was obtained from UNIPROT. The sequences of all proteins in the database were used to generate a dataset containing 15mer peptide units with 10 amino acid overlaps, mimicking the peptide microarray experiments. This dataset was also reduced to exclude peptides with cysteines at position 15. The C-Coord1 model with two neurons in the intermediate layer was applied on this dataset and the results were consolidated to get the binding ratio of each protein (R_{bind}, the number of binder peptides divided by the total number of peptides each protein contains). Highly ranked proteins were subjected to biological annotation enrichment analysis using Metascape (https://metascape.org)[55]. Network mapping of enrichment annotation terms from predicted survivin interaction partners was made using the Cytoscape platform[94].

Acknowledgements

This work has been funded by grants from the Röntgen-Ångström Cluster Framework of the Swedish Research Council (G.K., 2015-06099), the Swedish Research Council (M.I.B., 2017-03025), the Swedish Association against Rheumatism (M.I.B., R-566961, R-751351 and R-860371), the King Gustaf V:s 80-year Foundation (M.I.B.), the Regional agreement on medical training and clinical research between the Western Götaland county council and the University of Gothenburg (M.I.B., ALFGBG-717681 and ALFGBG-965623). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 964203 (Long-range electrodynamic INteractions between proteinS – LINkS). S.A.M. acknowledges financial support from HrZZ (Croatian Science Foundations, Grant ZFM – 0352Z32 – HZZ – POTENTIALS – 3456).

Author Contributions

G.K. and S.A.M. conceptualized the study outline and discussion. G.K. supervised the research. A.L.A., T.O.N., M.J. and M.J.-G.B. and G.K. analyzed the results from the peptide microarray. M.I.B. and S.P.W. contributed to the discussion regarding the cellular function of survivin. The manuscript was prepared by A.L.A., T.O.N., S.A.M., and G.K. with additional input from all authors.

Declaration of interest

G.K. and M.I.B. submitted a patent application for the machine learning method described in the paper. The remaining authors have no competing interests.

Inclusion and Diversity

We support inclusive, diverse, and equitable conduct of research.

References

1 Steiner, R. F. Reversible association processes of globular proteins. IV. Fluorescence methods in studying protein interactions. *Archives of Biochemistry and Biophysics* **46**, 291-311 (1953).

2 Oncley, J., Ellenbogen, E., Gitlin, D. & Gurd, F. Protein–protein interactions. *J. Phys. Chem.* **56**, 85-92 (1952).

3 Waugh, D. F. in *Advances in protein chemistry* Vol. 9 325-437 (Elsevier, 1954).

4 Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical journal* **49**, 463 (1951).

5 Lehner, B. *et al.* Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics* **83**, 153-167 (2004).

6 Suter, B., Kittanakom, S. & Stagljar, I. Two-hybrid technologies in proteomics research. *Current opinion in biotechnology* **19**, 316-323 (2008).

7 Sidhu, S. S., Fairbrother, W. J. & Deshayes, K. Exploring protein–protein interactions with phage display. *Chembiochem* **4**, 14-25 (2003).

8 Kodama, Y. & Hu, C.-D. Bimolecular fluorescence complementation (BiFC): a 5-year update and future perspectives. *Biotechniques* **53**, 285-298 (2012).

9 Cornett, E. *et al.* in *Method Enzymol* Vol. 574 31-52 (Elsevier, 2016).

10 Rothbart, S. B., Krajewski, K., Strahl, B. D. & Fuchs, S. M. in *Method Enzymol* Vol. 512 107-135 (Elsevier, 2012).

11 Chang, T.-W. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods* **65**, 217-223 (1983).

12 Breitling, F., Nesterov, A., Stadler, V., Felgenhauer, T. & Bischoff, F. R. High-density peptide arrays. *Mol Biosyst* **5**, 224-234 (2009).

13 Jensen, M. *et al.* Survivin prevents the polycomb repressor complex 2 from methylating histone 3 lysine 27. *iScience* **26**, 106976 (2023). https://doi.org/10.1016/j.isci.2023.106976

14 Kastritis, P. L. & Bonvin, A. M. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface* **10**, 20120835 (2013).

15 Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: amino acid index database. *Nucleic acids research* **27**, 368-369 (1999).

16 Schreiber, G. 78 edn 1-24.

17 Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-299 (1999).

18 Fox, J. M., Zhao, M., Fink, M. J., Kang, K. & Whitesides, G. M. The molecular origin of enthalpy/entropy compensation in biomolecular recognition. *Annual Review of Biophysics* **47**, 223-250 (2018).

19 Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nature Reviews Drug Discovery* **14**, 95-110 (2015).

20 Lafont, V. *et al.* Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chemical biology & drug design* **69**, 413-422 (2007).

21 VAN DAN BURG, B. *et al.* Protein stabilization by hydrophobic interactions at the surface. *European Journal of Biochemistry* **220**, 981-985 (1994). https://doi.org/https://doi.org/10.1111/j.1432-1033.1994.tb18702.x

22 Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1-9 (1998).

23 Pace, C. N. *et al.* Tyrosine hydrogen bonds make a large contribution to protein stability. *J Mol Biol* **312**, 393-404 (2001).

24 Reichmann, D. *et al.* Binding hot spots in the TEM1–BLIP interface in light of its modular architecture. *J Mol Biol* **365**, 663-679 (2007).

25 Reichmann, D. *et al.* The modular architecture of protein–protein binding interfaces. *Proceedings of the National Academy of Sciences* **102**, 57-62 (2005).

26 Cohen, M., Reichmann, D., Neuvirth, H. & Schreiber, G. Similar chemistry, but different bond preferences in inter versus intra-protein interactions. *Proteins: Structure, Function, and Bioinformatics* **72**, 741-753 (2008).

27 La, D., Kong, M., Hoffman, W., Choi, Y. I. & Kihara, D. Predicting permanent and transient protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **81**, 805-818 (2013).

28 Pál, G., Kouadio, J.-L. K., Artis, D. R., Kossiakoff, A. A. & Sidhu, S. S. Comprehensive and Quantitative Mapping of Energy Landscapes for Protein-Protein Interactions by Rapid Combinatorial Scanning*♦. *J. Biol. Chem.* **281**, 22378-22385 (2006).

29 Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021.2010. 2004.463034 (2021).

30 Yu, D., Chojnowski, G., Rosenthal, M. & Kosinski, J. AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer. *Bioinformatics* **39**, btac749 (2023).

31 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).

32 Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Pairing interacting protein sequences using masked language modeling. *bioRxiv*, 2023.2008. 2014.553209 (2023).

33 Heffernan, R., Yang, Y., Paliwal, K. & Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **33**, 2842-2849 (2017).

34 Zhang, B., Li, J., Quan, L., Chen, Y. & Lü, Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* **357**, 86-100 (2019).

35 Liu, J. & Gong, X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC bioinformatics* **20**, 1-11 (2019).

36 Cadet, F. *et al.* A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Scientific reports* **8**, 16757 (2018).

37 Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *J Mol Biol* **338**, 181-199 (2004).

38 Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* **13**, 190-202 (2004).

39 Bordner, A. J. & Abagyan, R. Statistical analysis and prediction of protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **60**, 353-366 (2005).

40 Zhou, H. X. & Shan, Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Bioinformatics* **44**, 336-343 (2001).

41 Hwang, H., Petrey, D. & Honig, B. A hybrid method for protein–protein interface prediction. *Protein Science* **25**, 159-165 (2016).

42 Xue, L. C., Dobbs, D., Bonvin, A. M. & Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS letters* **589**, 3516-3526 (2015).

43 van Mierlo, G., Veenstra, G. J. C., Vermeulen, M. & Marks, H. The Complexity of PRC2 Subcomplexes. *Trends Cell Biol* **29**, 660-671 (2019). https://doi.org/10.1016/j.tcb.2019.05.004

44 Kyte, J. & Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J Mol Biol* **157**, 105-132 (1982). https://doi.org/Doi 10.1016/0022-2836(82)90515-0

45 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015). https://doi.org/10.1038/nature14539

46 Oren, Y., Meister, N., Chatterji, N., Ladhak, F. & Hashimoto, T. B. Proving Test Set Contamination in Black Box Language Models. *ArXiv e-prints*, 2310.17623 (2023).

47 Stadler, V. *et al.* Combinatorial synthesis of peptide arrays with a laser printer. *Angewandte Chemie International Edition* **47**, 7132-7135 (2008).

48 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J Mach Learn Res* **9** (2008).

49 Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, e2 (2016).

50 Saitou, N. in *Introduction to Evolutionary Genomics* 109-148 (Springer International Publishing, 2018).

51 Jeyaprakash, A. A. *et al.* Structure of a Survivin-Borealin-INCENP core complex reveals how chromosomal passengers travel together. *Cell* **131**, 271-285 (2007). https://doi.org/10.1016/j.cell.2007.07.045

52 Fu, J., Bian, M., Liu, J., Jiang, Q. & Zhang, C. A single amino acid change converts Aurora-A into Aurora-B-like kinase in terms of partner specificity and cellular function. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 6939-6944 (2009). https://doi.org/10.1073/pnas.0900833106

53 Wang, H. *et al.* Acetylation directs survivin nuclear localization to repress STAT3 oncogenic activity. *J. Biol. Chem.* **285**, 36129-36137 (2010). https://doi.org/10.1074/jbc.M110.152777

54 Sgrignani, J. *et al.* Structural Biology of STAT3 and Its Implications for Anticancer Therapies Development. *Int J Mol Sci* **19** (2018). https://doi.org/10.3390/ijms19061591

55 Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications* **10**, 1523 (2019).

56 Rivadeneira, D. B. *et al.* Survivin promotes oxidative phosphorylation, subcellular mitochondrial repositioning, and tumor cell invasion. *Sci Signal* **8**, ra80 (2015). https://doi.org/10.1126/scisignal.aab1624

57 Hagenbuchner, J., Kuznetsov, A. V., Obexer, P. & Ausserlechner, M. J. BIRC5/Survivin enhances aerobic glycolysis and drug resistance by altered regulation of the mitochondrial fusion/fission machinery. *Oncogene* **32**, 4748-4757 (2013). https://doi.org/10.1038/onc.2012.500

58 Andersson, K. M. *et al.* Survivin controls biogenesis of microRNA in smokers: A link to pathogenesis of rheumatoid arthritis. *Biochim Biophys Acta Mol Basis Dis* **1863**, 663-673 (2017). https://doi.org/10.1016/j.bbadis.2016.11.033

59 Vovk, A. & Zilman, A. Effects of Sequence Composition, Patterning and Hydrodynamics on the Conformation and Dynamics of Intrinsically Disordered Proteins. *Int J Mol Sci* **24** (2023). https://doi.org/10.3390/ijms24021444

60 Belo, Y. *et al.* Unexpected implications of STAT3 acetylation revealed by genetic encoding of acetyl-lysine. *Biochim Biophys Acta Gen Subj* **1863**, 1343-1350 (2019). https://doi.org/10.1016/j.bbagen.2019.05.019

61 Poudyal, M. *et al.* Intermolecular interactions underlie protein/peptide phase separation irrespective of sequence and structure at crowded milieu. *Nature communications* **14**, 6199 (2023). https://doi.org/10.1038/s41467-023-41864-9

62 Ahlberg Gagner, V., Jensen, M. & Katona, G. Estimating the probability of coincidental similarity between atomic displacement parameters with machine learning. *Machine Learning: Science and Technology* **2**, 035033 (2021). https://doi.org/10.1088/2632-2153/ac022d

63 Gagnér, V. A. *et al.* Clustering of atomic displacement parameters in bovine trypsin reveals a distributed lattice of atoms with shared chemical properties. *Scientific Reports* **9**, 19281 (2019). https://doi.org/10.1038/s41598-019-55777-5

64 Ahlberg Gagner, V. *et al.* Femtosecond X-ray snapshots reveal correlated displacements of specific distal atoms in a protein crystal. *Manuscript submitted for publication* (2023).

65 Ebbing, D. D. & Gammon, S. D. *General Chemistry*. 484 (Houghton Mifflin, 2010).

66 Ensing, B. *et al.* On the origin of the extremely different solubilities of polyethers in water. *Nature communications* **10**, 2893 (2019). https://doi.org/10.1038/s41467-019-10783-z

67 Ezzat, K., Sturchio, A. & Espay, A. J. Proteins Do Not Replicate, They Precipitate: Phase Transition and Loss of Function Toxicity in Amyloid Pathologies. *Biology* **11**, 535 (2022).

68 van der Vegt, N. F. A. Length-Scale Effects in Hydrophobic Polymer Collapse Transitions. *The Journal of Physical Chemistry B* **125**, 5191-5199 (2021). https://doi.org/10.1021/acs.jpcb.1c01070

69 Panyukov, S. V. & Kuchanov, S. I. New statistical approach to the description of spatial inhomogeneous states in heteropolymer solutions. *J. Phys. II France* **2**, 1973-1993 (1992).

70 Bose, I. & Ghosh, S. Bifurcation and criticality. *J Stat Mech-Theory E* (2019). https://doi.org/10.1088/1742-5468/ab11d8

71 Tang, Q. Y., Hatakeyama, T. S. & Kaneko, K. Functional sensitivity and mutational robustness of proteins. *Phys Rev Res* **2** (2020). https://doi.org/10.1103/PhysRevResearch.2.033452

72 Lifshitz, I. M., Grosberg, A. Y. & Khokhlov, A. R. Some problems of the statistical physics of polymer chains with volume interaction. *Reviews of Modern Physics* **50**, 683-713 (1978). https://doi.org/10.1103/RevModPhys.50.683

73 Gasic, A. G. *et al.* Critical Phenomena in the Temperature-Pressure-Crowding Phase Diagram of a Protein. *Phys Rev X* **9**, 041035 (2019). https://doi.org/10.1103/PhysRevX.9.041035

74 Kosterlitz, J. M. & Thouless, D. J. Ordering, metastability and phase transitions in two-dimensional systems. *Journal of Physics C: Solid State Physics* **6**, 1181 (1973). https://doi.org/10.1088/0022-3719/6/7/010

75 Peierls, R. *Surprises in theoretical physics*. (Princeton University Press, 1979).

76 Pietronero, L. & Tosatti, E. Surface theory of melting. *Solid State Communications* **32**, 255-259 (1979). https://doi.org/https://doi.org/10.1016/0038-1098(79)90133-9

77 Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nat Phys* **11**, 899-904 (2015). https://doi.org/10.1038/Nphys3532

78 Bot, A., Dewi, B. P. C. & Venema, P. Phase-Separating Binary Polymer Mixtures: The Degeneracy of the Virial Coefficients and Their Extraction from Phase Diagrams. *ACS Omega* **6**, 7862-7878 (2021). https://doi.org/10.1021/acsomega.1c00450

79 Lin, Y. H., Song, J. H., Forman-Kay, J. D. & Chan, H. S. Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins. *J Mol Liq* **228**, 176-193 (2017). https://doi.org/10.1016/j.molliq.2016.09.090

80 Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102-112 (2015). https://doi.org/https://doi.org/10.1016/j.sbi.2015.03.008

81 Singh, P. B., Belyakin, S. N. & Laktionov, P. P. Biology and Physics of Heterochromatin-Like Domains/Complexes. *Cells* **9**, 1881 (2020).

82 Martin, E. W. & Mittag, T. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry* **57**, 2478-2487 (2018). https://doi.org/10.1021/acs.biochem.8b00008

83 Nott, Timothy J. *et al.* Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell* **57**, 936-947 (2015). https://doi.org/https://doi.org/10.1016/j.molcel.2015.01.013

84 Choi, J. M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annu Rev Biophys* **49**, 107-133 (2020). https://doi.org/10.1146/annurev-biophys-121219-081629

85 Berry, J., Brangwynne, C. P. & Haataja, M. Physical principles of intracellular organization via active and passive phase transitions. *Reports on Progress in Physics* **81**, 046601 (2018). https://doi.org/10.1088/1361-6633/aaa61e

86 Chen, C. R. & Makhatadze, G. I. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. *BMC Bioinformatics* **16**, 101 (2015). https://doi.org/10.1186/s12859-015-0531-2

87 Mezzasalma, S. A. *et al.* Light-driven Self-Oscillation of Thermoplasmonic Nanocolloids. *Adv. Matter.* **in press** (2023). https://doi.org/10.1002/adma.202302987

88 Seelig, J. & Schonfeld, H. J. Thermal protein unfolding by differential scanning calorimetry and circular dichroism spectroscopy Two-state model versus sequential unfolding. *Quarterly Reviews of Biophysics* **49** (2016). https://doi.org/10.1017/S0033583516000044

89 Labaj, P. P., Sykacek, P. & Kreil, D. P. An analysis of single amino acid repeats as use case for application specific background models. *Bmc Bioinformatics* **12** (2011). https://doi.org/10.1186/1471-2105-12-173

90 Jensen, M., Chandrasekaran, V., García-Bonete, M.-J., Li, S., Anindya, A. L., Andersson, K., Erlandsson, M. C., Oparina, N. Y., Burmann, B. M., Brath, U., Panchenko, A. R., Bokarewa I., M., & Katona, G. . (Zenodo, 2023).

91 Rackovsky, S. & Scheraga, H. Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids. *Macromolecules* **15**, 1340-1346 (1982).

92 Keras (2015).

93 Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *Ieee T Pattern Anal* **44**, 7112-7127 (2022). https://doi.org/10.1109/Tpami.2021.3095381

94 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003). https://doi.org/10.1101/gr.1239303

95 Serena, M., Bastos, R. N., Elliott, P. R. & Barr, F. A. Molecular basis of MKLP2-dependent Aurora B transport from chromatin to the anaphase central spindle. *J Cell Biol* **219** (2020). https://doi.org/10.1083/jcb.201910059

96 Elkins, J. M., Santaguida, S., Musacchio, A. & Knapp, S. Crystal structure of human aurora B in complex with INCENP and VX-680. *J Med Chem* **55**, 7841-7848 (2012). https://doi.org/10.1021/jm3008954

97 Diebold, M., Schonemann, L., Eilers, M., Sotriffer, C. & Schindelin, H. Crystal structure of a covalently linked Aurora-A-MYCN complex. *Acta Cryst D* **79**, 1-9 (2023). https://doi.org/10.1107/S2059798322011433

	NB	B	NB	B	NB	B	
C-Coord2	81.6 ± 0.3	76.0 ± 0.4	87.9 ± 0.3	65.8 ± 0.5	84.6 ± 0.2	70.5 ± 0.3	79.8 ± 0.2
C-Coord2 every 3rd peptide	**80.6 ± 0.6**	69.9 ± 1.4	83.4 ± 1.2	**65.5 ± 1.2**	81.9 ± 0.6	67.4 ± 0.8	76.8 ± 0.7
C-Pos	81.9 ± 0.3	77.7 ± 0.5	89.0 ± 0.3	66.1 ± 0.6	85.3 ± 0.1	71.4 ± 0.3	80.6 ± 0.2
C-Pos every 3rd peptide	80.4 ± 0.6	**75.3 ± 0.9**	**87.6 ± 0.7**	63.6 ± 1.3	**83.8 ± 0.2**	**68.8 ± 0.5**	**78.7 ± 0.3**
Seq	64.3 ± 0.3	60.2 ± 5.0	98.1 ± 0.4	5.6 ± 1.0	77.7 ± 0.2	10.0 ± 1.8	64.2 ± 0.3
UniSeq	63.7 ± 0.2	47.8 ± 1.5	96.6 ± 0.4	5.3 ± 0.5	76.8 ± 0.2	9.4 ± 0.9	63.1 ± 0.2
FT-Seq	63.2 ± 0.2	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	77.4 ± 0.2	0.0 ± 0.0	63.2 ± 0.2

Table 1. Performance indicators of the survivin binding peptide predictions. Uncertainty is represented by the standard error of the mean.

		Predicted label	**Every peptide**		**Predicted label**	**Every Third Peptide**
		Non-interacting	**Interacting**		**Non-interacting**	**Interacting**
Correct label	**Non-interacting**	1218±4	159±3	**Non-interacting**	379±2	77±2
	Interacting	228±4	1093±4	**Interacting**	110±3	333±3
Precision (%)		84.3±0.2	87.4±0.2		77.8±0.4	81.6±0.4
Recall (%)		88.5±0.2	82.7±0.3		83.2±0.5	75.2±0.6
F1-score (%)		86.3±0.1	85.0±0.1		80.2±0.2	78.0±0.3
Accuracy (%)		85.7±0.1			79.2±0.2	

Table 2. Confusion matrix for the prediction of survivin-microarray peptide interactions using the LSTM model using the cross-validation protocol. Uncertainty is represented by the standard error of the mean.

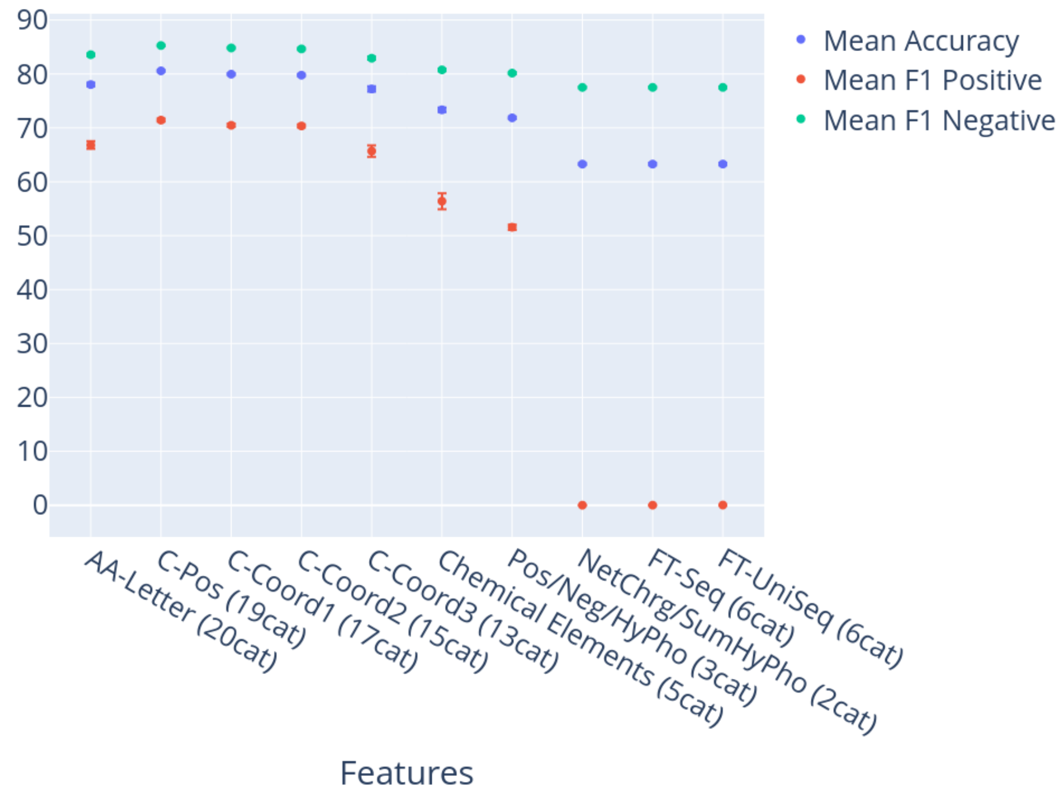


Figure 1. Performance evaluation of a multilayer perceptron classification with a single intermediate layer consisting of two perceptrons. Different sets of features describing the peptides on the peptide microarray were compared. Training was carried out using 50% of the data, repeated 100 times with varying randomly selected subsets, and testing was performed on the remaining half of the dataset.

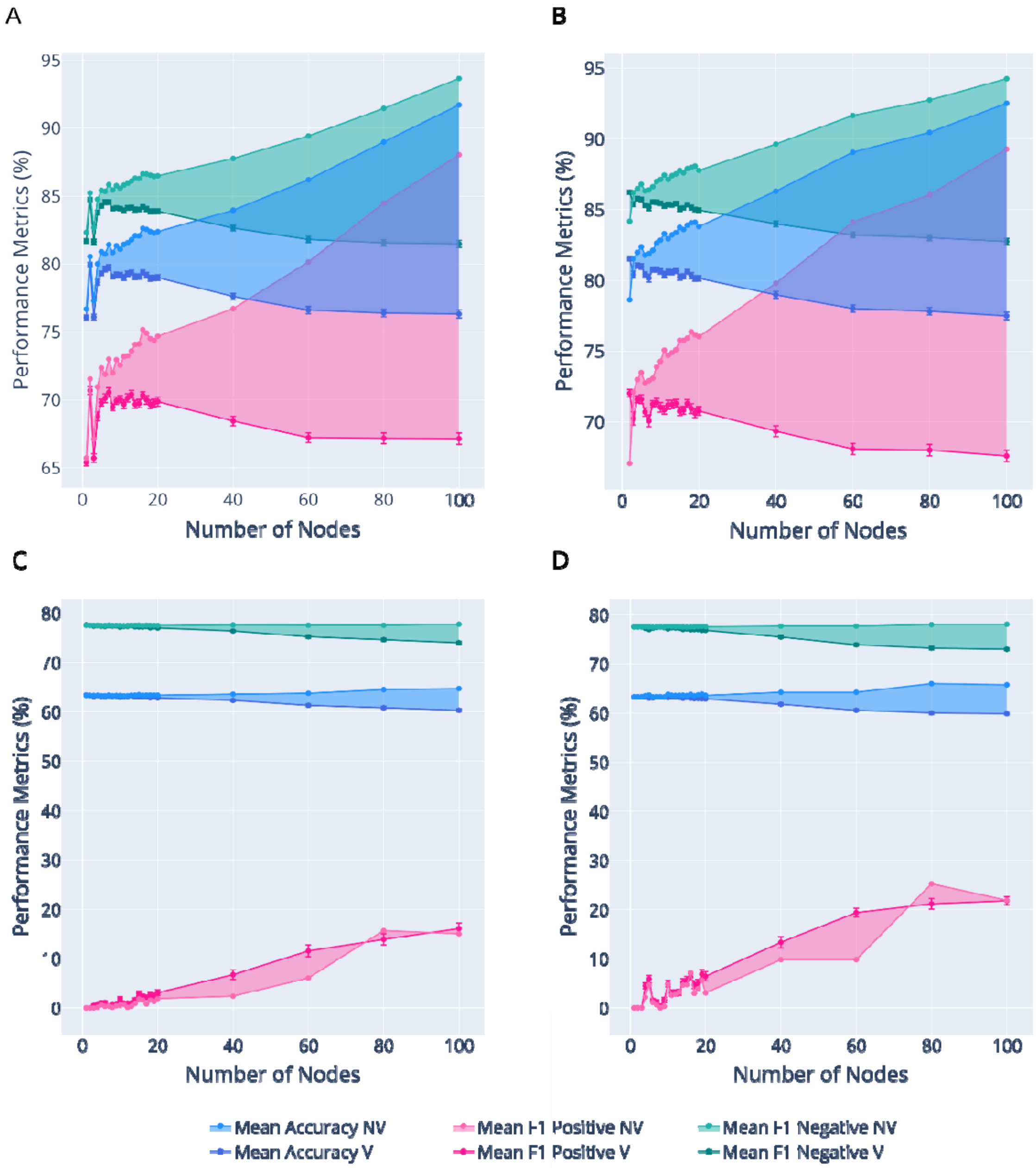


Figure 2. Performance metric value trends across a range of number of nodes (1 – 100) in different prediction models using: A. 15 atomic composition features (C-Coord2), B. 19 atomic composition features (C-Pos), C. Fourier Transform of non-uniform value representation of amino acids (FT-Seq), and D. Fourier Transform of uniform value representation of amino acids (FT-UniSeq). Performance metrics used are mean accuracy (blue), mean F1 values of 'binder' peptides (F1 Positive, in cyan), and mean F1 values of 'non-binder' peptides (F1 Negative, in pink). Filled areas between each performance metric show the difference between validated (V) and non-validated (NV) prediction results, with NV data points in lighter colours. All statistical values were calculated from 100 prediction runs; error bars represent the standard error of the mean.

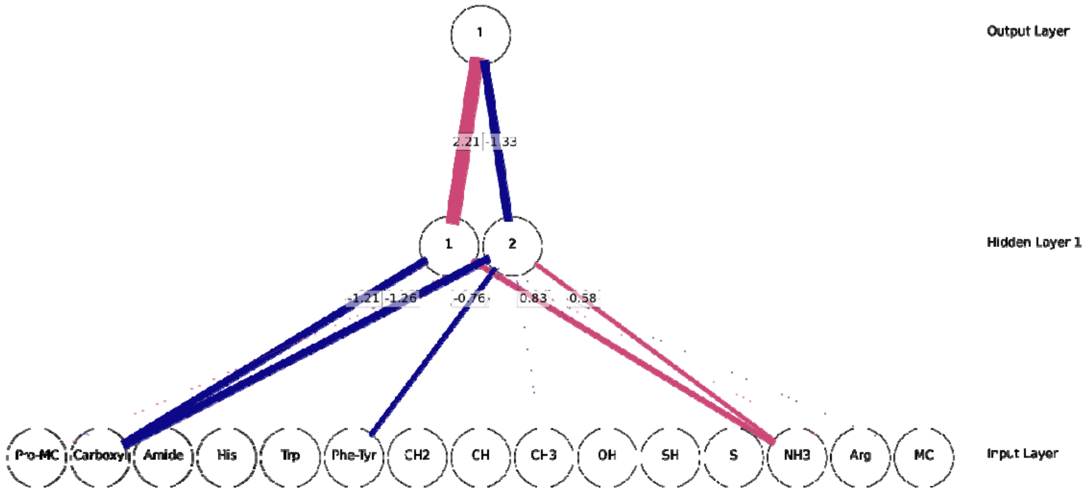

1
Output Layer
2.21
-1.33
1
2
Hidden Layer 1
-1.21
-1.26
-0.76
0.83
0.58
Pro-MC
Carboxyl
Amide
His
Trp
Phe-Tyr
CH2
CH
CH3
OH
SH
S
NH3
Arg
MC
Input Layer

Figure 3. Multilayer perceptron architecture (C-Coord2 based) used in the prediction model for survivin-microarray peptide interactions. The input layer nodes correspond to various features (atom types). These features, listed from left to right, encompass distinct atoms of the proline main chain, carboxyl group, amide group, imidazole ring, indol ring, phenyl ring, methanetriyl-, methylene-, methyl-, hydroxyl-, sulfhydryl group, thioether of methionine, amino group of lysine, guanidino group, and standard main chain atoms. Positive weights are represented by pink lines, and negative weights by blue lines.

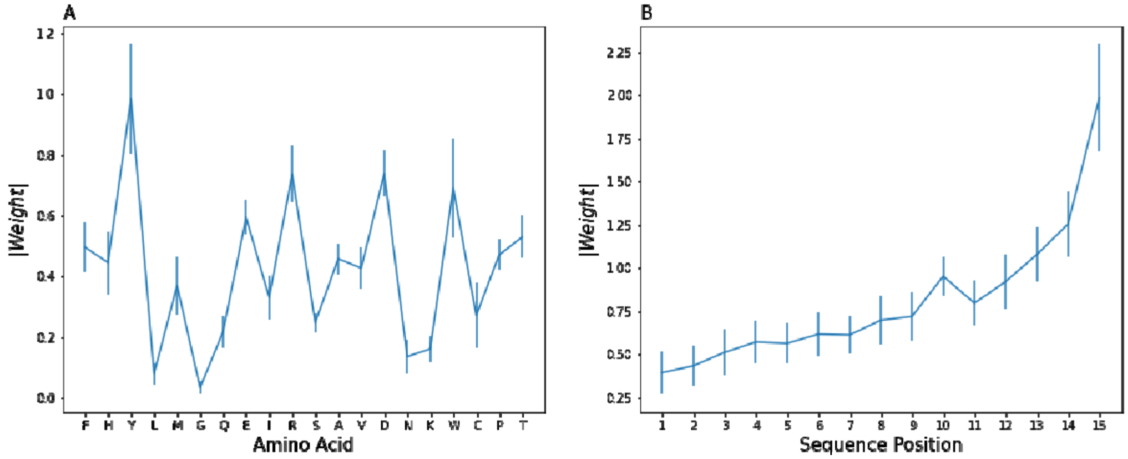


Figure 4. Position specific weights of the dense layer connecting the 2D embedding layer to the output. (A) Absolute value of the embedding quantity associate with every amino acid. (B) Absolute value of weights in the dense layer corresponding to the amino acid position from N to C terminus. The mean and the standard error of the mean are displayed based on 100 randomizations.

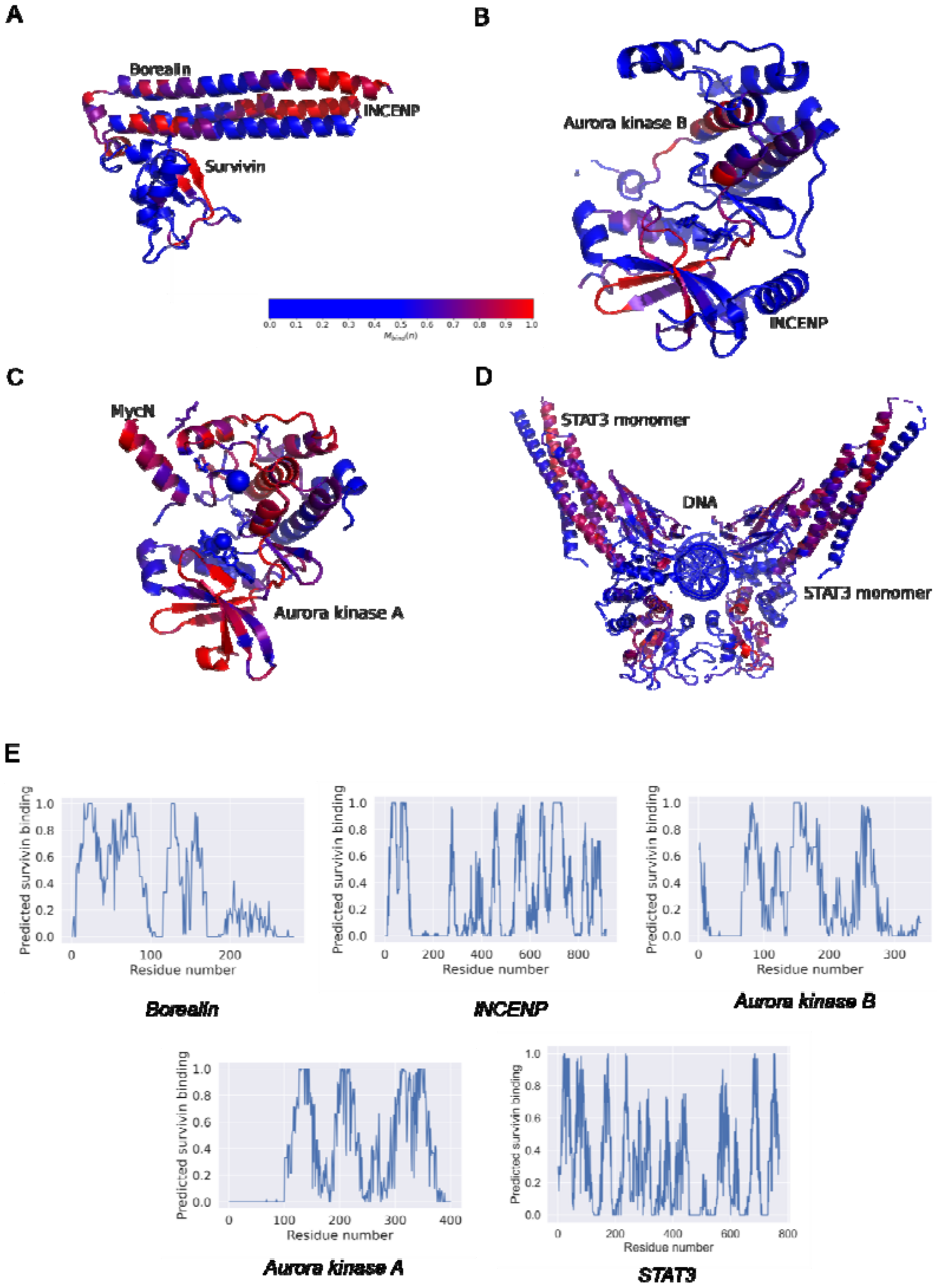
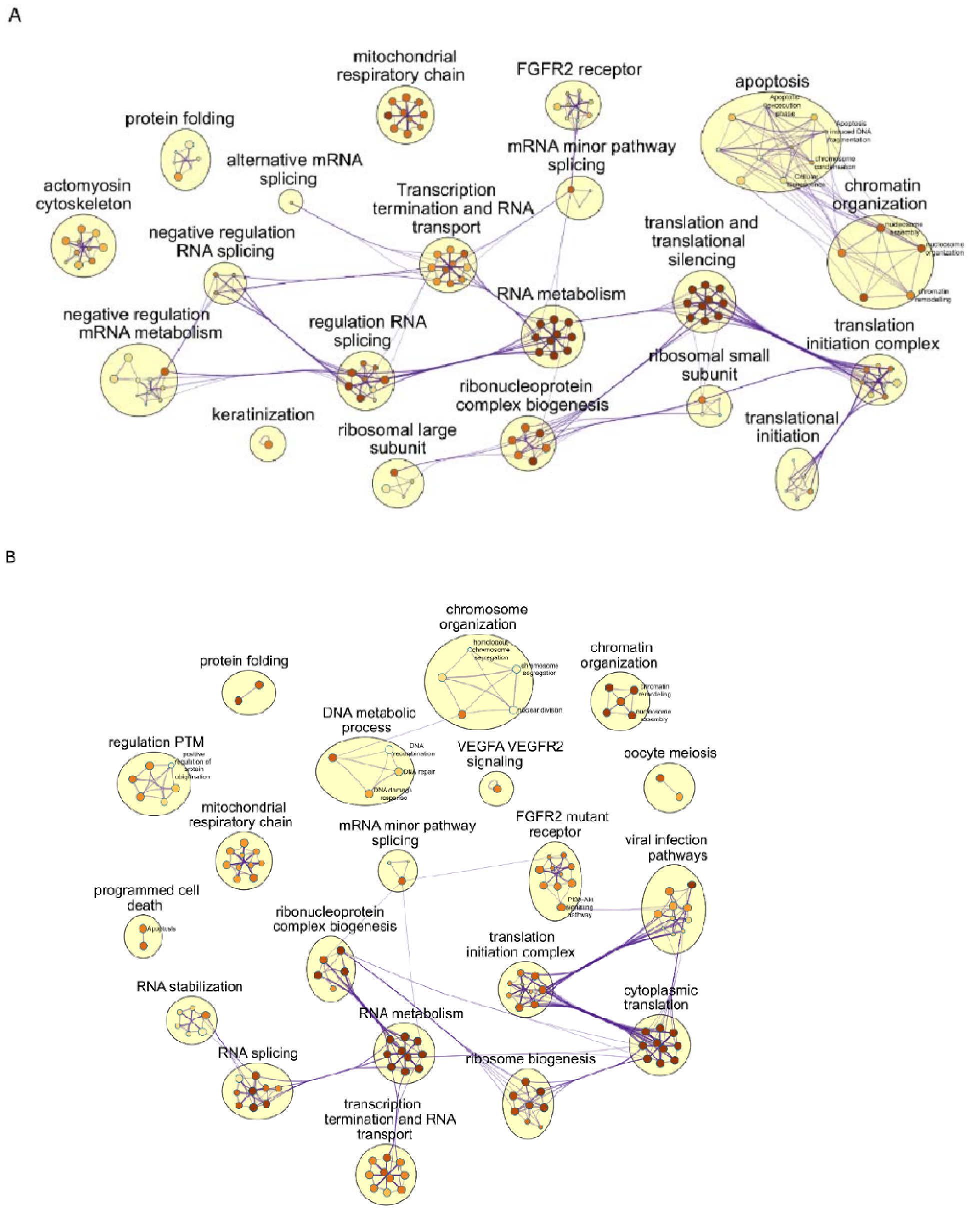
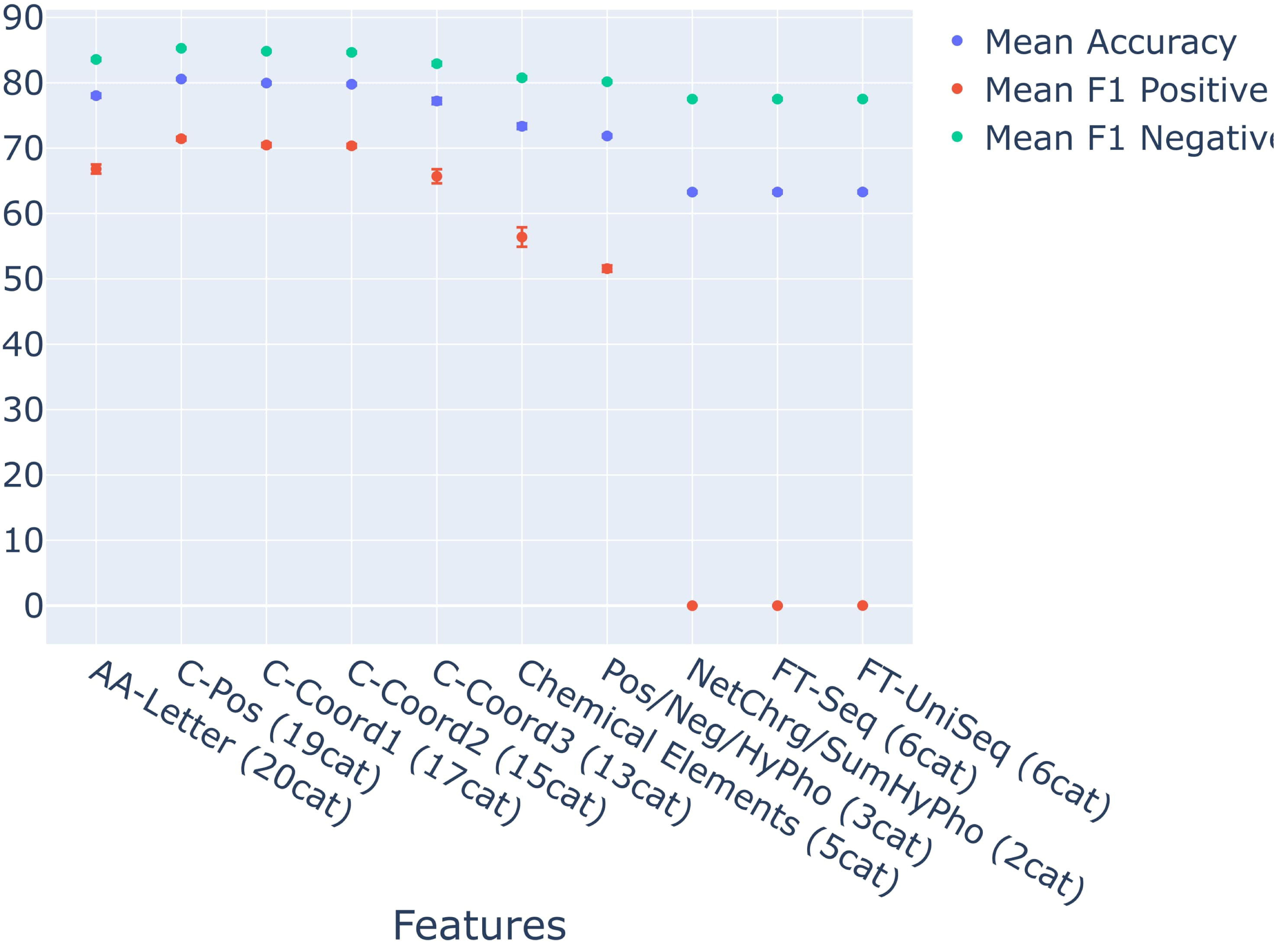


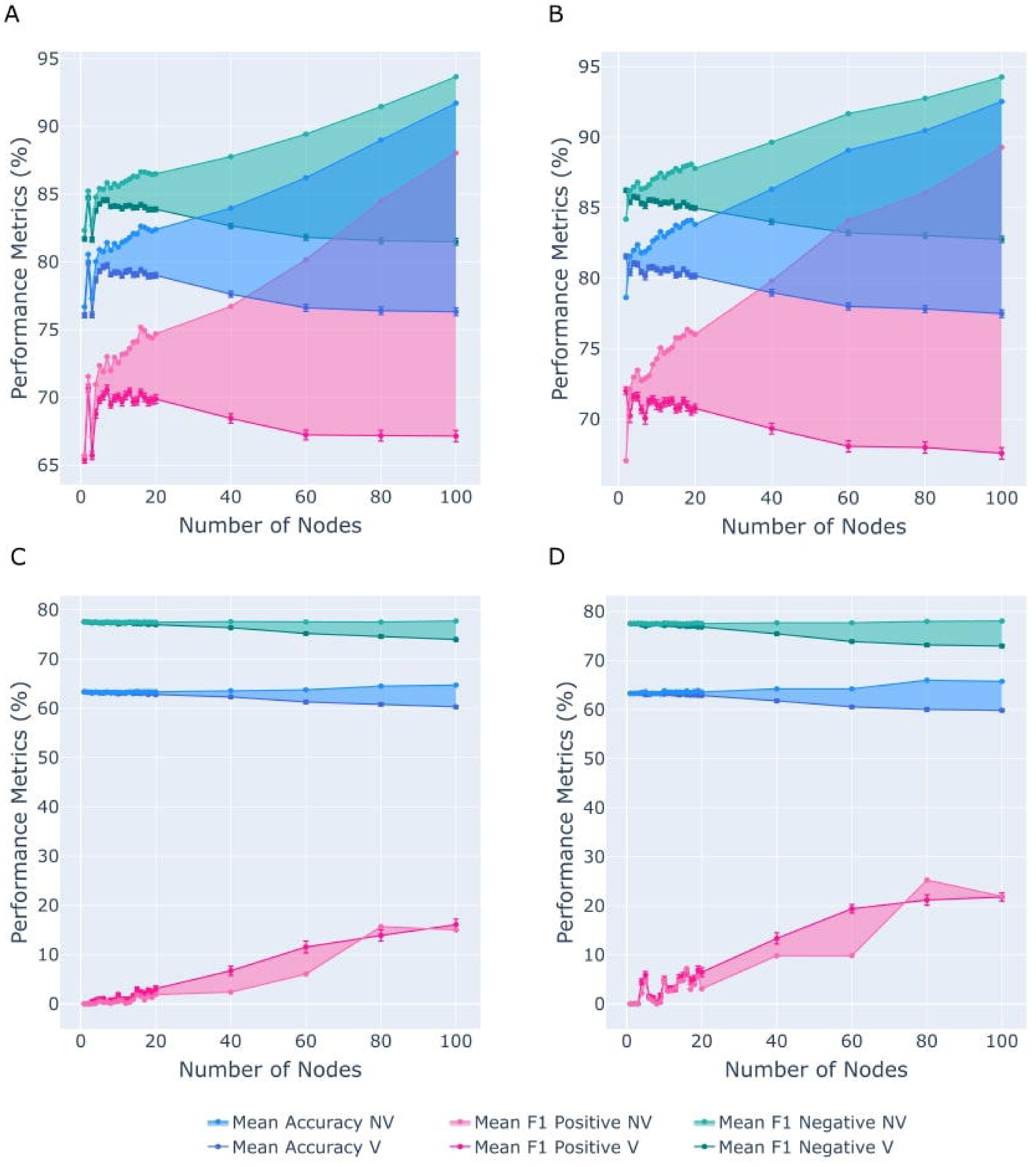
Figure 5. $M_{bind}(n)$ mapped on the ternary structure of known interaction partners of survivin: (A) CPC core consisting of borealin and INCENP (*PDB ID 6YIE* [95]), (B) INCENP and Aurora kinase B (*PDB ID 4AF3* [96]), (C) Aurora kinase A and MycN (*PDB ID 7ZTL* [97]), and (D) STAT3 bound to DNA (*PDB ID 6QHD* [60]). The blue-red color map visualizes the $M_{bind}(n)$ range from 0.5 to 1.0. (E) The one-dimensional plot of $M_{bind}(n)$ the subunits INCENP, Borealin, Aurora kinase A, Aurora kinase B and STAT3 is shown as a function of residues number.



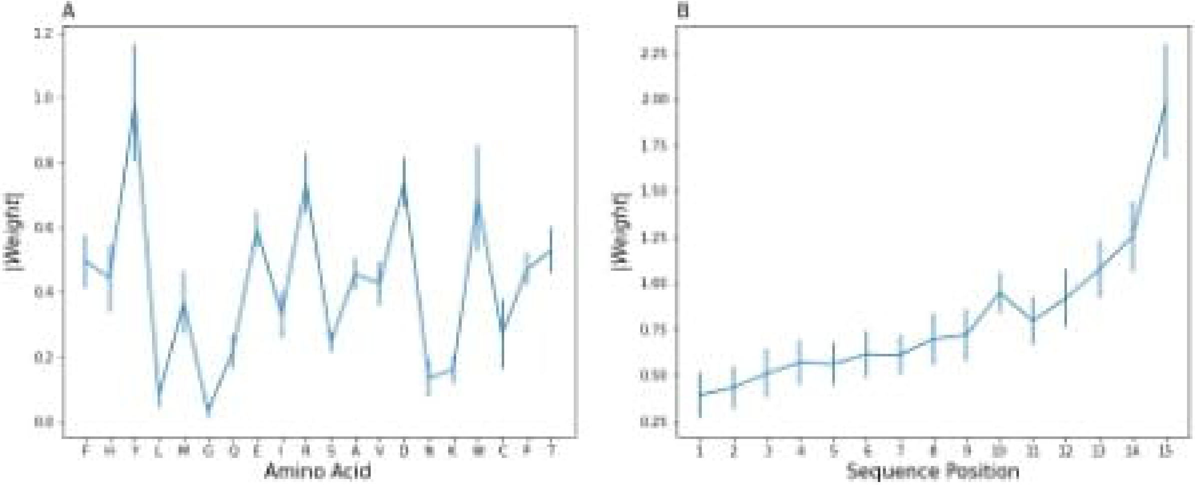
A
mitochondrial respiratory chain
FGFR2 receptor
apoptosis
Apoptotic execution phase
Apoptosis induced DNA fragmentation
chromosome condensation
Cellular Senescence
protein folding
alternative mRNA splicing
mRNA minor pathway splicing
actomyosin cytoskeleton
Transcription termination and RNA transport
chromatin organization
nucleosome assembly
nucleosome organization
chromatin remodeling
negative regulation RNA splicing
translation and translational silencing
RNA metabolism
negative regulation mRNA metabolism
regulation RNA splicing
translation initiation complex
ribosomal small subunit
ribonucleoprotein complex biogenesis
keratinization
translational initiation
ribosomal large subunit
B
chromosome organization
homologous chromosome segregation
chromosome segregation
nuclear division
chromatin organization
chromatin remodeling
nucleosome assembly
protein folding
DNA metabolic process
DNA recombination
DNA repair
DNA damage response
regulation PTM
VEGFA VEGFR2 signaling
oocyte meiosis
mitochondrial respiratory chain
mRNA minor pathway splicing
FGFR2 mutant receptor
PI3K-Akt signaling pathway
viral infection pathways
programmed cell death
Apoptosis
ribonucleoprotein complex biogenesis
translation initiation complex
RNA stabilization
cytoplasmic translation
RNA metabolism
RNA splicing
ribosome biogenesis
transcription termination and RNA transport

Figure 6. Annotation enrichment on human proteins by Metascape. (A) The protein with highest R_{bind} (top 865), excluding proteins shorter than 5 constituting peptides. (B) Metascape enrichment analysis of proteins with simultaneously higher than 10 predicted survivin binding 15mer regions and an R_{bind} higher than 0.5 (top 1724). These networks only include annotations with p-value less than 0.01 and an enrichment factor of at least 1.5.

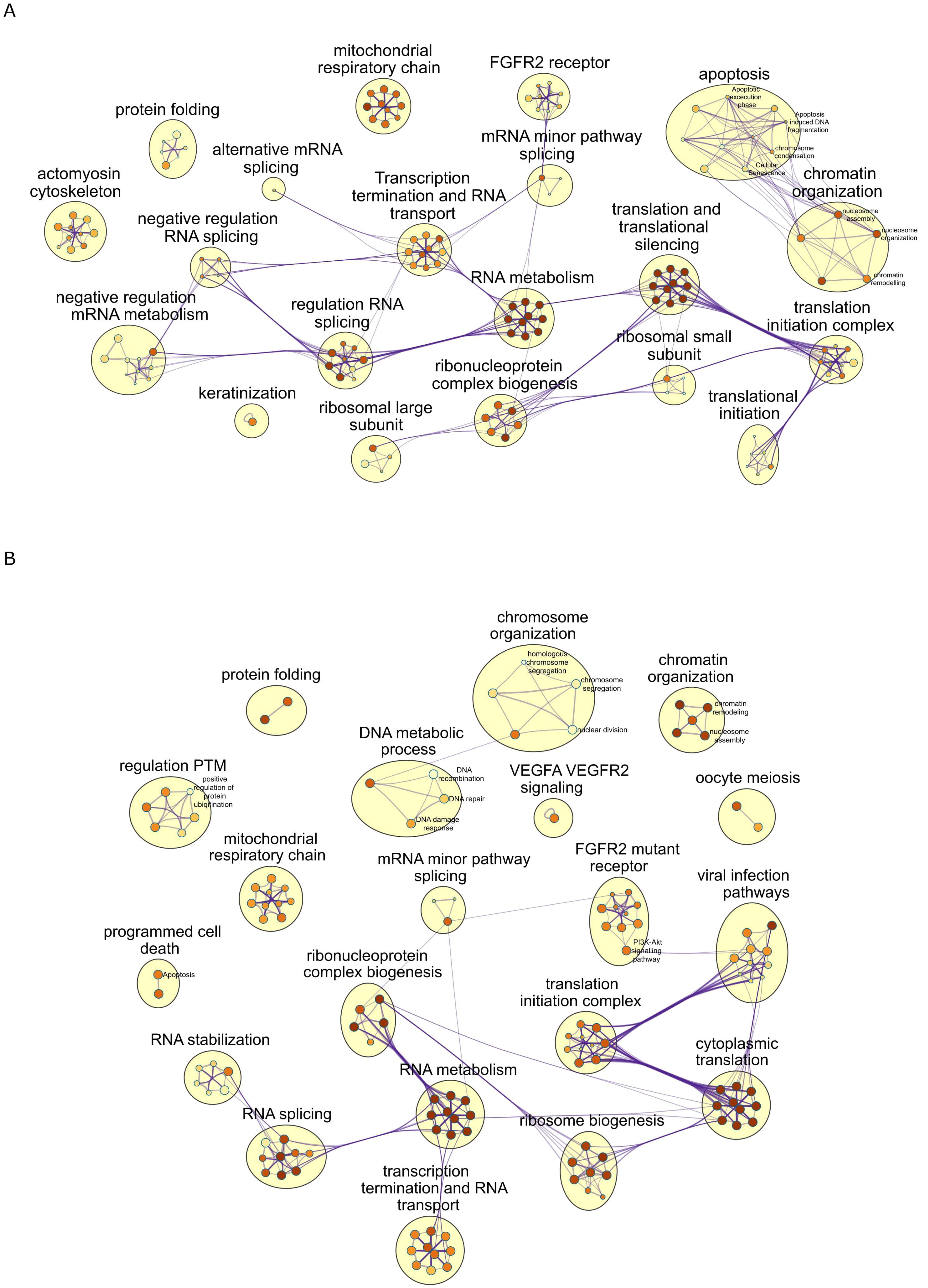

90
80
70
60
50
40
30
20
10
0
Mean Accuracy
Mean F1 Positive
Mean F1 Negativ
AA-Letter (20cat)
C-Pos (19cat)
C-Coord1 (17cat)
C-Coord2 (15cat)
C-Coord3 (13cat)
Chemical Elements (5cat)
Pos/Neg/HyPho (3cat)
NetChrg/SumHyPho (2cat)
FT-Seq (6cat)
FT-UniSeq (6cat)
Features

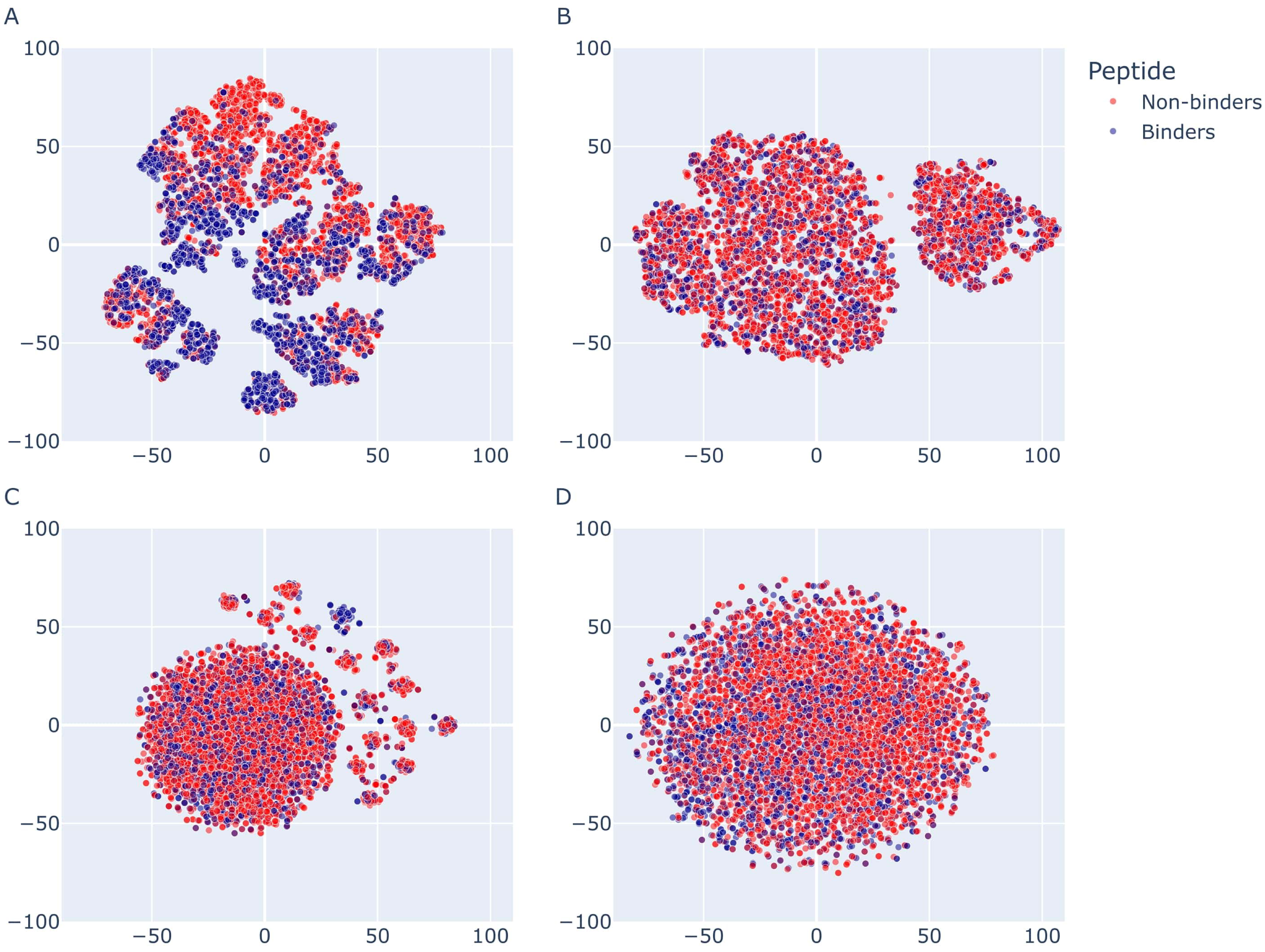

A
B
C
D
Performance Metrics (%)
Number of Nodes
95
90
85
80
75
70
65
0
10
20
30
40
50
60
80
100
Mean Accuracy NV
Mean Accuracy V
Mean F1 Positive NV
Mean F1 Positive V
Mean F1 Negative NV
Mean F1 Negative V

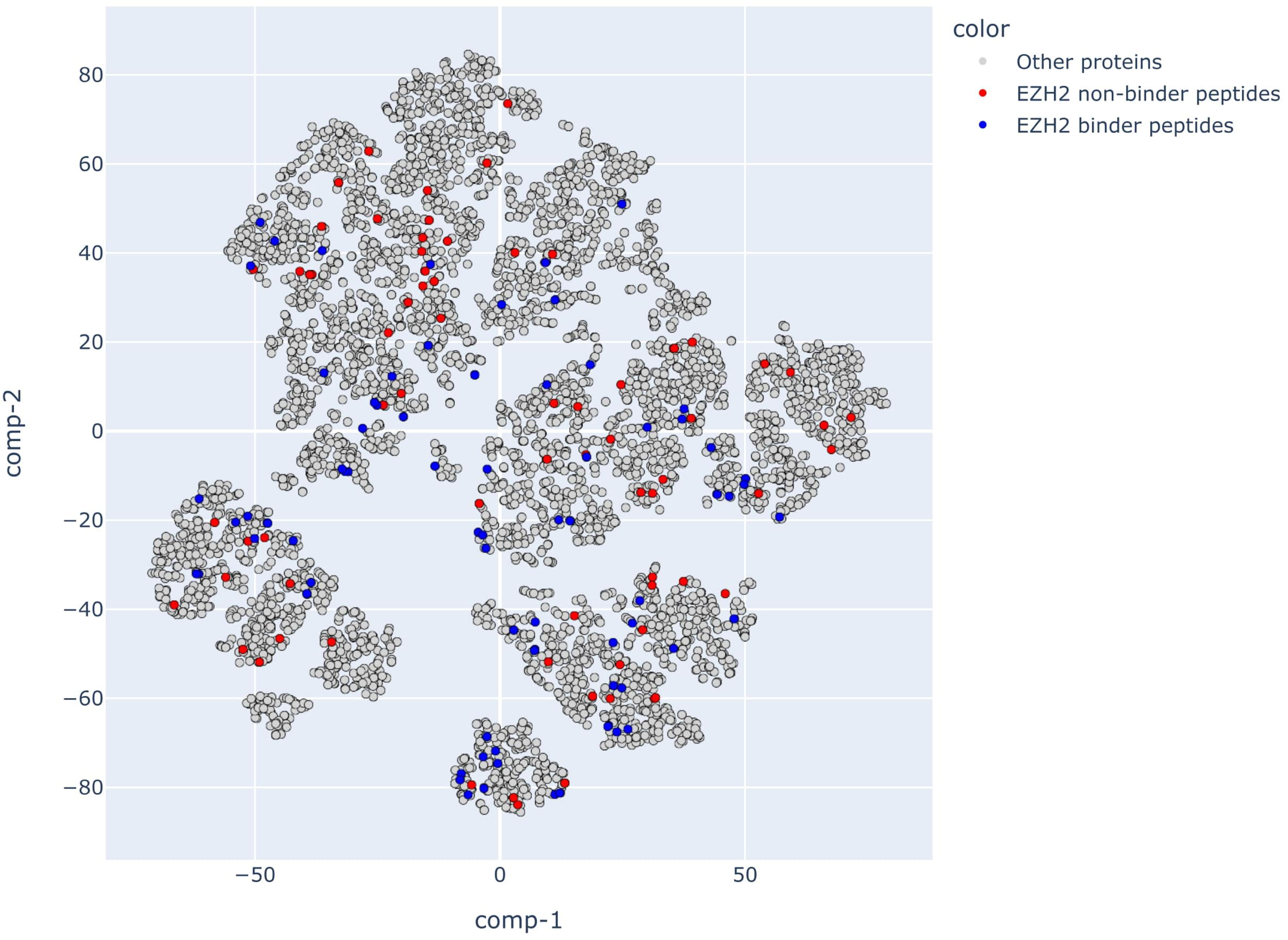
1
Output Layer
2.21
-1.33
1
2
Hidden Layer 1
-1.21
-1.26
-0.76
0.83
0.58
Pro-MC
Carboxyl
Amide
His
Trp
Phe-Tyr
CH2
CH
CH3
OH
SH
S
NH3
Arg
MC
Input Layer


A
1.2
1.0
0.8
0.6
0.4
0.2
0.0
|Weight|
F
H
Y
L
M
G
Q
E
I
R
S
A
V
D
N
K
W
C
P
T
Amino Acid
B
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
|Weight|
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Sequence Position

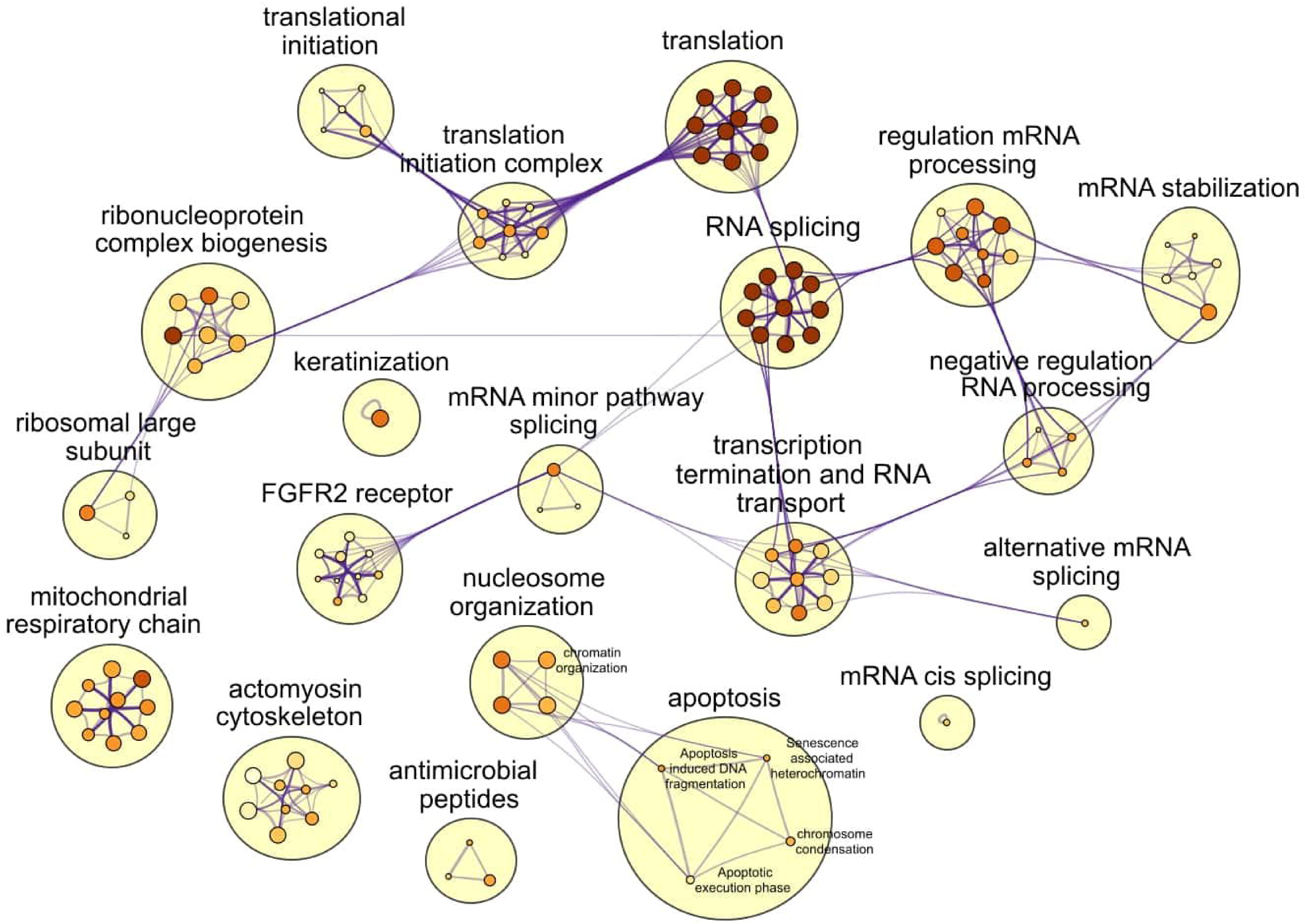
A
Borealin
INCENP
Survivin
B
Aurora kinase B
INCENP
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
$M_{bind}(n)$
C
MycN
Aurora kinase A
D
STAT3 monomer
DNA
STAT3 monomer
E
Predicted survivin binding
Residue number
Borealin
INCENP
Aurora kinase B
Aurora kinase A
STAT3
0.0 0.2 0.4 0.6 0.8 1.0
0 100 200 300 400 600 800


A
mitochondrial respiratory chain
FGFR2 receptor
apoptosis
Apoptotic excecution phase
Apoptosis induced DNA fragmentation
chromosome condensation
Cellular Senescence
protein folding
alternative mRNA splicing
mRNA minor pathway splicing
chromatin organization
nucleosome assembly
nucleosome organization
chromatin remodelling
actomyosin cytoskeleton
Transcription termination and RNA transport
negative regulation RNA splicing
translation and translational silencing
RNA metabolism
negative regulation mRNA metabolism
regulation RNA splicing
translation initiation complex
ribosomal small subunit
ribonucleoprotein complex biogenesis
keratinization
ribosomal large subunit
translational initiation
B
chromosome organization
homologous chromosome segregation
chromosome segregation
nuclear division
protein folding
chromatin organization
chromatin remodeling
nucleosome assembly
DNA metabolic process
DNA recombination
DNA repair
DNA damage response
regulation PTM
positive regulation of protein ubiquitination
VEGFA VEGFR2 signaling
oocyte meiosis
mitochondrial respiratory chain
mRNA minor pathway splicing
FGFR2 mutant receptor
PI3K-Akt signalling pathway
viral infection pathways
programmed cell death
Apoptosis
ribonucleoprotein complex biogenesis
translation initiation complex
RNA stabilization
cytoplasmic translation
RNA metabolism
RNA splicing
ribosome biogenesis
transcription termination and RNA transport


A
B
C
D
100
50
0
−50
−100
−50
0
50
100
Peptide
Non-binders
Binders


color
Other proteins
EZH2 non-binder peptides
EZH2 binder peptides
comp-2
80
60
40
20
0
−20
−40
−60
−80
comp-1
−50
0
50

A
Fluoresence intensity
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
comp-2
80
60
40
20
0
−20
−40
−60
−80
comp-1
−80
−60
−40
−20
0
20
40
60
80
B
Number of carboxyl groups
35
30
25
20
15
10
5
0
comp-2
comp-1


translational initiation
translation
translation initiation complex
regulation mRNA processing
mRNA stabilization
ribonucleoprotein complex biogenesis
RNA splicing
negative regulation RNA processing
keratinization
mRNA minor pathway splicing
ribosomal large subunit
transcription termination and RNA transport
FGFR2 receptor
alternative mRNA splicing
mitochondrial respiratory chain
nucleosome organization
chromatin organization
mRNA cis splicing
apoptosis
actomyosin cytoskeleton
Apoptosis induced DNA fragmentation
Senescence associated heterochromatin
antimicrobial peptides
chromosome condensation
Apoptotic execution phase