PAPER • **OPEN ACCESS**

Deciphering peptide-protein interactions via composition-based prediction: a case study with survivin/BIRC5

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 025081 https://doi.org/10.1088/2632-2153/ad5784

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
28 March 2024

REVISED
20 May 2024

ACCEPTED FOR PUBLICATION
12 June 2024

PUBLISHED
28 June 2024

PAPER

Deciphering peptide-protein interactions via composition-based prediction: a case study with survivin/BIRC5

Atsarina Larasati Anindya[1], Torbjörn Nur Olsson[1], Maja Jensen[1], Maria-Jose Garcia-Bonete[1], Sally P Wheatley[2], Maria I Bokarewa[3,4], Stefano A Mezzasalma[5,6] and Gergely Katona[1,*]

[1] Department of Chemistry and Molecular Biology, Faculty of Science, University of Gothenburg, Box 462, Gothenburg 40530, Sweden
[2] School of Life Sciences, Faculty of Medicine & Health Sciences, University of Nottingham, Nottingham NG7 2UH, United Kingdom
[3] Department of Rheumatology and Inflammation Research, Institute of Medicine, University of Gothenburg, Box 480, 40530 Gothenburg, Sweden
[4] Rheumatology Clinic, Sahlgrenska University Hospital, Gröna stråket 16, 41346 Gothenburg, Sweden
[5] Division of Materials Physics, Laboratory for Optics and Optical Thin Films, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[6] Institute for advanced Neutron and X-ray Science (LINXS), Lund University, IDEON Building: Delta 5, Scheelevägen 19, 22370 Lund, Sweden
[*] Author to whom any correspondence should be addressed.

E-mail: gergely.katona@gu.se

Keywords: composition-based prediction, multilayer perceptrons, classification, feature engineering, protein interactions, survivin, phase transition

Supplementary material for this article is available online

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 confirms the known cellular locations of survivin and its interaction partners, but also introduces novel insights into biological functionality. It becomes evident that electrostatic- and primary structure-based descriptions fall short in predictive power, leading us to speculate that protein interactions are orchestrated by the collective dynamics of functional groups.

1. 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 is not completely understood, and their deficiencies are often related with disease progression, such as 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 and van der Waals bonding 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 the primary amino acid sequence as the sole key to determine protein structure, function, and, by extension, how it recognize interaction partners, often oversimplifying 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.

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] research aimed at identifying protein binding partners and determining their binding affinities keeps rising. However, our understanding of these interactions is not keeping pace with this knowledge accumulation, and are generally explained by detailed short-range physical interactions. Black box machine learning of cognate partners runs the risk of merely transforming existing data into yet another (overfit) model representation.

Peptide microarrays, which have their roots in the method of antigen-antibody reaction determination [11], offer an effective tool in studying PPI due to its high-throughput nature. In peptide microarrays, the target protein is often divided into short peptide segments and orderly arranged on a solid surface, before being exposed to fluorescence-labelled ligand proteins or antibody in solution. The detected fluorescence intensity at each peptide spot correlates with the binding strength of that specific 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]. Survivin has only one characterized complex with other proteins (or more precisely, a complex with large fragments of other proteins). This complex is a part of the Chromosomal Passenger Complex (CPC) and includes fragments of borealin and INCENP [14]. The scarcity of structural information makes it is difficult to generalize the binding mechanism of survivin. For other known partners of survivin (for example aurora kinase B), one can only speculate about the binding mechanisms, as they may involve multiple ways of attachment, induced fit, or other types of interactions that are difficult to structurally characterize.

The rapid acquisition of PPI data from peptide microarray analyses as become a valuable factor to formulate patterns to identify and understand the driving forces behind. A major challenge to deriving prediction tools comes from inconsistencies in defining the 'cut-off' distance of non-covalent interactions [15], which are modelled based on amino acid physicochemical and biochemical properties; AAindex is a database containing various indices representing these properties [16].

Protein association is thought to start with encounter complex formation mediated by long-range electrostatic forces resulting in a favourable (exothermic) enthalpy change [17], which can be predicted by statistical coupling analysis [18]. Encounter complex formation implies an entropic loss from reduced conformational, rotational, and translational degrees of freedom leading to structural rearrangements and eventual desolvation of the interacting surfaces [19–21].

Certain characteristics are connected to protein interfaces: stable interactions have larger buried surface areas and more hydrophobic surfaces than transient ones [22]. Stable interactions also have more conserved, tyrosine and arginine residues, where most of the binding energy is concentrated [23, 24]. 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 [25, 26]. Side-chain interactions are also prominent in PPIs, whereas they only contribute to 36% of the internal bonds stabilizing a protein [27]. There is also a significant compositional difference between transient and permanent interactions (i.e. in quaternary structures or complexes) [28].

Despite knowing conserved residues forming PPIs, binding affinity prediction based solely on amino acid sequence have been somewhat ambiguous, as diverse sequences can have similar binding affinities [29]. 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 [30], AlphaPullDown [31], and RoseTTAFold [32]. These tools have limitations on proteins with no solved homologous structure, a common problem with intrinsically 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 [33], attention network, and long short-term memory (LSTM) models [32, 34–36]. 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 numerically encoded sequences, creating a unique pattern from residue frequency for each protein [37]. This study used a training dataset of single point mutations of epoxide hydrolase from *Aspergillus niger* and achieved 81% accuracy in validation dataset to predict epistatic interactions [37]. On the other hand, binding site predictions from the joint analysis of solvation potentials, amino acid composition, conservation, electrostatics, and hydrophobicity seems to yield promising results [38–43].

When in dynamic equilibrium chemical reactions, 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 the concentrations of the interacting partners and their binding affinity. The scarcity of explicitly repulsive PPI observations [44] points to a prevalence of attractive interactions among proteins or to an observational bias focusing on attractive interactions.

Considering the macromolecular size, complexity and diversity of 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 model target protein and its binding partners to illustrate how protein composition is a predictive factor for biological interactions.

Our work represents an exploration of the concept of 'composition,' serving as an initial approximation. This entails investigating which features contribute to a more useful description. Here, we examined amino acids, groups of amino acids (based on their charge and hydrophobicity), chemical elements, and functional groups within amino acids as features. Our definitions are sometimes ambiguous and not always rigorously following ideal principles. For example, when considering an imidazole ring as a feature, we acknowledge the inherent ambiguity that this functional group may be protonated or non-protonated. It has the potential to exist in both states, and we allow the training process to incorporate this ambiguity in decision-making, possibly by also considering positively and negatively charged residues within the same peptide. We anticipate that improved definitions of abstract elements underlying encodings will be developed in the future. In the absence of a more rigorous definition of 'composition', we establish the exact encodings explored in this study.

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 and the associated collective dynamics of features in determining protein function.

2. Results and discussion

2.1. PPI classification based on atom type composition and sequence

This study used a peptide microarray dataset containing peptides from 36 known and potential binding partners [13], including PRC2 [45] and CPC subunits. The peptides were then described by counting various chemical groups they contain as features, encompassing the standard main chain atoms (MC), methylene group from glycine $C\alpha$ atom (CA–Gly), side chain atoms (carboxyl, amide, phenolic hydroxyl, sulfhydryl groups, imidazole and indole rings) among others. A comprehensive list and explanation of feature development can be found in the Methods section, with cues to tables S1–S10. Briefly, the C–Coord model series categorizes aliphatic carbon atoms according to coordinating hydrogen atoms. C–Coord1 (table S1), C–Coord2 (table S2) and C–Coord3 (table S3) models contain 17, 15 and 13 features, respectively. The 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 the chemical formula model counts hydrogen, carbon, nitrogen, oxygen, and sulfur atoms (table S6). Electrostatic and hydrophobic properties are represented in Pos/Neg/HyPho model as binary properties of individual amino acids (table S7) and NetChrg/SumHyPho models representing the properties of the entire peptide (table S8) [46]. We also compared our approach with Fourier Transform patterns from AAindex numerical value representation of amino acids (RACS820104, Seq, table S9) [37]. The real part of the complex term yields a peptide spectrum (FT-Seq). We also used equally spaced intervals between values to encode the amino acids. (UniSeq, table S10). These features were combined with different machine learning architectures as shown in figure 1.

Figure 2 presents the performance indicators of various approaches, using a very simple multilayer perceptron (MLP) with two intermediate layer perceptrons. As previously reported, non-binders are slightly more common compared to binders [13]. 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 MLP model. We speculate that the AA-Letter feature set 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

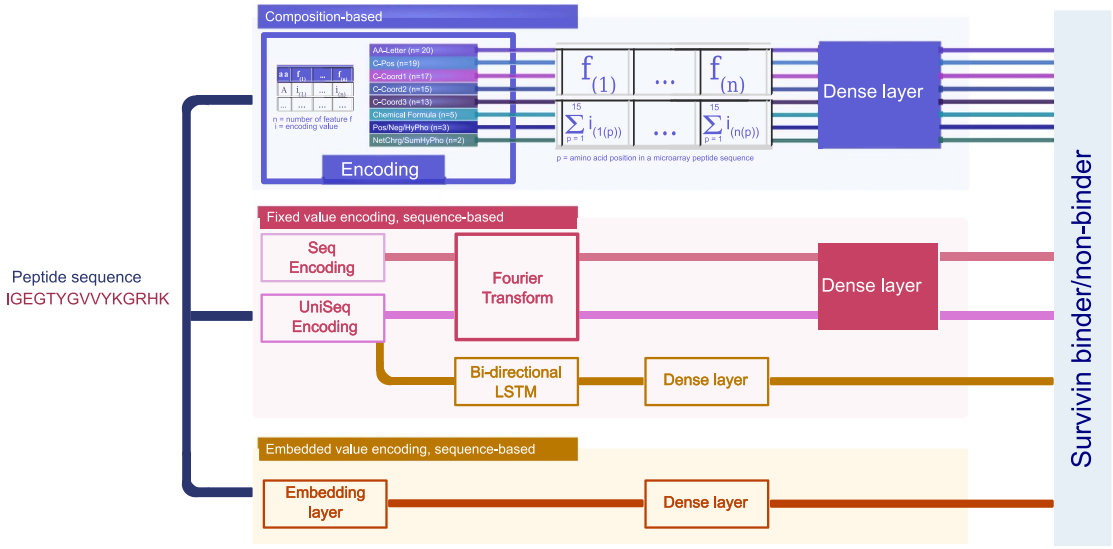

Figure 1. Features and machine learning architectures explored for predicting survivin binding.

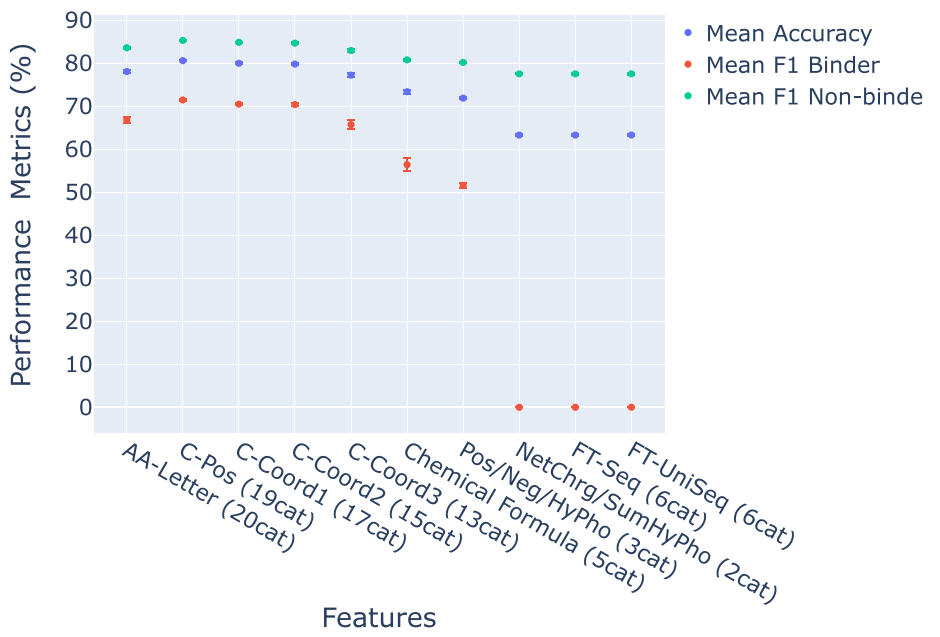

Figure 2. Performance evaluation of a MLP classification with a single intermediate layer consisting of two perceptrons. Different sets of features describing the peptides on the peptide microarray were compared. Training was carried out using 50% of the data, repeated 100 times with varying randomly selected subsets, and testing was performed on the remaining half of the dataset.

C–Pos, C–Coord1, and C–Coord2 notably align better with intuitive understanding of amino acid features with limited data.

Simplification resulted in similar performances when the 15-feature model (C–Coord2) was compared with a 17-feature model (C–Coord1). 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 slightly better performance of the C–Pos model implies the importance 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

Table 1. Performance indicators of the survivin binding peptide predictions. B and NB indicates binding and non-binding peptides. Two intermediate layer perceptrons were used. Uncertainty is represented by the standard error of the mean.

	Precision (%)		Recall (%)		F1-score (%)		
	NB	B	NB	B	NB	B	Accuracy (%)
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 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.

		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	

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.

2.2. The impact of network complexity on the prediction accuracy

Next, we investigated how the complexity of the MLP impacts prediction performance with different feature sets. To maintain simplicity, we employed a single intermediate layer of perceptrons and explored changing the number of hidden layer perceptrons, ranging from the default of 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 the 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 Binder score (figures 3(A) and (B)). Here 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 an appropriate balance between accuracy and generalization. Notably, a substantial non-monotonic variation in F1 score performance exist, 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 model with a single perceptron in the intermediate layer dismisses any overly simplistic contribution of diverse amino acids and atom types. Instead, it underscores the importance of considering their (anti-)correlated presence for precise binder detection.

As their F1 Binder score reaches a plateau around 0.20 with 100 intermediate layer perceptrons, sequence-based methods evidently require a more complex network to enhance binder detection (figures 3(C) and (D)). As the model complexity increases, the gap between the performance with and without cross-validation as model complexity widens suggesting that the best-performing model may lack sufficient generalizability.

We can only speculate about the underperformance of the FT-Seq model, even compared to the FT-UniSeq models (see figure 3). The FT-Seq model has an F1 Binder score below 20%, whereas FT-UniSeq achieves slightly higher than 20% when the model contains 100 intermediate layer perceptrons. First, Fourier transformation of a numerical sequence captures fundamentally different information than composition. The periodicity of the sequence is the primary factor that is detected, while the specific amino acids contributing to the repeating pattern are of secondary importance. If phase information is retained, no information is lost compared to the original numerical sequence. In principle, a sufficiently complex architecture could approximate the function 'reverse Fourier transform' along with any interpretation of the sequence given millions of data points provided. However, such an approximation is not feasible with

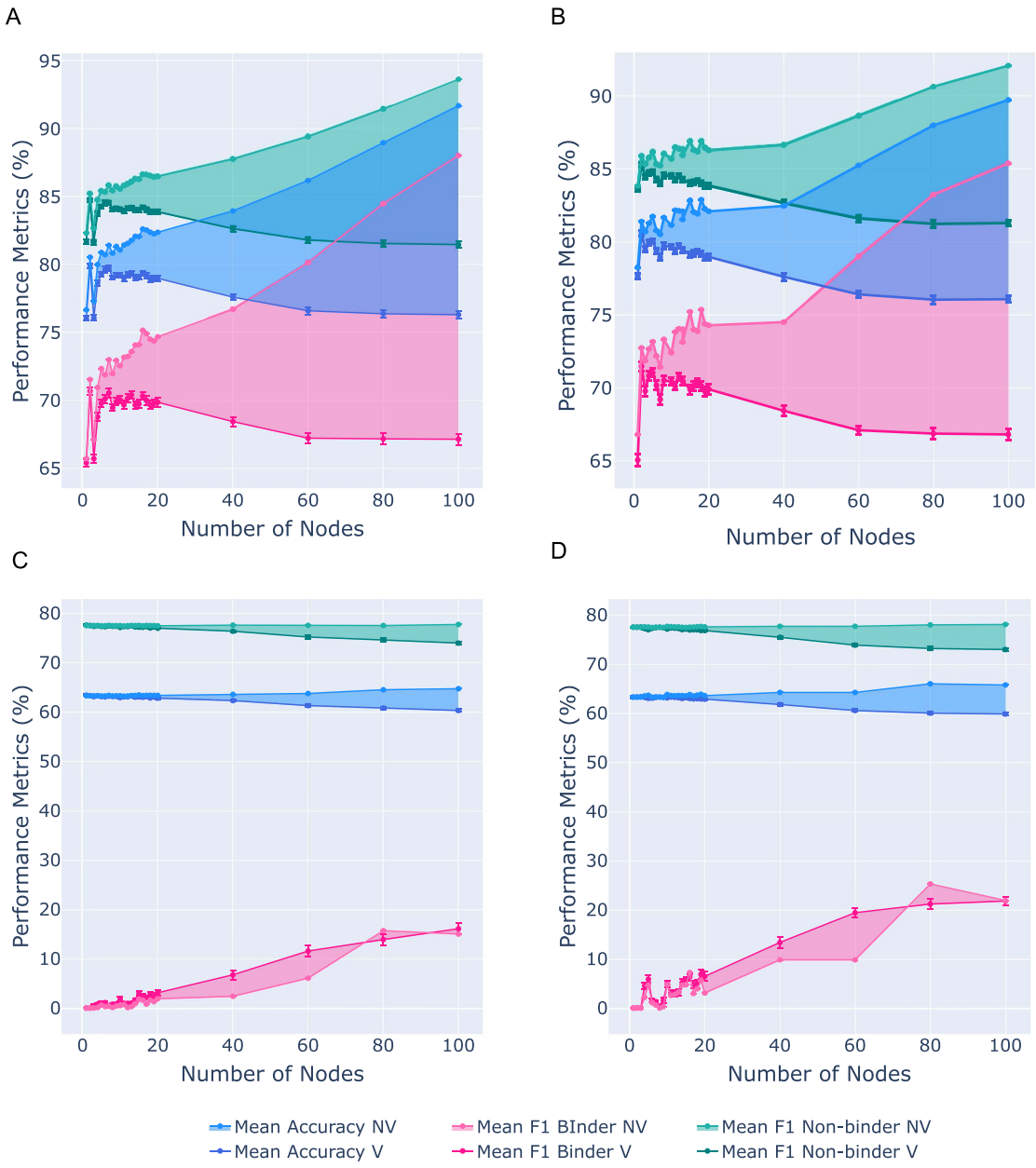

Figure 3. Trends of performance metrics across a range of number of nodes (1–100) in different prediction models using: (A). 15 atomic composition features (C–Coord2), (B). 19 atomic composition features (C–Pos), (C). Fourier Transform of non-uniform value representation of amino acids (FT-Seq), and (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 Binder, in cyan), and mean F1 values of 'non-binder' peptides (F1 Non-binder, 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.

shallow networks even if sufficient data are provided. Second, the RACS820104 index (Seq) assigns the exact same numerical value (1.31) to serine and tyrosine, making these amino acids indistinguishable. In the UniSeq encoding, some amino acids with similar compositions are grouped together (for example, *D* and *E* contain carboxyl groups, and *S*, *T*, and *Y* contain hydroxyl groups). However, in the RACS820104 index, such clustering is less apparent. Improving the order of amino acids by focusing specifically on compositional similarity is possible, but it is important to note that a one-dimensional representation cannot fully capture the compositional diversity of amino acids.

2.3. Intuitive interpretation of the peptide preferences of survivin

Due to the simplicity of the composition-specific MLPs, the optimized architecture can be interpreted directly and visualized in figure 4. The C–Coord2 model (figure 4(A)) highlights two distinct pathways that lead to survivin binding decisions for a given peptide. First, a pathway is created when many carboxyl groups and few amino groups are present which inhibit and activate the inhibitory perceptron 2, respectively.

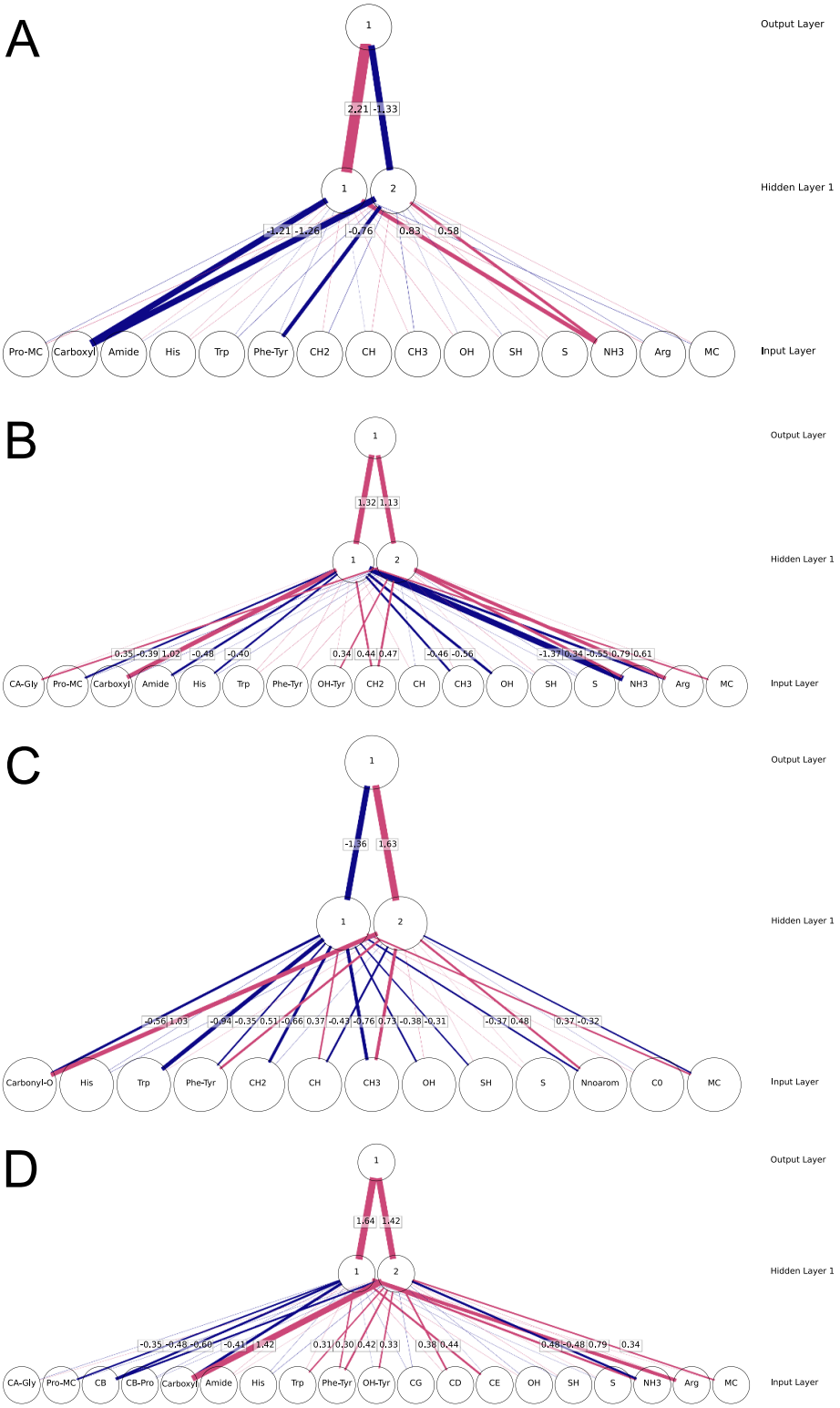

Figure 4. MLP architecture used in the prediction model for survivin-microarray peptide interactions. (A). C–Coord2, (B). C–Coord1, (C). C–Coord3, (D). C–Pos models with two intermediate layer nodes. The input layer nodes correspond to various features (atom types). Positive weights are represented by pink lines, and negative weights by blue lines. The MLPs were illustrated with the help of VisualizeNN library (https://github.com/jzliu-100/visualize-neural-network).

Second, amino groups activate the stimulating perceptron 1 and if the number of carboxyl groups is low, survivin binding is predicted. Interestingly, phenyl rings are supportive factors when paired with carboxyl groups as their presence inhibits perceptron 2, which in turn, also supports survivin binding decisions. All atom types contribute to the prediction, although to improve clarity, the weighted edges of features below and above a model specific threshold are not displayed in figure 4. Compositional features are inherently non-independent and tend to be negatively correlated since adding an amino acid necessarily implies the

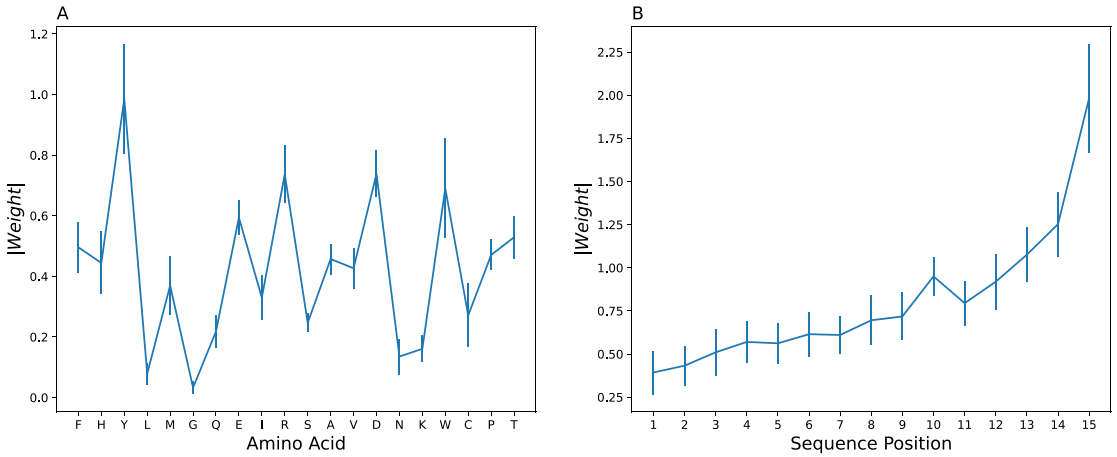

Figure 5. Position specific weights of the dense layer connecting the 1D embedding layer to the output. (A) Absolute value of the embedding quantity associate with every amino acid. (B) Absolute value of weights in the dense layer corresponding to the amino acid position from N to C terminus. The mean and the standard error of the mean are displayed based on 100 randomizations.

removal of another as the length of the peptide is fixed. Therefore, the adjustment of the weight of a single feature potentially impacts the contribution of all others. The sign of the weights can also vary depending on different randomization starting points. The different number of features, the sign variation of the weights and feature correlation makes the direct comparison between model edges challenging. 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.

2.4. The contribution of the peptide primary structure

To improve the sequence-based predictions, we employed bidirectional LSTM [47] networks and embedding techniques. The most effective architecture had two LSTM layers, with 90 neurons in the first layer and 70 neurons in the second (table 2). The model comprised 243 221 trainable parameters and demonstrated an accuracy of 85.7 $\pm$ 0.1% (cross-validated) and 89.7 $\pm$ 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 [48] due to peptide overlaps may also inflate the cross-validation performance, which can force the model to detect exact sequences instead of real physicochemical patterns. When using only every 3rd peptide the accuracy drops to 79.2 $\pm$0.2% (cross-validated) and 89.7 $\pm$0.2% (non-validated).

This 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 complex 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 model) in the composition-based network and 243 221 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 5, where the connection weights are extracted as a function of position. Despite having only 35 parameters, the accuracy is 79.7 $\pm$ 0.0% (cross-validated) and 80.1 $\pm$ 0.2% (non-validated), comparable to composition-based methods. Figure 5(A) shows that tyrosines, aspartates, tryptophanes and arginines contribute the most to the prediction, in partial agreement with the composition-based model (figure 4).

The gradual increase in weights from N- to C-terminus (figure 5(B)) 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 [49]. The relative importance of the C-terminus is unexpected because the surface exposed

N-terminus intuitively has closer contact with survivin in solution. Such a position specific pattern is unlikely to represent a universal mechanism guiding survivin interactions *in vivo*, but these inherent experimental biases need to be addressed in the composition-based analysis.

2.5. 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) [50], 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 S1(A)) with distinct separation between binders and non-binders, while only two clusters observed using the FT-Seq representation (figure S1(B)) with binder and non-binder peptides mostly overlap with each other. Varying the perplexity value between 5.0 and 50.0 did not have a significant impact on the data point distribution [51]. Peptides from a single protein chain do not necessarily share similar compositions and they can be present in more than five recognizable clusters, exemplified by protein EZH2 in figure S2. This suggests 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'.

2.6. Analysis of a known survivin complex and structurally characterized interaction partners

Using the composition-based model, we investigated the impact of point mutations within the context of the surrounding residues by creating a metric $M_{\text{bind}}(n)$, where 'n' represents a residue position along the sequence based on the number of point mutations that result in peptides containing the given residue position being predicted to bind to survivin. $M_{\text{bind}}(n)$ of one indicates a region always predicted to bind survivin even if the position is mutated to any other amino acid, whereas zero indicates that no point mutation of the site can convert the region to a survivin binding region.

Through this approach, the wild-type amino acid is not given preferential treatment because it is assumed that the choice of a particular amino acid at a given position is often (not always) arbitrary, and alternative amino acids may perform equally well. This is because functionally neutral point mutations occur frequently during evolution [52]. The composition-based view explains the functional robustness against individual point mutations. Generally, composition is resistant to change and evolves slowly over time. This is because it requires multiple point mutations with a clear drift direction before any substantial change in composition becomes apparent.

Figure 6 illustrates $M_{\text{bind}}(n)$ as a function of residue number, and it is evident that $M_{\text{bind}}(n)$ does not change smoothly along the sequence. The boundary between survivin binding and non-binding regions tend to be sharply defined and robust binding regions are focused to shorter segments. Discontinuous jumps are also prevalent. If an amino acid residue in the binding region has a significantly lower $M_{\text{bind}}(n)$ than its surroundings, it suggests that the wild-type residue at position n is crucial for maintaining the affinity of the surrounding region. This may lead to a loss of function (binding) with point mutations at that specific position. Conversely, if an amino acid residue in a non-binding region has a $M_{\text{bind}}(m)$ significantly higher than its surroundings, it indicates that the wild-type residue at position m is unsuitable for survivin binding within the given context and that position m may be associated with gain-of-function mutations resulting in survivin binding.

Apart from itself, of all identified survivin interaction partners, only the ternary complex of CPC involving survivin, borealin, and INCENP [14] has been structurally characterized using x-ray crystallography (figure 6(A)). 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 region of high-affinity (59–76), which is observed 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 many mitotic substrates, including the CPC subunits (figure 6(B)). Aurora B shares 71% identity in its catalytic domain with its family member Aurora A, whereas the identity is 57% across the entire protein. The C-termini of the aurora kinases are conserved, particularly from residues 383–387. It is possible to swap the C-terminus of Aurora A with that of Aurora B to restore the lost function of Aurora B [57]. Moreover the point mutation G198N can transform Aurora kinase A into B in terms of their specificity to interact INCENP and survivin [58]. The high $M_{\text{bind}}(198)$ of

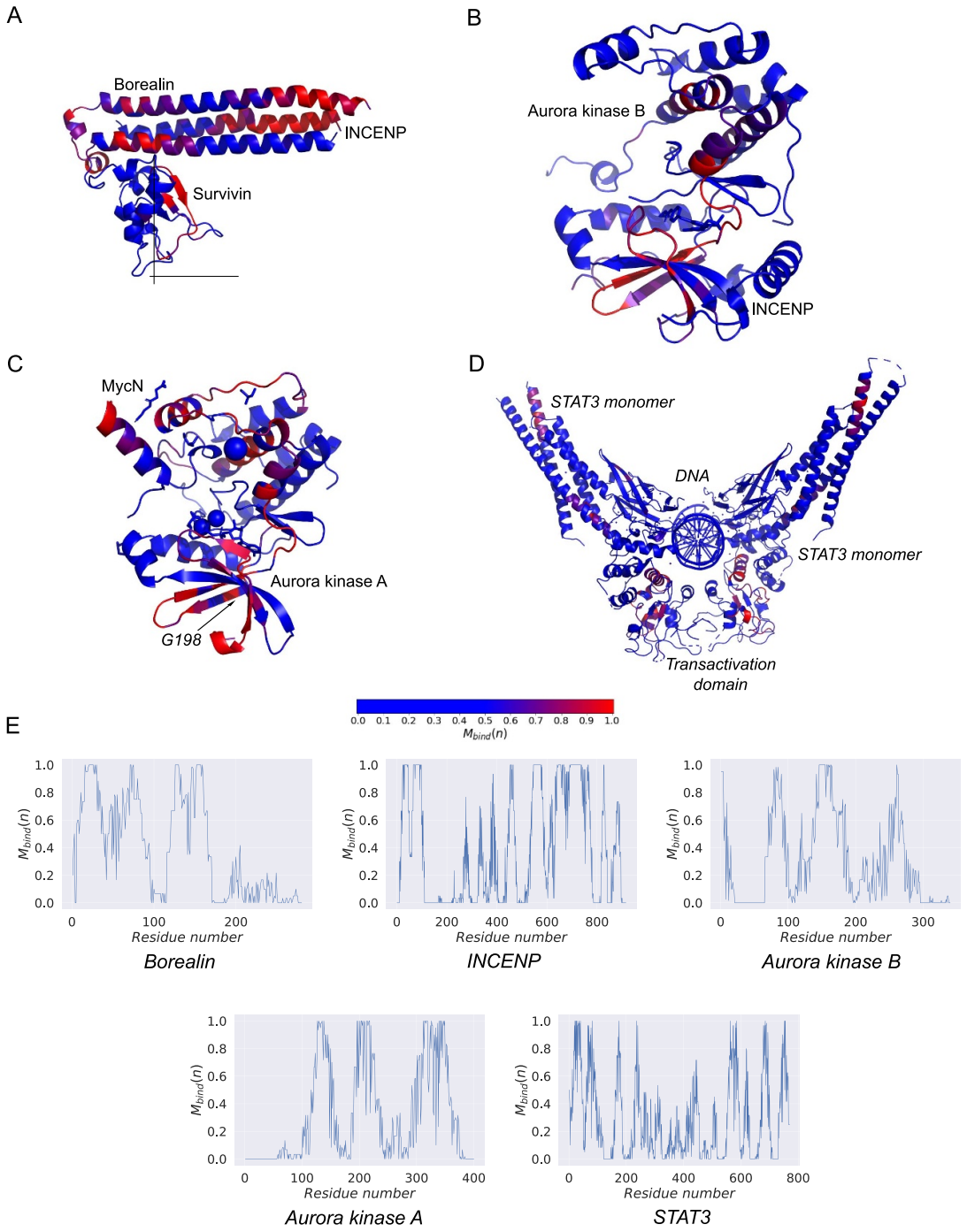

Figure 6. $M_{\text{bind}}(n)$ mapped on the ternary structure of known interaction partners of survivin: (A) CPC core consisting of borealin and INCENP (*PDB ID 6YIE* [53]), (B) INCENP and Aurora kinase B (*PDB ID 4AF3* [54]), (C) Aurora kinase A and MycN (*PDB ID 7ZTL* [55]), and (D) STAT3 bound to DNA (*PDB ID 6QHD* [56]). The blue-red color map visualizes the $M_{\text{bind}}(n)$ range from 0.5 to 1.0. (E) The one-dimensional plot of $M_{\text{bind}}(n)$ the subunits INCENP, Borealin, Aurora kinase A, Aurora kinase B, and STAT3 is shown as a function of residues number.

Aurora A (figure S4), flanked by substantially lower $M_{\text{bind}}(197)$ and $M_{\text{bind}}(199)$ from tyrosine residues (ILRL**Y**G**Y**FHDA), suggests that replacing glycine with many other residues would enhance survivin binding in that region, while removing even one tyrosine would be most likely to reduce survivin affinity.

STAT3 is likely another survivin interacting partner [59], which exhibits no extensive regions with high survivin binding, seeing almost no $M_{\text{bind}}(n)$ value reaching 1.0 (figure 6(D)). However, both the N– and C-terminal regions generally show elevated $M_{\text{bind}}(n)$, with the latter encompass the transactivation (responsible for survivin interaction [59]) and SH2 domains near the DNA [60], which is localized by $M_{\text{bind}}(n)$ with precision. The first helix of the coiled coil domain also displays heightened $M_{\text{bind}}(n)$ [60].

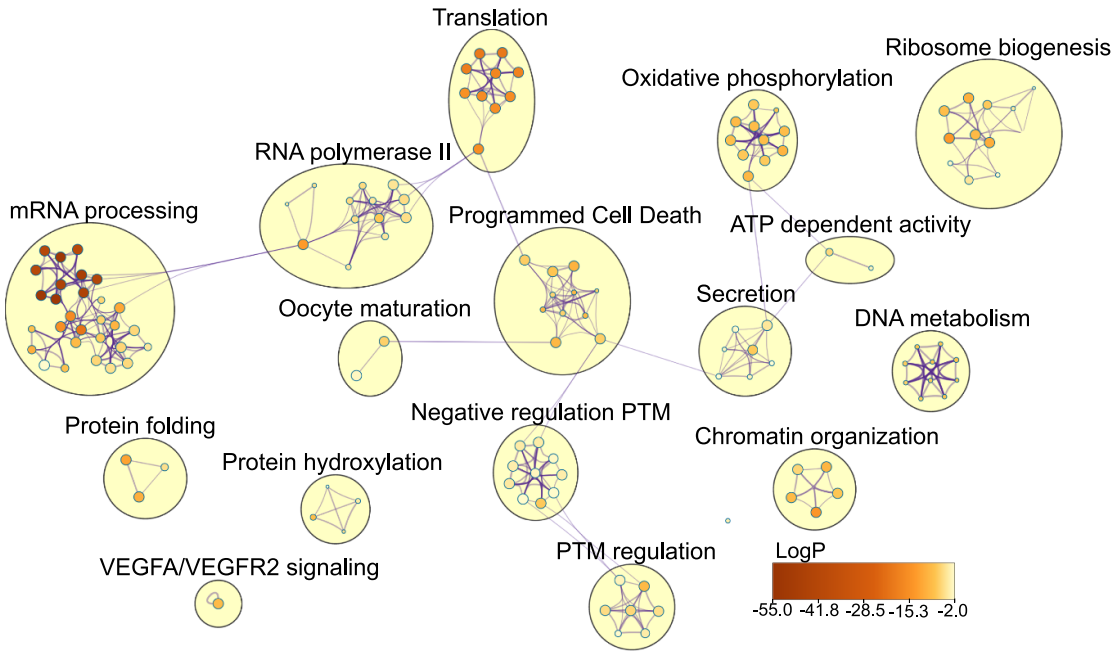

Figure 7. Annotation enrichment analysis conducted on human proteins using Metascape, focusing on proteins with an R_{bind} score exceeding 0.5, and excluding those with fewer than 5 constituting peptides (1396 proteins in total). The depicted network specifically highlights annotations with a p-value below 0.01 and an enrichment factor of at least 1.5.

2.7. Proteome-wide survivin binding prediction and overview of survivin 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 binders was used as a parameter to judge a higher likelihood for affinity to survivin. The proteins with the higher than R_{bind} were used in enrichment analysis of biological annotations using Metascape [61] and the network representation was mapped (figure 7). 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 [62]. 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), AKT1/AKT3, Stratifin, IL18, High mobility group box 1 and 2 (HMGB1/2), Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta (YWHAZ), SEM1 26S Proteasome Subunit and Occludin (OCLN). These apoptotic proteins are non-homologous, catalyze different reactions, and belong to different compartments, yet their composition renders them attractive to survivin. The mitochondrial Smac/DIABLO protein, known for its proapoptotic activity, has an R_{bind} of 0.27 below the median R_{bind} of the proteome (0.31).

Within the mitochondria, survivin influences oxidative phosphorylation [63, 64] and represented in figure 7 by ATP synthase (subunits ε, ATP5MJ and K), NADH:ubiquinone oxidoreductase (B2, B5, B12, B14.5b, C2 and assembly factors NDUFAF1), succinate dehydrogenase complex (assembly factor 2), cytochrome b/c1 complex (subunits VI, VII, IX and X), cytochrome c oxidase (PET100, VIb2, VIc, VIIa1) and ubiquinol-cytochrome-c reductase complex assembly factor 2.

Figure 7 also shows a significant number of translation-related proteins and proteins involved in RNA metabolism, transcription, and splicing, revealing putative functions of survivin in humans. Recent experimental insights on survivin in micro-RNA biogenesis [65] highlights the potential to be able to confirm binding predictions. Moreover, in *Caenorhabditis elegans*, the homologue of survivin, BIR-1, is primarily associated with transcriptional regulation and development [66], lacking evidence of involvement in apoptosis [67]. Indeed, the anti-apoptotic functions of survivin appear to be acquired in only specific evolutionary lineages. Interaction of survivin with the cellular components are further discussed in the accompanying text under the heading 'Additional observations concerning the Metascape analysis of predicted survivin binders of the proteome.'

A particular limitation of enrichment analysis stems from arbitrarily selecting the number of proteins enriched in survivin binding regions. Opting for too few may overlook important 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. With its small solvent-accessible surface, how survivin engages in specific pairwise interactions also poses 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.

2.8. Connection to fragment-based drug design (FBDD)

The composition of peptides and their success in predicting PPI raises very pertinent questions about the compositional description (fragments) of drug molecules. Both peptides and drugs can be seen as collections of functional groups, and the boundary between them may not be as distinct, both in functionality and compositional description. The compatibility of these fragments with a protein is assessed through screening, providing a fundamental determination of their binding potential. This assessment typically precedes the covalent connection of these fragments and the creation of the molecular structure of the drug. Further research is necessary to understand the physical basis of compatibility between functional groups in a protein and the extensive fragment libraries that organic synthesis can provide [68, 69].

2.9. Exploring the effectiveness of composition-based prediction

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. 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 [70], 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 IDPs and unstructured domains under physiological conditions [56]. 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 [71]). Understanding complex molecular 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).

Departing from structural information, our study initially focused on atom displacement analysis in folded proteins. The observation that atoms with similar chemical properties exhibit shared displacement patterns [72, 73] led to a dynamic model in second-quantization, revealing coherent motion of phononic and dipolar collective excitations within the protein crystal [74]. This coherence extends between different proteins over intermolecular distances, involving identical functional groups. Testing this unconventional idea on PPIs yielded a surprisingly simple and effective initial approximation for predicting which peptides bind to survivin [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 [75] for inferring water solubility in organic molecules. However, exceptions to this trend exist [76], 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 do not 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. In addition, non-interacting proteins may not be neutral and their incompatibility can drive active structuring/segregation into separate phases. This concept has analogies with phenomena in polymer, colloid and interface sciences, such as ‘irreversible aggregation’ or ‘reversible self-assembly’, with connections to amyloid fibril formation in various diseased states [77].

Understanding elementary mechanisms behind protein transitions remains difficult due to compensating energy−entropy effects at various length scales [78]. The realistic free energy behaviour of macromolecular systems with heterogeneous composition is a non-local function of monomer density distributions [79], 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 the bifurcation theory, predicting significant perturbations in atomic dynamics with small alterations in control parameters [80].

Given their exceptional functional sensitivity and mutational robustness, it has been suggested that proteins may operate near a critical point where distinct phases merge [81]. 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 [82], akin to a continuous coil-globule transition in synthetic polymers [83]. 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 [84].

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 [85], akin to the law governing the end-to-end distance in a polymer as a function of molecular weight. This relation is influenced by factors like equilibrium spacing, temperature, and force constants connecting atomic pairs, as discussed in the accompanying text under the heading 'Fluctuations in a Monoatomic Lattice'. We have generalized this problem to a linear diatomic lattice with long-range oscillations (acoustic branch), revealing that the collective fluctuations are influenced by the atomic composition, determined by the mass proportions of their respective species. The derivation of this model is elaborated in the supplemental text 'Fluctuations in a Diatomic Lattice', demonstrating that the total fluctuation illustrates a compositional-dependent random walk of single atomic fluctuations, weighted in probability by their masses. Deviations from a perfectly periodic array can therefore be estimated at the melting point of these structures, while still preserving compositional-dependent long-range order. In reality, melting is an unstable process initiated at the surface [86], typically occurring at a temperature approximately (20–30)% lower than the bulk value, but this observation does not undermine the former conclusions. Although diatomic lattices do not fully capture the complexity of peptides, it is crucial to note that even these basic prototype systems display composition-specific signatures. Conversely, the random-walk model we have developed can be extended to accommodate any number of atomic species, even when approaching a very large value.

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 [87], while EO is a (truncated) virial expansion [88]. 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 may be used for a patterning description of charged units, provided the transition is not driven by strong critical fluctuations [89]. 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) [90]. 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 [91]).

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 [89, 92]. Even a standard FH approach with average χ accurately predicts the critical temperature for the germ granule protein Ddx4 [93]. The phase transition rate is governed by concentration and conformational fluctuation, reaching the largest amplitudes at the critical point that induces spontaneous separation. Enhancing PPI results in multiple energy minima, defining the composition of separated phases along the volume fraction coordinate [92]. 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 [94]. 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 [95]. 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 [94].

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 $c_{\text{eff}} \approx n_F/(N_A l_P\ 135\ \text{Å}^3) = (1.7 - 12.5)\ M$, where n_F is the number of functional groups or atoms and l_P is the polypeptide length [96]. On average, this range remains valid even though there is only one peptide (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.

At the local level, the behaviour of peptide or protein solutions 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 which are 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 [97]. When associated with a bifurcation point, this process could facilitate the folding of these structures into a 3D configuration, akin to the principles governing crystal nucleation and growth, highlighting a dynamic attractor resistant to perturbations. Alternatively, it may give rise to distinct periodic and 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 [98] or protein unfolding through pulling [40].

Over billions of years, organisms have leveraged fundamental physical principles, exemplified by signal peptides like nuclear localization signals with significant compositional biases [99]. Despite low information content, such peptides 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 that are primarily determined by genetic material, are further influenced by covalent post-translational modifications like phosphorylation, acetylation, methylation, ubiquitination, nitrosylation, hydroxylation, sulfation and deamidation. Observed individually, or in patterns like hyperphosphorylation, these post-translational modifications also enable protein composition to adapt, triggering new collective dynamics based on introduced groups.

It is important to highlight the limitations of our study. In this research, survivin was the only target for which composition alone has been demonstrated to be sufficient for initially approximating binding partner peptides. It is also likely that in very specific lock-and-key type interactions between partners, a similar method is unlikely to be highly effective. In such interactions, the structural details of the interface and the pattern of attractive and repulsive short-range interactions cannot be ignored. Furthermore, we emphasize that many proteins do not possess ordered structures; instead, many exhibit substantial compositional bias in at least some segments. Therefore, our findings are more likely to be applicable to such protein systems. We acknowledge that there are very convenient tools available for feature selection in tree-based ML methods, and our decision to use an MLP-based analysis was somewhat arbitrary. We found MLPs to be a suitable tool for exploring our dataset. It was particularly important for us to ensure that the model generalizes well beyond the scope of the peptide microarray (training) dataset. We believe that a similar stepwise simplification of a tree-based model may be possible, and it would yield similar explanatory details.

In summary, this work has focused on the predictive influence of 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, a system underlying biological organization is beginning to emerge.

3. Methods

3.1. Peptide microarray experiments

The peptide microarray experiment was described previously [13] and corresponding data is the basis of this analysis [100]. 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^{-1} 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.

3.2. Development of feature scoring set for the representation of peptides

The C-Coord1 features encompassed the standard four non-hydrogen MC atoms, except glycine (3 non-hydrogen atoms) and proline (2 non-hydrogen atoms). Additionally, specific features included the glycine $C\alpha$ 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 $C\alpha$, 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 MC and a methylene group, thus eliminating the necessity for a separate atom type for the $C\alpha$ 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 $C\beta$ atom of lysine forms a quasi-rigid group with the $C\alpha$, 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 $C\delta$ 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 (18 if the perfectly correlated features Pro–CB and Pro–MC are merged) only marginally less than the most common 20 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 models with an elemental composition model, 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 [46], yielding two features.

3.3. 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 one value based on RACS820104 index [101] (Seq) from AAindex database [16]. Fourier transform 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.

3.4. T-SNE dimensionality reduction analysis of peptide microarray dataset

For dimensionality reduction, the t-SNE module from scikit-learn 1.0.1 package was applied to both atom type composition, Seq and UniSeq representations. In the latter cases, t-SNE was performed on the representation before (Seq and UniSeq, as in figures S1(C) and (D)) and after (FT-Seq, as in figure S1(B)) Fast Fourier Transform step. The following parameters were used: *n*_components = 2, perplexity = 30.0, verbose = 1, and random_state = 123.

3.5. 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 = 10 000, 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 3. Prediction statistics were obtained from the averages of 100 prediction runs.

3.6. Improving the sequence-based prediction of survivin binding using embedding and 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 [102] or ProtT5 encoding [103], 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. In contrast, our AA-Letter model has explicit summation of amino acids ignoring their positions.

3.7. 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 $M_{\text{bind}}(n)$, 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. Because of the 10-residue overlap among 15-residue-long peptides, each residue position is usually present in three peptides. Considering the 20 common, natural amino acid variants at each residue position, this resulted in 60 peptide variants, each associated with a binding prediction. The total number of variants predicted to be binders divided by the total number of peptides with variants at 'n' equals $M_{\text{bind}}(n)$. The C–Pos model with two neurons in the intermediate layer was applied on this dataset and 90% of the microarray data set was used. The peptides with C-terminal cysteines were excluded from the data set. The tested peptides were scaled using the microarray data distribution parameters. The $M_{\text{bind}}(n)$ values were mapped on the 3D structures using the cctbx [104] and pymol [105] libraries.

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

Human proteome database was obtained from UNIPROT [106]. 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. The tested peptides were scaled using the microarray data distribution parameters. The trained C–Pos model (90% of the data with C-terminal cysteines excluded) 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). The list of human proteins and their associated R_{bind} value is listed in table S10. The highly ranked proteins were subjected to biological annotation enrichment analysis using Metascape (https://metascape.org) [61]. Network mapping of enrichment annotation terms from predicted survivin interaction partners was made using the Cytoscape platform [107].

Data availability statement

All original code was deposited the Github database https://github.com/Katona-lab/Composition. The repository includes a snapshot of human proteome from the Uniprot data base. Other data that support the findings of this study are available upon reasonable request and by contacting the corresponding authors.

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.7774524.

Acknowledgments

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 the Croatian Science Foundation through the project IP-2022-10-3456. S P W acknowledges funding from BBSRC International partnership Grant.

Author contributions

G K and S A M conceptualized the study outline and discussion. G K supervised the research. A L A, T N O, M J, 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 N O, S A M, and G K with additional input from all authors.

Conflict 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.

ORCID iD

Gergely Katona https://orcid.org/0000-0002-2031-8716

References

[1] Steiner R F 1953 Reversible association processes of globular proteins. IV. Fluorescence methods in studying protein interactions *Arch. Biochem. Biophys.* **46** 291–311
[2] Oncley J, Ellenbogen E, Gitlin D and Gurd F 1952 Protein–protein interactions *J. Phys. Chem.* **56** 85–92
[3] Waugh D F 1954 *Advances in Protein Chemistry* vol 9 (Elsevier) pp 325–437
[4] Sanger F and Tuppy H 1951 The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates *Biochem. J.* **49** 463

[5] Lehner B, Semple J I, Brown S E, Counsell D, Campbell R D and Sanderson C M 2004 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** 153–67
[6] Suter B, Kittanakom S and Stagljar I 2008 Two-hybrid technologies in proteomics research *Curr. Opin. Biotechnol.* **19** 316–23
[7] Sidhu S S, Fairbrother W J and Deshayes K 2003 Exploring protein–protein interactions with phage display *Chembiochem* **4** 14–25
[8] Kodama Y and Hu C-D 2012 Bimolecular fluorescence complementation (BiFC): a 5-year update and future perspectives *Biotechniques* **53** 285–98
[9] Cornett E *et al* 2016 *Method Enzymol* (Elsevier) vol 574 pp 31–52
[10] Rothbart S B, Krajewski K, Strahl B D and Fuchs S M 2012 *Method Enzymol* vol 512 (Elsevier) pp 107–35
[11] Chang T-W 1983 Binding of cells to matrixes of distinct antibodies coated on solid surface *J. Immunol. Methods* **65** 217–23
[12] Breitling F, Nesterov A, Stadler V, Felgenhauer T and Bischoff F R 2009 High-density peptide arrays *Mol. Biosyst.* **5** 224–34
[13] Jensen M *et al* 2023 Survivin prevents the polycomb repressor complex 2 from methylating histone 3 lysine 27 *iScience* **26** 106976
[14] Jeyaprakash A A, Klein U R, Lindner D, Ebert J, Nigg E A and Conti E 2007 Structure of a Survivin-Borealin-INCENP core complex reveals how chromosomal passengers travel together *Cell* **131** 271–85
[15] Kastritis P L and Bonvin A M 2013 On the binding affinity of macromolecular interactions: daring to ask why proteins interact *J. R. Soc. Interface* **10** 20120835
[16] Kawashima S, Ogata H and Kanehisa M 1999 AAindex: amino acid index database *Nucleic Acids Res.* **27** 368–9
[17] Schreiber G 2020 Protein–protein interaction interfaces and their functional implications *Protein – Protein Interaction Regulators Drug Discovery* 78th edn ed S Roy and H Fu (The Royal Society of Chemistry) ch 1, pp 1–24
[18] Lockless S W and Ranganathan R 1999 Evolutionarily conserved pathways of energetic connectivity in protein families *Science* **286** 295–9
[19] Fox J M, Zhao M, Fink M J, Kang K and Whitesides G M 2018 The molecular origin of enthalpy/entropy compensation in biomolecular recognition *Annu. Rev. Biophys.* **47** 223–50
[20] Klebe G 2015 Applying thermodynamic profiling in lead finding and optimization *Nat. Rev. Drug Discovery* **14** 95–110
[21] Lafont V, Armstrong A A, Ohtaka H, Kiso Y, Mario Amzel L and Freire E 2007 Compensating enthalpic and entropic changes hinder binding affinity optimization *Chem. Biol. Drug Des.* **69** 413–22
[22] Van Dan Burg B, Dijkstra B W, Vriend G, Van Dar Vinne B, Venema G and Eijsink V G H 1994 Protein stabilization by hydrophobic interactions at the surface *Eur. J. Biochem.* **220** 981–5
[23] Bogan A A and Thorn K S 1998 Anatomy of hot spots in protein interfaces *J. Mol. Biol.* **280** 1–9
[24] Pace C N, Horn G, Hebert E J, Bechert J, Shaw K, Urbanikova L, Scholtz J M and Sevcik J 2001 Tyrosine hydrogen bonds make a large contribution to protein stability *J. Mol. Biol.* **312** 393–404
[25] Reichmann D, Cohen M, Abramovich R, Dym O, Lim D, Strynadka N C J and Schreiber G 2007 Binding hot spots in the TEM1–BLIP interface in light of its modular architecture *J. Mol. Biol.* **365** 663–79
[26] Reichmann D, Rahat O, Albeck S, Meged R, Dym O and Schreiber G 2005 The modular architecture of protein–protein binding interfaces *Proc. Natl Acad. Sci.* **102** 57–62
[27] Cohen M, Reichmann D, Neuvirth H and Schreiber G 2008 Similar chemistry, but different bond preferences in inter versus intra-protein interactions *Proteins: Struct. Funct. Bioinf.* **72** 741–53
[28] La D, Kong M, Hoffman W, Choi Y I and Kihara D 2013 Predicting permanent and transient protein–protein interfaces *Proteins: Struct. Funct. Bioinf.* **81** 805–18
[29] Pál G, Kouadio J-L K, Artis D R, Kossiakoff A A and Sidhu S S 2006 Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning *J. Biol. Chem.* **281** 22378–85
[30] Evans R *et al* 2010 Protein complex prediction with AlphaFold-Multimer *biorxiv* (https://doi.org/10.1101/2021.10.04.463034)
[31] Yu D, Chojnowski G, Rosenthal M and Kosinski J 2023 AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer *Bioinformatics* **39** btac749
[32] Baek M *et al* 2021 Accurate prediction of protein structures and interactions using a three-track neural network *Science* **373** 871–6
[33] Lupo U, Sgarbossa D and Bitbol A-F 2023 Pairing interacting protein sequences using masked language modeling *bioRxiv* (https://doi.org/10.1101/2023.08.14.553209)
[34] Heffernan R, Yang Y, Paliwal K and Zhou Y 2017 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** 2842–9
[35] Zhang B, Li J, Quan L, Chen Y and Lü Q 2019 Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network *Neurocomputing* **357** 86–100
[36] Liu J and Gong X 2019 Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction *BMC Bioinform.* **20** 1–11
[37] Cadet F, Fontaine N, Li G, Sanchis J, Ng Fuk Chong M, Pandjaitan R, Vetrivel I, Offmann B and Reetz M T 2018 A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes *Sci. Rep.* **8** 16757
[38] Neuvirth H, Raz R and Schreiber G 2004 ProMate: a structure based prediction program to identify the location of protein–protein binding sites *J. Mol. Biol.* **338** 181–99
[39] Caffrey D R, Somaroo S, Hughes J D, Mintseris J and Huang E S 2004 Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13** 190–202
[40] Bordner A J and Abagyan R 2005 Statistical analysis and prediction of protein–protein interfaces *Proteins: Struct. Funct. Bioinf.* **60** 353–66
[41] Zhou H X and Shan Y 2001 Prediction of protein interaction sites from sequence profile and residue neighbor list *Proteins: Struct. Funct. Bioinf.* **44** 336–43
[42] Hwang H, Petrey D and Honig B 2016 A hybrid method for protein–protein interface prediction *Protein Sci.* **25** 159–65
[43] Xue L C, Dobbs D, Bonvin A M and Honavar V 2015 Computational prediction of protein interfaces: a review of data driven methods *FEBS Lett.* **589** 3516–26
[44] Dumetz A C, Snellinger-o'brien A, M, Kaler E W and Lenhoff A M 2007 Patterns of protein protein interactions in salt solutions and implications for protein crystallization *Protein Sci.* **16** 1867–77
[45] van Mierlo G, Veenstra G J C, Vermeulen M and Marks H 2019 The complexity of PRC2 subcomplexes *Trends Cell Biol.* **29** 660–71
[46] Kyte J and Doolittle R F 1982 A simple method for displaying the hydropathic character of a protein *J. Mol. Biol.* **157** 105–32
[47] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

[48] Oren Y, Meister N, Chatterji N, Ladhak F and Hashimoto T B 2023 Proving test set contamination in black box language models (arXiv:2310.17623)
[49] Stadler V *et al* 2008 Combinatorial synthesis of peptide arrays with a laser printer *Angew. Chem., Int. Ed.* **47** 7132–5
[50] Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[51] Wattenberg M, Viégas F and Johnson I 2016 How to use t-SNE effectively *Distill* **1** e2
[52] Saitou N 2018 *Introduction to Evolutionary Genomics* (Springer) pp 109–48
[53] Serena M, Bastos R N, Elliott P R and Barr F A 2020 Molecular basis of MKLP2-dependent Aurora B transport from chromatin to the anaphase central spindle *J. Cell Biol.* **219** e201910059
[54] Elkins J M, Santaguida S, Musacchio A and Knapp S 2012 Crystal structure of human aurora B in complex with INCENP and VX-680 *J. Med. Chem.* **55** 7841–8
[55] Diebold M, Schonemann L, Eilers M, Sotriffer C and Schindelin H 2023 Crystal structure of a covalently linked Aurora-A-MYCN complex *Acta Cryst.* D **79** 1–9
[56] Belo Y, Mielko Z, Nudelman H, Afek A, Ben-David O, Shahar A, Zarivach R, Gordan R and Arbely E 2019 Unexpected implications of STAT3 acetylation revealed by genetic encoding of acetyl-lysine *Biochim. Biophys. Acta Gen. Subj.* **1863** 1343–50
[57] Scrittori L, Skoufias D A, Hans F, Gerson V, Sassone-Corsi P, Dimitrov S and Margolis R L 2005 A small C-terminal sequence of Aurora B is responsible for localization and function *Mol. Biol. Cell* **16** 292–305
[58] Fu J, Bian M, Liu J, Jiang Q and Zhang C 2009 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. USA* **106** 6939–44
[59] Wang H, Holloway M P, Ma L, Cooper Z A, Riolo M, Samkari A, Elenitoba-Johnson K S J, Chin Y E and Altura R A 2010 Acetylation directs survivin nuclear localization to repress STAT3 oncogenic activity *J. Biol. Chem.* **285** 36129–37
[60] Sgrignani J, Garofalo M, Matkovic M, Merulla J, Catapano C V and Cavalli A 2018 Structural biology of STAT3 and its implications for anticancer therapies development *Int. J. Mol. Sci.* **19** 1591
[61] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi A H, Tanaseichuk O, Benner C and Chanda S K 2019 Metascape provides a biologist-oriented resource for the analysis of systems-level datasets *Nat. Commun.* **10** 1523
[62] Wheatley S P and Altieri D C 2019 Survivin at a glance *J. Cell Sci.* **132** jcs223826
[63] Rivadeneira D B, Caino M C, Seo J H, Angelin A, Wallace D C, Languino L R and Altieri D C 2015 Survivin promotes oxidative phosphorylation, subcellular mitochondrial repositioning, and tumor cell invasion *Sci. Signal* **8** ra80
[64] Hagenbuchner J, Kuznetsov A V, Obexer P and Ausserlechner M J 2013 BIRC5/Survivin enhances aerobic glycolysis and drug resistance by altered regulation of the mitochondrial fusion/fission machinery *Oncogene* **32** 4748–57
[65] Andersson K M *et al* 2017 Survivin controls biogenesis of microRNA in smokers: a link to pathogenesis of rheumatoid arthritis *Biochim. Biophys. Acta Mol. Basis Dis.* **1863** 663–73
[66] Kostrouchova M, Kostrouch Z, Saudek V, Piatigorsky J and Rall J E 2003 BIR-1, a Caenorhabditis elegans homologue of Survivin, regulates transcription and development *Proc. Natl Acad. Sci. USA* **100** 5240–5
[67] Fraser A G, James C, Evan G I and Hengartner M O 1999 Caenorhabditis elegans inhibitor of apoptosis protein (IAP) homologue BIR-1 plays a conserved role in cytokinesis *Curr. Biol.* **9** 292–301
[68] Shulga D A, Ivanov N N and Palyulin V A 2022 In silico structure-based approach for group efficiency estimation in fragment-based drug design using evaluation of fragment contributions *Molecules* **27** 1985
[69] Kirsch P, Hartman A M, Hirsch A K H and Empting M 2019 Concepts and core principles of fragment-based drug design *Molecules* **24** 4309
[70] Vovk A and Zilman A 2023 Effects of sequence composition, patterning and hydrodynamics on the conformation and dynamics of intrinsically disordered proteins *Int. J. Mol. Sci.* **24** 1444
[71] Poudyal M *et al* 2023 Intermolecular interactions underlie protein/peptide phase separation irrespective of sequence and structure at crowded milieu *Nat. Commun.* **14** 6199
[72] Ahlberg Gagner V, Jensen M and Katona G 2021 Estimating the probability of coincidental similarity between atomic displacement parameters with machine learning *Mach. Learn. Sci. Technol.* **2** 035033
[73] Gagnér V A, Lundholm I, Garcia-Bonete M-J, Rodilla H, Friedman R, Zhaunerchyk V, Bourenkov G, Schneider T, Stake J and Katona G 2019 Clustering of atomic displacement parameters in bovine trypsin reveals a distributed lattice of atoms with shared chemical properties *Sci. Rep.* **9** 19281
[74] Ahlberg Gagner V *et al* 2023 Femtosecond x-ray snapshots reveal correlated displacements of specific distal atoms in a protein crystal *bioRxiv* (https://doi.org/10.1101/2024.05.29.596429) (Accessed 2 June 2024)
[75] Ebbing D D and Gammon S D 2010 *General Chemistry* vol 484 (Houghton Mifflin)
[76] Ensing B, Tiwari A, Tros M, Hunger J, Domingos S R, Pérez C, Smits G, Bonn M, Bonn D and Woutersen S 2019 On the origin of the extremely different solubilities of polyethers in water *Nat. Commun.* **10** 2893
[77] Ezzat K, Sturchio A and Espay A J 2022 Proteins do not replicate, they precipitate: phase transition and loss of function toxicity in amyloid pathologies *Biology* **11** 535
[78] van der Vegt N F A 2021 Length-scale effects in hydrophobic polymer collapse transitions *J. Phys. Chem.* A **125** 5191–9
[79] Panyukov S V and Kuchanov S I 1992 New statistical approach to the description of spatial inhomogeneous states in heteropolymer solutions *J. Phys.* II **2** 1973–93
[80] Bose I and Ghosh S 2019 Bifurcation and criticality *J. Stat. Mech.: Theory* E **2019** 043403
[81] Tang Q Y, Hatakeyama T S and Kaneko K 2020 Functional sensitivity and mutational robustness of proteins *Phys. Rev. Res.* **2** 033452
[82] Lifshitz I M, Grosberg A Y and Khokhlov A R 1978 Some problems of the statistical physics of polymer chains with volume interaction *Rev. Mod. Phys.* **50** 683–713
[83] Gasic A G, Boob M M, Prigozhin M B, Homouz D, Wirth A J, Daugherty C M, Gruebele M and Cheung M S 2019 Critical phenomena in the temperature-pressure-crowding phase diagram of a protein *Phys. Rev.* X **9** 041035
[84] Kosterlitz J M and Thouless D J 1973 Ordering, metastability and phase transitions in two-dimensional systems *J. Phys.* C **6** 1181
[85] Peierls R 1979 *Surprises in Theoretical Physics* (Princeton University Press)
[86] Pietronero L and Tosatti E 1979 Surface theory of melting *Solid State Commun.* **32** 255–9
[87] Brangwynne C P, Tompa P and Pappu R V 2015 Polymer physics of intracellular phase transitions *Nat. Phys.* **11** 899–904
[88] Bot A, Dewi B P C and Venema P 2021 Phase-separating binary polymer mixtures: the degeneracy of the virial coefficients and their extraction from phase diagrams *ACS Omega* **6** 7862–78
[89] Lin Y H, Song J H, Forman-Kay J D and Chan H S 2017 Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins *J. Mol. Liq.* **228** 176–93

[90] Das R K, Ruff K M and Pappu R V 2015 Relating sequence encoded information to form and function of intrinsically disordered proteins *Curr. Opin. Struct. Biol.* **32** 102–12
[91] Singh P B, Belyakin S N and Laktionov P P 2020 Biology and physics of heterochromatin-like domains/complexes *Cells* **9** 1881
[92] Martin E W and Mittag T 2018 Relationship of sequence and phase separation in protein low-complexity regions *Biochemistry* **57** 2478–87
[93] Nott T J *et al* 2015 Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles *Mol. Cell* **57** 936–47
[94] Choi J M, Holehouse A S and Pappu R V 2020 Physical principles underlying the complex biology of intracellular phase transitions *Annu. Rev. Biophys.* **49** 107–33
[95] Berry J, Brangwynne C P and Haataja M 2018 Physical principles of intracellular organization via active and passive phase transitions *Rep. Prog. Phys.* **81** 046601
[96] Chen C R and Makhatadze G I 2015 ProteinVolume: calculating molecular van der Waals and void volumes in proteins *BMC Bioinform.* **16** 101
[97] Mezzasalma S A, Kruse J, Merkens S, Lopez E, Seifert A, Morandotti R and Grzelczak M 2023 Light-driven self-oscillation of thermoplasmonic nanocolloids *Adv. Mater.* **35** 2302987
[98] Seelig J and Schonfeld H J 2016 Thermal protein unfolding by differential scanning calorimetry and circular dichroism spectroscopy two-state model versus sequential unfolding *Q. Rev. Biophys.* **49** e9
[99] Labaj P P, Sykacek P and Kreil D P 2011 An analysis of single amino acid repeats as use case for application specific background models *BMC Bioinform.* **12** 1–10
[100] Jensen M *et al* 2023 *Zenodo*
[101] Rackovsky S and Scheraga H 1982 Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids *Macromolecules* **15** 1340–6
[102] Chollet F Keras 2015 (available at: https://keras.io)
[103] Elnaggar A *et al* 2022 ProtTrans: toward understanding the language of life through self-supervised learning *IEEE Trans. Pattern Anal.* **44** 7112–27
[104] Grosse-Kunstleve R W, Sauter N K, Moriarty N W and Adams P D 2002 The computational crystallography toolbox: crystallographic algorithms in a reusable software framework *J. Appl. Crystallogr.* **35** 126–36
[105] Schrodinger L L C 2015 The PyMOL molecular graphics system, version 1.8
[106] Consortium T U 2022 UniProt: the universal protein knowledgebase in 2023 *Nucleic Acids Res.* **51** D523–31
[107] Shannon P, Markiel A, Ozier O, Baliga N S, Wang J T, Ramage D, Amin N, Schwikowski B and Ideker T 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks *Genome Res.* **13** 2498–504