Polycation–π Interactions Are a Driving Force for Molecular Recognition by an Intrinsically Disordered Oncoprotein Family

Jianhui Song1*, Sheung Chun Ng2*, Peter Tompa3,4, Kevin A. W. Lee2*, Hue Sun Chan1*

1 Departments of Biochemistry, Molecular Genetics, and Physics, University of Toronto, Toronto, Ontario, Canada, 2 Division of Life Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong S.A.R., China, 3 VIB Department of Structural Biology, Vrije Universiteit Brussel, Brussels, Belgium, 4 Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungary

Abstract

Molecular recognition by intrinsically disordered proteins (IDPs) commonly involves specific localized contacts and target-induced disorder to order transitions. However, some IDPs remain disordered in the bound state, a phenomenon coined “fuzziness”, often characterized by IDP polyvalency, sequence-insensitivity and a dynamic ensemble of disordered bound-state conformations. Besides the above general features, specific biophysical models for fuzzy interactions are mostly lacking. The transcriptional activation domain of the Ewing’s Sarcoma oncogene family (EAD) is an IDP that exhibits many features of fuzziness, with multiple EAD aromatic side chains driving molecular recognition. Considering the prevalent role of cation–π interactions at various protein-protein interfaces, we hypothesized that EAD-target binding involves polycation–π contacts between a disordered EAD and basic residues on the target. Herein we evaluated the polycation–π hypothesis via functional and theoretical interrogation of EAD variants. The experimental effects of a range of EAD sequence variations, including aromatic number, aromatic density and charge perturbations, all support the cation–π model. Moreover, the trend orders observed are well captured by a coarse-grained EAD chain model and a corresponding analytical model based on interaction between EAD aromatics and surface cations of a generic globular target. EAD-target binding, in the context of pathological Ewing’s Sarcoma oncogenes, is thus seen to be driven by a balance between EAD conformational entropy and favorable EAD-target cation-π contacts. Such a highly versatile mode of molecular recognition offers a general conceptual framework for promiscuous target recognition by polyvalent IDPs.

Introduction

Understanding the sequence-function relationship of a protein and how it might malfunction is central to biomedical research. While many proteins function in their folded states, recently it became clear that intrinsically disordered proteins (IDPs) also play key functional roles [1,2] in transcription, translation and cell cycle regulation that, when altered, frequently lead to cancer [3]. Indeed, ∼70% of proteins implicated in cancer are predicted to have significant disordered regions [3,4]. Molecular recognition by IDPs typically involves target-induced folding. Intriguingly, however, certain IDPs engage in protein-protein interaction without coupled folding and binding [5] such that the IDP remains disordered even when bound to a globular target. This phenomenon has been termed “fuzziness” [6] and is characterised by IDP polyvalency, sequence-insensitivity and lack of strict geometric complementarity for binding [6]. Important examples of fuzziness include transcription factors [7], linker histones [8], prion-like proteins [9] and Sic1-Cdc4 in yeast [10].

To gain insight into “fuzzy” interactions, we have studied the Ewing’s Sarcoma (EWS)-activation domain (EAD) in the TET family of RNA-binding proteins [11] and Ewing’s family of oncogenes (EFPs). EAD is a ∼280 residue polyvalent IDP comprised mainly of a degenerate repeat motif SYGQQS. Studies of EAD have mostly focused on its role in naturally occurring EFPs in which it is fused to various transcription factor partners. EFPs are potent EAD-dependent transcriptional activators, resulting in distinct phenotypes of the associated Ewing’s family of tumors [12,13] which are largely dictated by the DNA-binding domain of the EWS fusion partner. Progress in understanding EAD has been hindered by its IDP properties [14] and a general lack of biophysical/biochemical insights [15]. Another barrier is the paucity of information regarding cognate EAD-interacting proteins. Because native EWS interacts with a highly complex array of proteins at a network hub [16,17] or potentially as a scaffold protein [18], it is probable that EAD has numerous partners.

Functional studies of EFPs have provided a foundation for understanding sequence-function relationship of EAD. Most
Author Summary

Understanding how proteins recognize each other is central to deciphering the inner workings of living things and for biomedical research. It has long been known that the sequence of a protein, which is a string of different amino acids, can dictate how a protein molecule folds into a well-defined shape required for biological tasks. Many folded proteins recognize and bind with each other by a tight geometric fit similar to that between a lock and its key. Recently, however, it has become clear that some proteins function as a flexible string, in constant motion, without forming a stable shape. Understanding how such “disordered” proteins work is challenging. To gain insight, we studied a disordered protein region that causes a large family of human cancers. Employing an innovative combination of experimental and theoretical techniques, we describe a new mode of protein interaction based on multiple simple contacts between one type of amino acid (aromatic) in the disordered protein and another type (positively charged) on the partner protein. Because this mechanism also underlies the ability of the disordered protein to cause cancer, further investigation of this unprecedented mode of protein-protein interaction may open up new avenues for cancer therapy.

notably, the transcriptional and oncogenic activity of EAD is conferred by multiple tyrosine (Y) residues due to their aromaticity but not hydrophobicity [14]. EAD function is also markedly sequence-insensitive [14], although a permissive overall composition is apparently required. This type of interaction shares features with other systems that exploit polyvalent IDP phosphorylation, as in autoinhibition of CFTR [19], auto-regulation of Ets-1 transcription factor [20,21] and interaction of Cdk inhibitor Sic1 with its E3 ubiquitin ligase Cdc4 [10]. Sic1 has nine low-affinity Cdc4-binding sites and a threshold number of phosphorylated sites induces highly cooperative “polyelectrostatic” binding of Sic1 to a single positively charged pocket in Cdc4 [10,22–24]. Similarly, EAD activity requires cooperative action of multiple aromatic moieties in a disordered structure [14,25], though it does not require phosphorylation. Thus molecular recognition by EAD was coined “polyaromatic” [26]. However, the physical basis for polyaromatic EAD function has not been elucidated.

In light of the versatile roles of cation-π interactions in protein folding and protein-protein interactions [27–37], we hypothesized that a major contribution to molecular recognition by EAD (within EFPs) is the attraction between numerous unconstrained aromatic residues (π’s) on the EAD and basic residues (cations) on the target. We tested this idea experimentally and also theoretically in molecular simulations that are based on cation-π contacts between the EAD and a generic folded target. We found broad agreement between EAD functionality and simulated EAD binding. Thus our findings strongly support the polycation-π model and suggest that similar mechanisms might also be exploited by other IDPs.

Results

Rationale of the investigation

In view of the abundance of aromatic residues in EAD (38 Ys in the native EAD) and the significant strength of cation-π interactions [38], we posit cation-π as a highly plausible and probably most significant type of EAD-target contacts. This leaves open the possibility of additional contributions but these are likely to be secondary [14]. To probe the nature of EAD-target interactions we designed EAD mutants with different numbers, placements and types of aromatic residues in conjunction with EAD charge variations (Fig. S1). In vivo transcriptional activity of EAD mutants was compared, relatively, with computational predictions of binding probability assuming the polycation-π hypothesis. In the absence of specific knowledge about EAD targets, a generic globular target with appropriate surface charge was used for simulations to provide a minimalist physical model for the proposed interactions.

Functional effect of Y-dosage is consistent with simulated EAD-target binding via cation-π interactions

The intact EAD spans ~280 highly repetitive residues but such a long sequence is not particularly amenable to mutagenesis and is also quite impractical for computational studies. This hurdle can be overcome by exploiting small EAD regions (~40 residues) that faithfully mimic the salient features of the intact EAD and whose transcriptional activity (transactivation) can be readily tested using a multisite reporter [14,25]. To establish a framework for experiments, we began by functionally testing a 66-residue polypeptide (10Yn) with ten Y residues, which is closely related to part of the native EAD, and the corresponding series of mutants (4Yn-9Yn) varying only in Y number (ny) but retaining the same Y density (Fig. 1A). Transactivation was quantified by a well-established transient assay in Jeg3 cells with EAD sequences fused to the DNA-binding domain of zta protein and a zta reporter plasmid (pZ7bac) [39] (Methods). As for other EAD sequences studied before [25], transactivation rises in a nonlinear manner with ny (Fig. 1C, red circles), demonstrating that multiple Ys act together in a cooperative manner.

To assess the polycation-π idea, we constructed a coarse-grained chain simulation model that embodies the hypothesis. The EAD is represented by a flexible Cα chain and a generic globular target protein is modeled as a sphere with surface charge distribution (Fig. S2A,B) resembling that of the RNA polymerase II subunits Rpb4/Rpb7 (PDB id: 2C35; Fig. S2C), which was reported to bind to EAD [40,41]. Binding is driven by EAD-target cation-π contacts (Fig. S2D), the interaction energies (Fig. 1B and Fig. S2E) of which are consistent with published estimates of cation-π potentials of mean force in aqueous environments, with attractive well depths ~3.0 to ~5.5 kcal/mol [32,33,38]. In accordance with PDB data [30,32], contacts between one cation and multiple aromatics or between one aromatic and multiple cations are allowed; but the orientation dependence [42] and nonadditivity [43] of cation-π interactions are neglected. Because EAD-target cation-π interactions are suggested to be highly dynamic with bound EAD remaining disordered, we included an average solvation effect [38] rather than considering the discrete water configurations that impact on cation-π interactions [44]. Our model also incorporates electrostatic and intra-EAD hydrophobic effects (Fig. S2F) by using potential functions similar to those developed for coarse-grained protein folding simulations [45,46]; but EAD-target hydrophobic interactions were not considered because of insufficient knowledge about the real target. Binding probability (Pb) was determined using Monte Carlo sampling (see Methods and Supporting Text S1 for details).

Fig. 1C shows that the simulated Pb values rationalize the functional data regarding the effect of Y (aromatic) number. A similar agreement with model simulation was also observed for the activities of a set of previously studied EAD sequences (Fig. S3). Noting that the EAD-zta proteins used for determining activity are dimers whereas EAD monomers were used in our simulations, we also verified that the EAD monomer and dimer Pb values have a
similar Y-number dependence (Fig. S4), indicating that EAD monomer simulations are adequate for capturing behavioral trends of the corresponding EAD dimers.

We emphasize that the experimental-theoretical comparisons in Fig. 1 and subsequent figures are between relative experimental activities and relative P₈s. The model binding free energy ΔGₐ = −k_BT ln [P₈ / (1 − P₈)], where k_BT is Boltzmann constant and T is absolute temperature (Fig. 1D), is dependent upon the effective EAD concentration (see below). However, the latter is unknown experimentally and our simple model does not account for every physical interaction between the real EAD and its target. Thus, it is not meaningful to compare absolute P₈ against absolute experimental activity. Nonetheless, by assuming that putative unknown factors affect different EAD sequences similarly (Text S1), one may compare the differences in simulated ΔGₐ for various EAD sequences with the corresponding differences in EAD activity. Doing so yielded a good agreement between experiment and theory for the 4Yn–10Yn sequences (Fig. 1D), lending support to the polycation-π hypothesis.

An analytical model of polycation-π mediated IDP binding to a folded target

To better understand how EAD binding might be affected by various assumptions about the target and multisite IDP binding in general, we developed a simple analytical model to complement the chain simulations. Briefly, our analytical model considers an IDP chain of n contour length units with Nₓ equally spaced aromatic residues that are k units apart, and a target with Nc cations. When the IDP is distant from the partner, it can adopt Ωₓ(n) conformations with any residue fixed in space; that residue in turn can access a volume Vᵢ, i.e., the IDP concentration is 1/V. Binding is favored by an energy Eₓ(Δn) for each IDP-target cation-π contact. A bound IDP has ±1 such contact, with NₓNc possible pairings for the first contact. Because the volume accessible to the first contacting residue is reduced from V to a small “capture” volume δV and the number of IDP conformations is reduced from Ωₓ(n) to a smaller Ωₓ(n) because of IDP-target excluded volume, it follows that the change in free energy upon forming the first contact is ΔEₓ = ±k_BT ln (V/δV) + ln Ωₓ(n)/Ωₓ(n) − ln NₓNc.

\[ \Delta G_b = \frac{E_x}{k_BT} + \ln \left( \frac{V}{\delta V} \right) + \ln \left( \frac{\Omega_x(n)}{\Omega_x(n)} \right) - \ln N_c \]

where we have used the number of conformations \( \Omega_x(n) \) with a mid-chain attachment for \( \Omega_x(n) \), neglecting the small variation in \( \Omega_x(n) \) that depends on the attaching point (Text S1); thus \( \Omega(k_l, R_l | n) / \Omega_x(n) \) is the conformational reduction factor for forming an IDP loop. \( \{l_i\} \) in \( \Sigma(l_i) \) represents all \( 2^{N_x} - N_x - 1 \) possible sets of 2±2 aromatic residues that can contact the target (resulting in 1 to \( N_x - 1 \) loops). The \( \Sigma(l_i) \) term vanishes when
$N_p = 1$. $\Pi_i$ is over the different loops for a given set of contacting residues. We assumed that the loops are independent and neglected the excluded volume repulsion among them. Exact enumeration of self-avoiding lattice fields [47] (Figs. S3C–E, S6, Supporting Tables S1, S2, S3) and extrapolations of such data (Fig. S7) were applied to estimate the conformational entropy terms involving $\Delta G_k$ in Eq. (1). Further details of the model are provided in Text S1.

Salient features of the analytical model are shown in Fig. 2. An essentially linear dependence of $\Delta G_k$ on $N_p$ is seen (Fig. 2A) as for the simulation results (Fig. 1D). As expected, a stronger (more negative) $E_{\text{ex}}$ leads to tighter binding (more negative $\Delta G_k$). The binding equilibrium is governed by a balance between favorable cation-$\pi$ contacts on one hand and translational and conformational entropy on the other (Fig. 2A, inset). Binding increases with aromatic density $1/k$, IDP concentration C (Fig. 2B,D; $C \sim 1/V$), and target cation density (Fig. 2C). Fig. 2A shows that the $\Delta G_k$ trend for $E_{\text{ex}} = -3.5k_BT$ in our analytical model matches approximately the behavior of $\Delta G_k$ in the chain simulation in Fig. 1D. This value of $E_{\text{ex}} = -2.1 \text{ kcal/mol}$ (for $T = 300 \text{ K}$ used in this study) is comparable but weaker than the average pairwise cation-$\pi$ energy ($\approx -3.3 \text{ kcal/mol}$) we determined from our simulation using a cation-$\pi$ potential energy well depth of $\approx 3.6 \text{ kcal/mol}$ (Fig. 1B). This discrepancy is not unexpected because excluded volume effects among the loops are neglected in Eq. (1), resulting in an overestimation of binding probability. Nonetheless, the overall trend exhibited by the chain simulation model is well reflected by the analytical model.

**Efficacy of different cation-$\pi$ strengths and intramolecular competition by EAD cations supports the model**

In addition to accounting for Y-number dependence (Fig. 1), the cation-$\pi$ hypothesis also rationalizes EAD activity of mutants with Y substituted by phenylalanine (F) or tryptophan (W). Statistical analysis of PDB structures [32] and quantum calculations [48] have indicated that the cation-Y and cation-F strengths are similar, with F slightly weaker [48], but cation-W is significantly stronger (Text S1). Consistent with this trend, Fig. 3A shows that the experimental activity of 5Fn is slightly lower than that of 5Yn [25], but the activity of 5Wn is ~8 fold that of 5Yn. Simulated $P_B$s for these sequences using the corresponding cation-$\pi$ energies in Fig. 1B mirror these experimental observation, lending further credence to the cation-$\pi$ hypothesis.

We next investigated the effect of altering EAD charge. First, we changed anion composition by introducing aspartic acid (D) residues (Fig. 3B). Adding 3 Ds to 10Yn (10Y3D) or adding 5 Ds to the minimally active 5Y (5Y5D) barely changes activity. The fact that anion additions do not enhance EAD activity rules out favorable contacts between EAD anions and target cations as a major driving force for binding. Second, we changed cation composition by introducing arginine (R) residues (Fig. 3C). Inasmuch as the 66-residue EAD peptides are flexible as posited by our chain simulation model, the inserted Rs would allow intra-EAD cation-$\pi$ contacts and thus reduce activity by competition. Fig. 3C shows that an EAD with 10 Ys and 5 Rs (10Y5R) is indeed much less active than one with only 8 Ys and zero Rs (8Yn) and is comparable with a protein containing only 5 Ys. Similarly, 8Y2RAD (containing 8 Ys and 2 Rs) is comparable with 6YAD (6 Ys and zero Rs) and both EADs are approximately 3-fold less active than 8YAD (8 Ys and zero Rs). Apparently, the addition of R residues within the EAD functionally counteracts Ys in an essentially one-to-one manner. This finding is highly suggestive of Y-R contacts between EAD and real target proteins and thereby strongly supports the cation-$\pi$ hypothesis.

The relative simulated $P_B$ values broadly capture the activity trends for charge variations (Fig. 3C). Quantitative agreement between simulation and experiment is seen for 10Y3D, 5Yn, 6YAD, and 8Y2RAD. Simulation also accounts for the near-irrelevance of anion number for 5Y and 5Y5D activities (Fig. 3B). Simulations did however slightly overestimate the decrease in activity caused either by reduction of Y number from 10 to 5 (Fig. 3B, compare 10Yn with 5Y or 5Y5D) or by introduction of cations into 10Yn (Fig. 3C, compare 8Yn with 10Y5R). The average EAD-target electrostatic energy is essentially neutral or very slightly repulsive in our model (+0.2 kcal/mol). Because of the dominance of cation-$\pi$ over electrostatic interactions (Fig. S2E,F), $P_B$s of 10Yn and 10Y3D are very similar; but there is some EAD-target electrostatic repulsion due to the anions on 10Y3D, resulting in a slightly weaker average EAD-target cation-$\pi$ interactions in Oncoproteins

**Figure 2. IDP-target binding in the analytical model.** To match the chain simulation model, we used $\delta V = 2e^{-3/|R_c|}$ as in the simulations [hence ln($V$/$\delta V$) = 13.1]. Inset: The energy ($\Delta G_k$) and entropy ($T\Delta S_k$) components of $\Delta G_k$ for $E_{\text{ex}}/k_BT = -3.5$. Results in (B–D) are also for $E_{\text{ex}}/k_BT = -3.5$. (B) Effects of k and V on binding: $N_p = 32$; $C_b = 1/(600 \text{ Å}^3)$ is used as a reference IDP concentration. The black curves show $\Delta G_k$ at $C_b$ for hypothetical sequences with $k = 9, 8, 7, 6, 5, 4$, and 3 (from top to bottom), $n = 66$ for $k = 6$ and $n = 10N_p$ for $k \geq 7$. The blue curves are for the $k = 6$ sequences for three IDP concentrations $C_C$, with $C_b = 0.25, 3.0, 10.0$ (from top to bottom). (C) $\Delta G_k$ for $k = 6$ sequences at $C_b = 1$ on different targets of the same size with different $N_p = 8, 16, 32, 48, 64$, and 80 (from left to right; see Text S1 and Fig. S5B). (D) $P_B$s of the $k = 6$ sequences at different IDP concentrations $C = 1.00, 5.0, 4.0, 3.0, 2.0, 1.0, 0.5, 0.33$, and 0.25 (from top to bottom).

doi:10.1371/journal.pcbi.1003239.g002

PLOS Computational Biology | www.ploscompbiol.org 4 September 2013 | Volume 9 | Issue 9 | e1003239
energy for 10Y3D compared with 10Yn (−22.2 vs −23.4 kcal/mol). Intra-EAD cation-π interactions in the unbound state are strong in R-containing mutants, amounting on average to −31.9 kcal/mol for 10Y3R and −67.6 kcal/mol for 10Y5R and are slightly weaker in the bound state (−24.9 and −63.6 kcal/mol respectively). The favorable EAD-target cation-π energy acquired upon binding is −18.5 kcal/mol for 10Y3R and −7.0 kcal/mol for 10Y5R on average, indicating that the weaker binding of 10Y5R is caused by increased competition from intra-EAD cation-π interactions due to the larger number of Rs present.

Interplay between number of cation-π contacts and EAD conformational entropy determines activity

As shown in Fig. 2B, the polycation-π hypothesis envisions that EAD activity depends on both aromatic number and density. We tested this prediction using EAD sequences with constant Y

![Figure 3](image-url)
Figure 4. Effect of Y density and distribution on EAD activity. (A, B) The EAD peptides (left) were tested for relative transactivation (black) and simulated $P_b$ (grey), shown in the same style as in Fig. 3. (A) 7Yn (see Fig. 1A) with Y density denoted normal (n or $k = 6$) was compared with 7Yn/2 (Y density $−1/2$ of 7Yn, $k = 12$) and 7Yn/4 (Y density $−1/4$ that of 7Yn, $k = 24$). The actual simulated $P_b$ for 7Yn is 0.11. (B) 10Yn (see Fig. 1A; $k = 6$) was compared with 5Y ($k = 12$) and the sequence 5YP which has 5 pairs of sequentially adjacent Ys. The asterisk indicates that 5Y activity is overstated due to relatively higher expression of 5Y protein. (C) Analysis using our analytical model. All $\Delta G_{b}$s were for $N_c = 32$, $C = 1$, and $E_{cp} = −3.5 k_B T$ except the data point plotted as open circle ($\Delta G_b = −2.6 k_B T$) was for $E_{cp} = −5.1 k_B T$. The solid line shows results for $k = 6$ and $n = 66$. The upper and lower dashed lines provide results for $k = 12$ with chain lengths $n = 66$ and $n = 71$ respectively. The diamonds show results (from bottom to top) for 7Yn, 7Yn/2, and 7Yn/4 in (A), which have chain lengths $n = 66, 86$, and 156 respectively. To facilitate comparison with the $n_c = 7$ data in (A), $N_c = 7$ is marked by the vertical dotted line. The squares show results for 5Y ($N_c = 5$; $\Delta G_b = 4.0 k_B T$) and 10Yn ($N_c = 10$; $\Delta G_b = −3.2 k_B T$) in (B), both with $n = 66$. As discussed in Text S1, the model represented by the open circle may be applied to 5YP in (B) with $−5.1 k_B T$ as the interaction energy between a cation and two adjacent aromatic residues.

doi:10.1371/journal.pcbi.1003239.g004

Discussion

A distinctive fuzzy protein-protein interaction

The significance of protein disorder in the bound state or “fuzziness” has only recently emerged [6]. Theoretical modeling of IDPs [22,49–67], especially for fuzzy complexes [22], is also in its infancy but provides powerful tools for understanding dynamic conformer ensembles. Our integrated functional and computational approach has culminated in a distinctive model for fuzzy interactions (Fig. 5) that may contain the core features of a more general mode of protein-protein interaction. The model involves a simple biophysical contact (cation-$\pi$), strong cooperativity stemming from both IDP and target polyvalency, and a highly flexible and dynamic IDP conformer population in the bound state. Stable binding requires a sufficient number of cation-$\pi$ contacts but allows kinetic exchanges between myriad bound states. Notably the molecular recognition events studied here are particular to the diseased state of EFP-induced malignancies and are therefore of immediate biomedical interest.

Robustness of the polycation-$\pi$ model

Our hypothesis is intuitive given that cation-$\pi$ interactions have wide and versatile biological roles, the interaction is strong [38] and EAD is highly polyvalent. The native intact EAD is also virtually devoid of cationic residues and thus especially amenable to trans cation-$\pi$ interactions with target proteins. Here, our interrogation of the polycation-$\pi$ model covered a wide range of EAD sequence properties (variations of Y number, cation-$\pi$ strength, charge, Y density, and Y distribution) as well as simulation parameters (physically relevant variations of the cation-$\pi$, hydrophobic, and electrostatic interaction strengths; see

PLOS Computational Biology | www.plosoemb.org 6  
September 2013 | Volume 9 | Issue 9 | e1003239
In all these tests, the polycation-\(\pi\) hypothesis provides a consistent biophysical account of the experiments. Other types of interactions are much less likely to contribute dominantly to EAD-target binding and our experiments address some of these. Of particular interest is the stoichiometric intramolecular blocking of \(Y_s\) by \(R_s\) within EAD (Fig. 3C). This observation argues against alternative EAD-target aromatic interactions such as \(\pi-\pi\) stacking which are, in any event, probably of insufficient strength [68] in the absence of proximate metal ions [69] to account for the slope of \(Y\) number dependence of EAD activity (Fig. 1). One may also imagine a scenario in which EAD compaction is induced by \(Y\)-dependent hydrophobic interactions such that EAD-target contacts may involve poorly defined non-aromatic entities. But this possibility is strongly contraindicated by the high degree of EAD disorder [14] and also by our finding that EAD compaction by introduction of \(R\) residues (Text S1) actually decreases activity. Potential hydrogen bonding effects are not addressed in our model due to insufficient experimental data. Intuitively, hydrogen bonding effects are not addressed in our model due to insufficient experimental data. Intuitively, hydrogen bonding involving prevalent EAD residues (Gln, Ser, and Thr) may well contribute to molecular recognition by EAD, although previous data [14,25] together with the current study indicate that cation-\(\pi\) interactions are the essential driving force. More refined studies will be required to uncover secondary and more subtle contributions to EAD-target binding, including potential couplings between hydrogen bonding and cation-\(\pi\) interactions [70]. We also stress that our results do not preclude additional effects due to EAD posttranslational modifications, including tyrosine phosphorylation and O-GlcNAcylation [71], that might sometimes be manifest for particular EFPs and/or in specific physiological circumstances.

We have assumed a globular target because the biophysical aspects of the proposed model strongly predict that a large number of real globular proteins interact with EAD. Nonetheless, a disordered (IDP) target that enables favorable cation-\(\pi\) contacts with the EAD is also possible (Fig. S8) although so far the fuzzy complexes known to involve two IDPs are homodimers [72,73]. We cannot infer how many cation-\(\pi\) contacts are required for EAD binding to real targets. It is also likely that particular interactions will deviate in some manner from our generic model.

One can envision a variety of target determinants that might have an impact, including, for example, number and/or density of cations, acute geometric constraints imposed by residues flanking target cations, and the contribution of other aromatic side chain interactions such as hydrogen bonding.

**Comparison of polyelectrostatic and polycation-\(\pi\) interactions**

Polyelectrostatic (Sic1/Cdc4) and polycation-\(\pi\) interactions share some similarities. Each may well reflect a general mode of interaction for polyvalent IDPs. In contrast to Sic1-Cdc4, however, the properties of the EAD studied herein are related to the diseased state [14] and our study points to several significant biophysical differences between EAD and Sic1/Cdc4. First, Sic1/Cdc4 binding involves a single Cdc4 site while EAD binding in our model invokes multiple simultaneous contacts. Second, Sic1/Cdc4 interaction is switch-like, reflecting the biological need for acute response to cell cycle kinase levels, whereas the EAD is constitutively polyvalent [14,25]. Third, like most other polyvalent IDPs, Sic1 has short sequence-specific or linear motifs [74,75], a single copy of which can mediate suboptimal or high-affinity Sic1/Cdc4 binding.
Cde4 binding [10]. Such elements are almost certainly absent in EAD [14]. Fourth, the multiple cation-π contacts that underpin EAD binding in our model entail transient restrictions of EAD conformations (though they remain disordered), whereas a Sic1 bound to a single Cde4 pocket at a given instant is not subject to such conformational restriction [22–24].

Biological implications

The molecular recognition events studied here are related to pathological EAD function and, accordingly, are not obviously shaped by evolution [14]. Some aspects of EAD malfunction are an indirect consequence of loss of the EWS RNA-binding domain (RBD) or gain of a foreign DNA-binding domain in EFPs. In relation to our study, it is intriguing that the EWS RBD contains highly disordered regions with reiterated RGG that antirepress EAD [76], quite possibly via intramolecular masking [40]. The polycation-π perspective may offer a rationalization for this behavior. The simulated binding between a disordered EWS peptide containing multiple RGG boxes and the 10Yn EAD indeed reveals a strong interaction (Fig. S8). Intramolecular cation-π interactions between EAD and RGG have high potential to impact native EWS function by competing out aberrant interactions between EAD and the putative globular proteins relevant to EAD malfunction in oncogenesis.

In this regard, knowledge of EAD-target interface might provide therapeutic avenues [77] for Ewing’s family tumors with poor prognosis. Several small molecule inhibitors of EWS/Fli1 have been identified. Interestingly, they all have aromatic character [78–80] or, in one case, a very basic short peptide sequence [81]. Whether any of them target the EAD portion of EWS/Fli is unknown. Due to their likely being effective cation-π competitors, it will be of great biomedical interest to explore this possibility.

How may polycationic molecular recognition by EAD relate to normal EWS protein function? This is a challenging question given the strong evolutionary conservation of EWS [82] that includes several EAD properties: a positionally conserved Gln two residues C-terminal to Y, Y phosphorylation sites [83], and SH2/SH3 interaction sites. However, none of these features are required, at least in some cases, for oncogenic EAD function [14]. Perhaps the mode of EAD action in EFP oncoproteins reflects a primordial polycationic function that was subsequently tailored by evolution to fulfill normal cellular roles. For example, Y phosphorylation can dramatically increase the aromatic-cation interactions required for peptide inhibitors of Src [84], indicating that phosphorylation of only a limited number of Ys in EAD could have profound effects on EAD-target interactions that are important for normal EWS.

To conclude, the proposed model for molecular recognition by EAD expands the seemingly endless modalities for IDP function and malfunction. The hitherto unrecognized polycation-π mode of IDP-target binding can be versatile. It offers a highly plausible biophysical basis for EAD and perhaps other scaffold/networking proteins to interact with many distinct target proteins [16–18]. The present methodology and results can also be extended to facilitate the exciting search for real EAD targets.

Methods

Experiment

Plasmids: pZΔE [25] and pZ7Luc [39] are previously described. All other plasmids expressing EAD variants were derived from the mammalian expression vector pShlencer 4.1-CMV neo (Applied Biosystems). Proteins: pZΔE expresses a protein lacking EAD sequences and containing only the ATF1 region and zta bZIP domain [25]; see Fig. S1. Transactivation assays and Western blotting. Transfections, trans-activation assays and quantification of transactivation under linear assay conditions were performed as previously described [25]. Activity values were corrected for background activity determined by including the EAD-negative protein ZΔE in transfections. Details for plasmid and EAD construction and the assays are provided in Text S1.

Simulation

The EAD is modeled as a Cn chain. Pairwise interactions between amino acid residues depend on whether they are aromatic, hydrophobic, charged, or polar (see Text S1 and Fig. S2E,F for definition). The generic EAD-binding target is a sphere of radius 16.0 Å with 32 positively and 32 negatively charges on its surface (Fig. S2A). The total energy of the model system $E_T = E_{\text{intrachain}} + E_{\text{chain-target}}$ is the sum of the intramolecular energy $E_{\text{intrachain}}$ within the EAD and the intermolecular energy $E_{\text{chain-target}}$ between the EAD and its target. The expressions for these energy functions, other modeling details, and control simulations are provided in Text S1.

Supporting Information

Figure S1 Proteins and EAD sequences used in the present study. Transcriptional activator proteins (Top) contain the experimental sequences related to the N-terminal 66 residues of EAD1-66 (box with purple Ys), the region of ATF1 present in the EWS ATFT1 oncogene and the DNA-binding domain of zta protein (ZΔTDBD). In (A)–(C), amino acid residues are denoted by the standard one-letter code. Sequences for Figs. 1, 3, and 4 in the main text are listed, respectively, under (A), (B) and (C).

Figure S2 The chain simulation model. (A) The generic EAD binding target (partner) is a sphere of radius $R_p = 16$ Å with essentially evenly distributed positive and negative charges (represented by blue and red beads respectively). (B) An EAD sequence is modeled as a Cn chain (beads on a string) that can engage in cation-π, electrostatic, hydrophobic, and excluded-volume interactions as specified in the main text and Text S1. In this figure and subsequent supporting figures, aromatic (Y in this drawing) and hydrophobic (w) residues are shown in magenta and orange, respectively, whereas positively and negatively charged residues are shown in blue and red respectively. All other residues are shown in grey. (C) The distribution of positively charged residues on the heterodimer of the Rpb4/Rpb7 subunits of human RNA polymerase II was used as a reference for the design of the charge density on the generic EAD binding target. The histogram here shows the shortest distance from each of the 32 positively charged amino acid residues (R or K) on Rpb4/Rpb7 (16 each along the Rpb4 and Rpb7 chains) from another positively charged residue, based on the X-ray crystal structure (PDB ID: 2C3S) determined by Meka et al. ref. [10] of Text S1. The distances are measured between the atoms that have the positive charges. The red dashed horizontal line marks the average shortest distance which is $\approx 9.4$ Å. (D) EAD-target binding is defined in the model as having at least one EAD aromatic residue (magenta circle) within a capture radius $R_c = 6$ Å from a positive charge (blue circle) on the target. One such cation-π contact between an EAD sequence (brown string connecting magenta circles) and the target (large shaded circle with embedded blue circles) is shown in this schematic drawing. (E,F) Energetic components of the interaction potential, the horizontal variable $r$ here corresponds to $r_y$ in Eq. (S1) or $r_x$ in Eq. (S2). (E) Model cation-π interaction
polycation-π hypothesis from a re-analysis of early experiments on 33-residue EAD sequences. Sequences and experimental data were taken from ref. [1] of Text S1. Simulations were conducted using the same chain model as described in Text S1 and the main text in a (600 Å)³ simulation box. (A) The sequences are defined in the above reference. The experimental relative activities and the simulated relative binding probabilities are represented by the black and grey bars respectively. (B) The sequences in (A) are grouped according to their Y number n_Y. Plotted are the simulated binding probability (solid squares) and the relative experimental activity (open circles) averaged over sequences belonging to each given n_Y. For the simulation results, the averages are over all possible permutations of Y positions for a given n_Y, including those not studied by experiments. Note that both Y number and Y density are varied among this set of sequences (unlike the set in Fig. 1 that varies only the Y number while keeping Y density constant). Error bars show variation among sequences with the same n_Y. Lines joining the solid squares are merely a guide for the eye.

Figure S4 Simulated binding probabilities of monomer and dimer EAD sequences follow similar trends. Similar dependences on n_Y are observed for cis-duplication of small EAD elements in a single dimer. The monomer sequences used in the present simulations are the same 33-residue sequences based on the construction by Feng and Lec [1] of Text S1 studied in Fig. S3. As for the simulations in Fig. S3, all possible permutations of Y positions are considered. Each dimer was constructed by joining the C-terminus of a given monomer sequence to the C-terminus of another copy of the same monomer sequence by a linker chain. The linker consists of six residues that are neither charged nor hydrophobic; all reference bond angles within the linker are equal to 165° with a stiff bond-angle force constant equal to 10.0k_BT. Thus, in this figure, a dimer sequence with Y number n_Y is equivalent to two identical monomer sequences with Y number n_Y connected by such a linker. (A) A snapshot of an n_Y = 5 monomer bound to the target. (B) A snapshot of the corresponding n_Y = 10 dimer bound to the target. The EAD chains are depicted in a tube representation with the color code for different residue types specified in Fig. S2B. (C) Free energies of binding were computed under the same conditions as those used for Fig. S3. ΔGₘ values averaging over sequences with the same n_Y are plotted.

Figure S5 Components of the analytical model. (A) Schematic of cation-π contacts along the IDP. Here we only consider IDP chains with evenly spaced aromatics that are k residues apart; thus the contour length between two cation-contacting aromatics is always in the form of kl, where l is a positive integer. Three example contact patterns are shown, wherein the aromatics and cations are depicted as magenta and blue circles respectively. (B) Distribution of cation-cation distance R_i on the target. Each R_i value is the distance in Å from a given cation to a different cation, measured on the spherical surface of the model target (left drawing). The distribution n_c(R_i) is shown (histograms) for three different targets of the same size but different cation densities. As for the target with N_c = 32 cations in most of our simulations, the cations are essentially evenly distributed on the surface for the N_c = 8 and N_c = 96 targets. The approximately even distribution of charges on the target sphere was achieved by a numerical algorithm (see Text S1). As can be seen from the histograms, only a few of the R_i values are exactly identical. (C) An example conformation configured in the simple cubic lattice with one end of the chain touching a plane. The number of such conformations is referred to as Ω_0^A(n) in this work. (D) An example simple cubic lattice conformation with two of its mid-chain sites in contact with a plane. We denote the number of such conformations as Ω_0^C(n). (E) Change in conformational entropy (in units of the Boltzmann constant k_B) upon bringing a free lattice conformation to form a contact at a chain end (squares) or at mid-chain (circles) with an infinite impenetrable plane that imposes excluded volume on the other side of the plane (the space underneath the plane is not accessible to the chain). The data points (squares or circles) were computed using exact enumeration data in Table S1. The curves through the data points were generated by fitting the assumed relation y = ln[A exp(-αn) + B exp(-βn)]. The fitting parameters here are A = 0.5365, B = 0.53139, α = 0.02786, and β = 0.33604 for y = ln[Ω_0^A(n)/Ω_0^B(n)]; and A = 0.40915, B = 1.12627, α = 0.05373, and β = 0.39353 for y = ln[Ω_0^C(n)/Ω_0^B(n)].

Figure S6 Conformational entropy loss upon loop formation. The quantity Ω(l,R_i/n) is the number of simple cubic lattice conformations of length n (n is the total number of beads along the chain) that have one chain end (bead number 1) touching an excluded-volume plane at a given point (as in Fig. S5C) and, at the same time, bead number l+1 also making a contact with a given point on the plane at a distance R_i from where bead number l touches the plane, thus forming a loop of length l that spans a distance R_i on the plane (top left drawing). Note that conformations that form other chain-plane contact(s) in addition to these two are included in the Ω(l,R_i/n) count. As discussed in the main text and in Text S1, the vertical variable ln[Ω(l,R_i/n)/Ω_0^C(n)] for the plots in this figure corresponds approximately to the conformational entropy change, in units of k_B, upon making an additional chain-plane contact to form a loop of length l along a chain that has already made at least one contact with the plane. Each of the plotting panels provides the conformational entropy change upon forming a loop of a given length l as a function of R_i. Both l and R_i are shown in units of the lattice bond length (nearest distance between two beads on the
simple cubic lattice). Data points (open circles) in the plotting panels were computed by exact enumeration of lattice conformations with chain lengths from \( n = 4 \) through \( n = 12 \) (see Text S1 and Tables S2 and S3). Multiple data points for the same \( R_j \) value represent results from different \( n \) values. The continuous curves are quadratic fits in the form of

\[
\ln \left( \frac{\Omega(l,R_j|n)}{\Omega^{\text{m}}(n)} \right) = -a(l)(R_j-b(l))^2 + c(l).
\]

The \( l \)-dependent fitting parameters \( a(l), b(l), \) and \( c(l) \) are provided in Fig. S7. In view of the clustering of data points from different \( n \) values, we have made an approximation in the analytical model that

\[
\ln \left( \frac{\Omega(l,R_j|n)}{\Omega^{\text{m}}(n)} \right) \text{ is independent of } n.
\]

**Figure S7** Applying the lattice conformational entropy estimates to the analytical model. (A–C) The fitting parameters \( a(l), b(l), \) and \( c(l) \) for the conformational entropy changes shown in Fig. S6 are provided as data points in (A), (B), and (C), respectively. The continuous fitting curves are given by

\[
A(l) = A + B \exp(-C l), \quad B = 7.04181, \quad C = 0.52115; \quad (B) \quad b(l) = A + B \ln(C l), \quad A = 0.97499, \quad B = 0.93564; \quad \text{and} \quad (C) \quad c(l) = A + B \exp(-C l - D l), \quad A = -5.19530, \quad B = 2.98286, \quad C = 0.31975, \quad D = 2.79004.
\]

These expressions were used to estimate \( \ln \left( \frac{\Omega(l,R_j|n)}{\Omega^{\text{m}}(n)} \right) \) for \( l > 16 \) by extrapolation. (D) The extrapolated \( \ln \left( \frac{\Omega(l,R_j|n)}{\Omega^{\text{m}}(n)} \right) \) function (black curve) is compared against the corresponding random-flight expression

\[
\ln \left( \frac{3/2\pi l N}{\exp \left( -3R_j^2/2l \right)} \right) \text{ (red dashed curve) for } l = 60. \quad \text{(E) Two methods for estimating the entropic cost of loop formation in the analytical model are compared. Potted are the binding free energies of the model EAD chains in Fig. 1 for } E_{\text{ex}} = -3.5k_BT. \quad \text{The black data points (circles) were computed by using entropy estimates from exact enumeration for } \xi = 16 \text{ and extrapolated estimates for } l > 16, \text{ whereas the red data points (triangles) were obtained by using entropy estimates from exact enumeration for } \xi = 16 \text{ but random-flight estimates for } l > 16. \quad \text{The plot here shows that the predicted } \Delta G_N \text{ values based on the two different loop entropy estimates are very similar.}
\]

**Figure S8** Exploring other EAD-target binding scenarios. The EAD sequences are the same as those in Fig. 1. (A) Simulated EAD binding probability \( P_N \) with a hypothetical target in which the surface charges are not evenly distributed but confined to a patch. Two such hypothetical patch partners were considered, both with 12 cations localized on a patch with the same local cation density as the generic target with 32 cations (Fig. S2A) that we have used for most of the simulations. One of the targets (referred to as the positive patch target) contains 12 cations and no anions on the patch whereas the other (referred to as the neutral patch target) contains 12 cations and 12 anions. Plotted here are the simulated binding probabilities for the positive (squares) and neutral (circles) patch targets in either a simulation box of size of \( 300 \text{ Å}^3 \) (black symbols) or \( 600 \text{ Å}^3 \) (blue symbols). (B) A snapshot of an \( n = 10 \) EAD sequence (tube representation) bound to the neutral patch target. (C) Simulated EAD binding probability \( P_N \) with hypothetical disordered (IDP) partners. The EAD sequences and simulation conditions are the same as those in Fig. 1B,C, using a simulation box of size \( 600 \text{ Å}^3 \). During the binding simulations, both the EAD and the hypothetical IDP target were allowed to sample all accessible conformations while the center of mass of the IDP target was kept at a fixed position in the center of the simulation box. We considered a class of such targets, each of which is a chain consisting of 64 alternating cations and anions (32 cations and 32 anions). The adjacent cation and anion are connected by a 5 Å virtual bond with a stiff bond-angle force constant equal to \( 10.0k_BT \). Shown here are binding probabilities for four different such IDP targets with equilibrium bond angles that equal, respectively, to 105° (crosses), 120° (diamonds), 135° (squares) and 150° (circles). A general trend of increasing binding with increasing \( \sigma_N \) is observed for all four hypothetical IDP targets. Not surprisingly, the quantitative details of this trend are sensitive to the persistence length of the IDP target. Binding increases with the flexibility of the IDP target. Also included for comparison (blue triangles) are the simulated probabilities of EAD binding with the RGG3 sequence in the Ewing’s sarcoma RNA-binding domain GGDGRGGRGGMRGGRGLMDRGGPGMGFRGGGRGDRGFRGRGMGDGRGGGGRRGGPGG (refs. [27,28] in Text S1). Here the RGG3 sequence was modeled as a Cs chain using the same modeling scheme as that for the EAD sequences. (D) A snapshot of an \( n = 10 \) EAD sequence (tube representation) bound to a hypothetical IDP target (red and blue beads) with 150° bond angles.

**Table S1** Numbers of conformations, or self-avoiding flights, on the simple cubic lattice. Conformational counts as functions of chain length (number of beads) \( n \) are obtained by exact enumeration. A chain with \( n \) beads has \( n-1 \) bonds. Here, \( D_0 \) is the number of unconstrained conformations; \( \Omega^{\text{m}} \) is the number of conformations that have one chain end anchored onto an impenetrable plane (Fig. S5C); and \( \Omega^m_\text{na} \) is the number of conformations that have the mid-chain bead \( (n/2) \)th bead if \( n \) is even, \( \{ (n+1)/2 \} \)th bead if \( n \) is odd) making a contact with an impenetrable plane (Fig. S5D).

**Table S2** Loop probabilities determined by exact lattice conformational enumeration. Tabulated here are examples (not a complete list) of conformational counts \( \Omega(l,R_j|n) \) used in Fig. S6. Here one chain end is always in contact with the origin \((0,0)\) of a two-dimensional coordinate system for the impenetrable plane. In this table, the positions on the impenetrable plane where another contact with the chain existed are indicated by the \( (x,y) \) coordinates. In the present treatment of our analytical model, \( R_j \) values from all combinations of \( x,y \) (where \( x < y \)) that have nonzero \( \Omega(l,R_j|n) \) counts for \( n \leq 17 \) were used to estimate the conformational entropic cost of loop formation (Figs. S6 and S7).

**Table S3** Exact lattice enumeration data for loop formation probability. Tabulated here are examples as the exact \( \Omega(l,R_j|n) \) counts for \( l = 16 \) and \( n = 17 \). The horizontal and vertical labels correspond, respectively, to the \( x \) and \( y \) coordinates of the positions on the impenetrable plane. One end of the chain (first bead) is always anchored at the origin \((0,0)\). In this table, the entry at a given position \((x,y)\) is the number of conformations that have the chain’s last \((n/2)\)th bead contacting the given position and thus making a loop with \( R_j = \sqrt{x^2 + y^2} \). Data are shown only for \( x \leq y \) because of the obvious rotational symmetry.

**Text S1** Experimental and Computational Details and Rationale. (PDF)
Acknowledgments

J.S. and H.S.C. thank Julie Forman-Kay for discussion, and are grateful to Mitch Kovarik and the colleagues at SciNet of Compute Canada for computational support. K.L. thanks Liang Feng, Mingjie Zhang and Xuhui Huang for discussions, Kim Li for technical support, and the staff of TOP Gene Technologies Montréal for excellent gene synthesis.

References


Author Contributions

Conceived and designed the experiments: KAWL HSC. Wrote the paper: PT KAWL HSC. Performed the experiments: JS SCN KAWL HSC. Analyzed the data: JS SCN PT KAWL HSC.


Text S1

Supporting Information

for PLoS Comput Biol article

Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family

Jianhui Song,1† Sheung Chun Ng,2† Peter Tompa,3,4 Kevin A. W. Lee,2* and Hue Sun Chan1*

1Departments of Biochemistry, Molecular Genetics, and Physics, University of Toronto, Toronto, Ontario M5S 1A8 Canada; 2Division of Life Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong S.A.R., China; 3VIB Department of Structural Biology, Vrije Universiteit Brussel, Building E, 1050 Brussels, Belgium; and 4Institute of Enzymology, Hungarian Academy of Sciences, H-1113 Budapest, Hungary.

* To whom correspondence may be addressed. Email: bokaw@ust.hk or chan@arrhenius.med.toronto.edu
† These authors contributed equally to this work.

(Higher-resolution versions of the supporting figures and tables on pages 10-19 of this document are also provided separately as individual files)

Experimental and Computational Details and Rationale

Experimental details

General aspects of the methodology have been described previously [1,2]. Details that are specific to the present study are as follows.

Plasmids. For the EAD variants that were derived from the mammalian expression vector pSiencer 4.1-CMV neo (Applied Biosystems), pCMVvec contains unique HindIII and BglII sites between the CMV promoter and coding sequence of the ATF1 region present in the EWS/ATF1 oncoprotein except that the ATF1 bZIP domain is replaced with the zta bZIP domain as previously described [3]. For construction of EAD mutants, HindIII/BglII ended synthetic DNA fragments were obtained by commercial gene synthesis (TOP Gene Technologies, Montreal, QC, Canada) and directly inserted into pCMVvec digested with HindIII/BglII. All proteins also contained the KT3 monoclonal epitope PPPEPET [4] at the C-terminus adjacent to the zta bZIP domain.

Proteins. The complete amino acid sequences of all EAD mutants are shown in Fig. S1. 10Yn protein contains an EAD peptide identical to EAD N-terminal residues 1-66 except that the position of four of the ten Ys present is exchanged with nearby residues to give approximately even spacing of Y residues (Fig. 1 in the main text and Fig. S1A). For 9Yn-4Yn the corresponding number of Y residues in 10Y are substituted with prevalent residues in EAD (namely Ala, Gly, Thr, Ser or Gln) to maintain the overall composition and have regularly spaced Ys located in the middle of the peptide at the same density as 10Yn. 5Fn and 5Wn proteins correspond to 5Yn with all Ys replaced by F and W respectively. 5Y protein corresponds to native EAD1-66 with alternate Y residues replaced with prevalent residues in EAD. 7Yn/2 and 7Yn/4 are related to 7Yn and contain seven evenly spaced Ys with linear Y density (given by the distance between the first and last Ys) of approximately half (7Yn/2) and one quarter (7Yn/4) of that in 7Yn. Extra sequences present in 7Yn/2 and 7Yn/4 were derived from 7Yn (minus Ys) to maintain overall composition. Other EAD mu-
tants shown in Fig. S1 are 10Y3D, 5Y5D, 10Y3R, 10Y5R, 8YΔD, 8Y2RΔD and 6YΔD.

Transactivation assays and Western blotting. Activity values were corrected for background activity determined by including the EAD-negative protein ZΔE in transfections. In cases where EAD mutations resulted in significant changes in protein levels, protein expression in vivo was normalized by using different amounts of plasmid for transfection. Luciferase assays were performed at 40 h post-transfection. Western blotting using primary antibody KT3 [4] and alkaline phosphatase conjugated anti-mouse secondary antibody (DAKO) were as previously described [5].

Rationale and computational details of the chain simulation model

Intra-EAD and EAD-target interaction potentials. As outlined in Method of the main text, our chain simulation model describes binding of various EAD sequences (Fig. S1) with a generic target (binding partner) which is a sphere of radius $R_{0} = 16.0$Å. The total potential energy of the model system $E_{t} = E_{\text{intrachain}} + E_{\text{chain-target}}$ is the sum of the intramolecular energy $E_{\text{intrachain}}$ and the intermolecular chain-target energy $E_{\text{chain-target}}$, where

$$
E_{\text{chain-target}} = \sum_{i=1}^{n} \varepsilon_{\theta}(\theta_{i} - \theta_{0})^{2} + \sum_{i=1}^{n} \varepsilon_{\theta} \left( \sum_{j=1}^{n} \frac{r_{\exp ij}}{r_{ij}} \right)^{12} + \varepsilon_{\text{ex}} \left( \frac{\sigma_{\text{ex}}}{r_{ij}} \right)^{6} + \varepsilon_{\text{ex}} \left( \frac{\sigma_{\text{ex}}}{r_{ij}} \right)^{12} + \frac{1}{4\pi\varepsilon_{0}\varepsilon_{d}} \frac{q_{i}q_{j}}{r_{ij}} \exp \left( -\frac{r_{ij}}{\lambda_{D}} \right),
$$

where $n$ is the number of residues in the EAD sequence; $i,j$ label residue positions; $\theta_{i}$ is the $i$th virtual bond angle (in radian), $\theta_{0} = \pi/2$ is the reference virtual bond angle, $r_{ij}$ is the spatial distance between the $i$th and $j$th residues. $\varepsilon_{\theta}$, $\varepsilon_{\text{ex}}$, and $\varepsilon_{\text{bp}}$ are the interaction strengths for bond angle, excluded volume and hydrophobic interactions, and are set to 1.0, 1.0, and $-3.0 \ k_{B}T$ respectively (where $k_{B}$ is Boltzman constant and $T$ is absolute temperature). The range of excluded volume repulsion $r_{\text{exp ij}} = 5.0$ Å if both $i$ and $j$ are charged residues; otherwise $r_{\text{exp ij}} = 4.0$ Å for all other residue pairs as in our previous protein chain models [6,7]. $\kappa_{i}$ is 1 for hydrophobic residues, $\kappa_{i} = 0$ otherwise; and $2n = b = 3.8\AA$ sets the range of the hydrophobic interactions. $\varepsilon_{ij}$ is the cation-π interaction strength that depends on the aromatic residue and is nonzero only when $(i,j)$ is a (cation, π) or (π, cation) pair; and $\sigma_{\text{ex}}$ is set to 4.0 Å. In the electrostatic (last) term, $q_{i}$ ($= 0$ or $\pm 1$) is the charge of residue $i$; $\varepsilon_{0}$ is vacuum permittivity, $\varepsilon_{d} = 40$ is the dielectric constant; and $\lambda_{D}$ is the screening length, which we set to 10.0Å to mimic physiological conditions. The EAD-target interaction is given by

$$
E_{\text{chain-target}} = \sum_{i=1}^{n} \varepsilon_{\theta}(\theta_{i} - \theta_{0})^{2} + \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_{\text{ex}} \left( \frac{r_{\exp ij}}{r_{ij}} \right)^{12} + \varepsilon_{\text{ex}} \left( \frac{\sigma_{\text{ex}}}{r_{ij}} \right)^{6} + \varepsilon_{\text{ex}} \left( \frac{\sigma_{\text{ex}}}{r_{ij}} \right)^{12} + \frac{1}{4\pi\varepsilon_{0}\varepsilon_{d}} \frac{q_{i}q_{j}}{r_{ij}} \exp \left( -\frac{r_{ij}}{\lambda_{D}} \right),
$$

which is a sum of energy terms for EAD-partner excluded volume, cation-π, excluded volume between charged residues ($r_{\text{exp}} = 5.0$ Å), and electrostatic interactions (in the order given in the above equation). The index $v$ labels the charges $q_{i,j}$ on the binding partner, $r_{v}$ is the distance between residue $i$ in the EAD and charge $v$ on the partner, and $r_{ci}$ is the distance between the center of the partner and EAD residue $i$.

The model globular target. As outlined in the main text, the generic target is modeled by a sphere of radius $R_{0}$ with positive and negative charges embedded on its surface (Fig. S2A). The EAD sequences are modeled as $C_{n}$ chains (Fig. S2B). Taking a simple, minimalist approach, we
assume that the generic target is electrically neutral with equal numbers of positive and negative charges, and that the charges are evenly distributed on the surface of the generic target. We employed the “Golden Section Spiral Algorithm” [8,9] implemented in MATLAB (MathWorks, Natick, Massachusetts) to construct essentially even distributions for the charges. Most of the EAD binding simulations reported in this work – unless noted specifically otherwise – are for a generic target with 32 positive charges and 32 negative charges, wherein the distances between two nearest neighboring charges of the same sign and of opposite signs are, respectively, 9.4 Å and 5.1 Å. These distances were designed to mimic the charge distribution of RNA polymerase II subunits Rpb4/Rpb7 (PDB ID: 2C35) [10], for which the shortest distances between positive-positive, negative-negative, and negative-positive pairs are 9.4 Å (Fig. S2C), 7.7 Å, and 5.7 Å, respectively.

Model EAD chains. The Cα chain model for an EAD may be envisioned as a string of beads wherein the distance between the centers of two adjacent beads is equal to the Cα–Cα virtual bond length \( b = 3.8\text{Å} \) (Fig. S2B). Similar coarse-grained Cα models have provided much biophysical insight into protein folding and dynamics (see, e.g., [7] for a recent review). In the present EAD chain model, we set the reference virtual bond angle \( \theta_0 \) in Eq. (S1) to \( \pi/2 \) radian (90°), which coincides approximately with the peak value of the distribution of virtual bond angles in the Protein Data Bank (PDB) [11] but is somewhat smaller than the 106.3° [12] or 105° [13] used in other Cα chain models for proteins. To allow for more chain flexibility and efficient sampling of a large number of possible EAD conformations, we adopted a weak interaction strength \( \varepsilon_\theta = 1.0k_BT \) for the virtual bond angle term in Eq. (S1). Consequently, every virtual bond angle \( \theta \) can sample a range from \( \pi/4 \) (45°) to \( 3\pi/4 \) (135°) quite freely because it entails an energetic cost of at most \(-0.6k_BT\).

Types of interaction in the model. Both the intra-EAD and EAD-partner interactions [Eqs. (S1) and (S2) respectively] are dependent upon the EAD sequence. Pairwise interactions between amino acid residues (represented by their Cα positions) depend on whether they are aromatic (Y, F, W), hydrophobic (A, V, L, I, M, W, F, Y, P) — which include the aromatics [14], charged (D, E, R, K), or polar (N, C, Q, G, H, S, T). An EAD chain is considered to be bound to the generic target if at least one aromatic residue along the EAD sequence is spatially within a capture radius \( R_c = 6.0\text{Å} \) from a target cation (Fig. S2D). In the present model, the N- and C-termini of the EAD chain carry a positive and a negative charge respectively. These charges participate in intra-EAD and EAD-target electrostatic interactions, but the N-terminal positive charge does not engage in cation-π interactions.

Strengths of cation-π contacts in the model. In the present modeling setup, the total interaction between an aromatic residue and a cation is the sum of one of the 12-6 Lennard-Jones potentials in Fig. S2E and a general excluded volume term, which is equal to \( \varepsilon_{\text{ex}}(r_{\text{ep},ij}/r_{ij})^{12} \) for an intra-EAD cation-π interaction [Eq. (S1)] and \( \varepsilon_{\text{ex}}[(r_0 + r_p)/r_{cd}]^{12} \) for an EAD-target cation-π interactions [Eq. (S2)]. The effects of these two general excluded volume terms are similar and are relatively small, each amounting to a decrease of well depth by \( \approx 0.16 \text{ kcal/mol} \) relative to the potentials in Fig. S2E. The total intra-EAD cation-π potentials in Fig. 1B are practically identical to the corresponding total EAD-target cation-π potentials along the radial direction of the target. As discussed in the main text, the well depths of our model cation-π interactions (Fig. 1B) are in line with published estimates of cation-π potentials of mean force in aqueous environments with well depths \( \approx -3.0 \) to \(-5.5 \text{ kcal/mol} \) [15–17]. The differences in well depth among our model cation-Y, cation-F, and cation-W interactions were designed in accordance with PDB statistics. We relied largely on the PDB cation-π contact frequencies compiled by Gallivan and Dougherty [15] in this regard. Because the PDB structures were determined in aqueous environments, PDB statistics are more directly relevant to the aqueous cation-π potentials of mean force we aim to model than experimental and theoretical data on cation-π interactions in the absence of solvation effects [18]. In the dataset considered by Gallivan and Dougherty [15], the frequencies of R, K, Y, F, and W are, respectively, \( p(R) = 10,919, \ p(K) = 13,446, \ p(Y) = 8,309, \ p(F) = 9,162, \) and \( p(W) = 3,412; \) and the frequencies of R-Y, R-F, R-W, K-Y, K-F, and K-W contacts are, respectively, \( p(R-Y) = 749, \ p(R-F) = 630, \ p(R-W) = 609, \ p(K-Y) = 438, \ p(K-F) = 285, \) and \( p(K-W) = 283 \) (Table 1 of [15]). Using a simple formulation for statistical potential [19] and \( k_BT \approx 0.6 \text{ kcal/mol} \) for \( T \approx 300\text{K} \), the difference in cation-Y and cation-F contact energy (former minus latter) may be es-
estimated as $-k_B T \ln \left\{ \frac{[p(R-Y)/p(R-F)][p(F)/p(Y)]}{[p(W)/p(Y)]} \right\} = -0.16 \text{kcal/mol}$ for an R cation and as $-k_B T \times \ln \left\{ \frac{[p(K-Y)/p(K-F)][p(F)/p(Y)]}{[p(W)/p(Y)]} \right\} = -0.32 \text{kcal/mol}$ for a K cation. These energy differences are consistent with the $-0.07$ to $-0.37 \text{kcal/mol}$ range of well depth differences between cation-Y and cation-F interactions we adopted in Fig. 1B. Similarly, the difference in cation-Y and cation-W contact energy (former minus latter) may be estimated as $-k_B T \ln \left\{ \frac{[p(R-Y)/p(R-W)]}{[p(W)/p(Y)]} \right\} = +0.41 \text{kcal/mol}$ for an R cation and $-k_B T \ln \left\{ \frac{[p(K-Y)/p(K-W)]}{[p(W)/p(Y)]} \right\} = +0.27 \text{kcal/mol}$ for a K cation. These energy differences are quite similar to the $+0.42 \text{kcal/mol}$ well depth difference between cation-Y and cation-W in Fig. 1B. In the PDB analysis of Crowley and Golovin [17], W is also seen to interact significantly stronger with R than Y or F for protein complexes and homodimers, but the trend is less clear for the interaction of W, Y, or F with K (Table III in [17]).

**Hydrophobic and electrostatic interactions.** The functional forms for the hydrophobic and electrostatic potentials in the present model (Fig. S2F) are similar to those used in previous coarse-grained modeling studies [20,21]. For most of the present simulations, we used a hydrophobic interaction strength $\epsilon_{hp} = -3.0 \text{kcal/mol}$ and a dielectric constant $\epsilon_d = 40$. We chose an $\epsilon_d$ value intermediate between the dielectric constant $\approx 78.5$ for bulk water and $\approx 2 - 4$ for the interior of a folded protein [22] because physically both the intra-EAD and EAD-target electrostatic interactions take place in an aqueous environment (not the interior of a folded protein) but with significant effective local protein (IDP) concentration.

**Varying the hydrophobic interaction strength.** To assess the robustness of our model predictions and to better delineate the conditions for the validity of the predictions, we have conducted control simulations using alternative values of $\epsilon_{hp}$ and $\epsilon_d$. We found that strengthening the hydrophobic interaction strength from $\epsilon_{hp} = -3.0 \text{kcal/mol}$ to $\epsilon_{hp} = -7.0 \text{kcal/mol}$ while keeping other modeling parameters unchanged did not have much effect on the binding of the $4\text{Yn} - 10\text{Yn}$ sequences in Fig. 1B. However, binding became very weak when hydrophobic interaction was strengthened to $\epsilon_{hp} = -13.0 \text{kcal/mol}$ (well depth $\approx 2.0 \text{kcal/mol}$) because in that case the aromatics would interact strongly with other aromatics and/or other hydrophobic residues and are sequestered in the interior of compact conformations instead of undergoing cation-$\pi$ interactions with the target.

**Varying the electrostatic interaction strength.** We have also considered two alternative $\epsilon_d$ values while keeping other modeling parameters unchanged. First, application of the distance-dependent dielectric constant of Jha and Freed, viz., $\epsilon_d(r) = (\epsilon_d^0 - \epsilon_d) \left[ (sr)^2 + 2sr + 1 \right] \exp(-sr/2)$, where $\epsilon_d^0 = 78.5$ is the dielectric constant of bulk water, $\epsilon_d^0 = 1.77$ and $s = 0.274$ [23] led to an even shallower attractive well for the electrostatic interactions than $\epsilon_d = 40$. Using this $\epsilon_d(r)$ led only to small changes to the simulated binding probabilities for the $4\text{Yn} - 10\text{Yn}$ sequences in Fig. 1B. Second, we tested $\epsilon_d = 20$. This $\epsilon_d$ value led to a deeper attractive well of $\approx 1.5 \text{kcal/mol}$ for the electrostatic interactions (blue dashed curve in Fig. S2F), which is $\approx 1.5/3.6 - 40\%$ of the cation-$\pi$ well depth in our model. This ratio of $\approx 40\%$ is relevant because a quantum mechanical calculation of the methylammonium-acetate and the methyl-ammonium-benzene potentials of mean force in water by Gallivan and Dougherty suggests the same ratio $(-2.2 \text{kcal/mol} / -5.5 \text{kcal/mol} = 40\%)$ of well depths between typical salt-bridge and cation-$\pi$ interactions [16]. For a test set of EAD sequences that include the highly charged $10\text{Y5R}$, $10\text{Y3R}$, and $10\text{Y3D}$ in Fig. 3 and Fig. S1, the simulated binding probabilities using $\epsilon_d = 20$ are quite similar to the corresponding probabilities simulated using $\epsilon_d = 40$. Specifically, the simulated $P_b$ values of $4\text{Yn}$, $5\text{Yn}$, $6\text{Yn}$, $7\text{Yn}$, $8\text{Yn}$, $9\text{Yn}$, and $10\text{Yn}$ are, respectively, 0.011, 0.025, 0.057, 0.109, 0.196, 0.297, and 0.427 for $\epsilon_d = 40$ (data plotted in Fig.1B) and are 0.008, 0.029, 0.059, 0.098, 0.165, 0.312, and 0.471 for $\epsilon_d = 20$. For the charged sequences $10\text{Y3D}$, $5\text{Y5D}$, $10\text{Y3R}$, $10\text{Y5R}$, $8\text{YAD}$, $6\text{YAD}$, and $8\text{Y2RAD}$, the simulated $P_b$ values are, respectively, 0.370, 0.003, 0.070, 0.005, 0.057, 0.017, and 0.019 for $\epsilon_d = 40$ and 0.386, 0.003, 0.084, 0.006, 0.049, 0.021, and 0.025 for $\epsilon_d = 20$. Taken together, results from our control simulations suggest that the trends predicted by our model should be robust inasmuch as the EAD does not undergo a hydrophobic collapse and that cation-$\pi$ interactions are significantly stronger than salt-bridge interactions as stipulated by theoretical considerations [16].
Conformational sampling. Monte Carlo sampling of EAD conformations was conducted at \( T = 300\text{K} \) in a \( 600\times600\times600\text{Å} \) simulation box with periodic boundary conditions and the target fixed at its center. The EAD is defined to be bound if at least one aromatic on it is within a capture radius \( R_c = 6.0\text{Å} \) from a target cation (Fig. S2D). Binding probabilities were computed accordingly. Four types of chain moves were used for conformational sampling with equal attempt probabilities: rigid rotations, pivot moves, kink-jumps [24,25] and translation moves. In a translation move, a random direction and a random distance \( \leq 1\text{Å} \) is selected. This random vector of displacement is then applied to every bead of the chain. All four types of attempted chain moves were accepted or rejected by applying the standard Metropolis criterion [26] to the total energy \( E_T = E_{\text{intrachain}} + E_{\text{chain-target}} \) in Eqs. (S1) and (S2). The acceptance rate is \(~50\%\). In a typical binding simulation with the generic target, \(~20\text{ million attempted chain moves were used for initial equilibration and data collection} \) was conducted during the subsequent \(~80\text{ million attempted chain moves}. \) For each of the binding simulation with an IDP target (see below), \(~40\text{ million attempted chain moves were used for initial equilibration and} \) \(~40\text{ million} \) subsequent attempted chain moves were used for data collection. Because the excluded volume and hydrophobic terms are of short spatial range, a \( 10.0\text{Å} \) cutoff was applied to the \( \epsilon_{\text{ex}} \) and \( \epsilon_{\text{hp}} \) terms in Eq. (S1) and a \( 2\epsilon_0 + R_p = 19.8\text{Å} \) cutoff was applied to the \( \epsilon_{\text{ex}} \) term in Eq. (S2) for computational efficiency.

Radii of gyration of the model EAD chains.

For the \( 4Y_n-10Y_n \) EAD sequences in Fig. 1, the simulated mean radii of gyration (using \( \epsilon_{\text{hp}} = -3.0 \ k_BT \) and \( \epsilon_0 = 40 \)) for unbound EADs are \(~21\text{Å} \), whereas the simulated mean radii of gyration for bound EADs vary slightly from \(~22.5\text{Å} \) for \( 4Y_n \) to \(~20.3\text{Å} \) for \( 10Y_n \). For the EAD sequences with positively charged Rs in Fig. 3 of main text, intra-EAD cation–π contacts lead to more compact unbound conformations. For 8Y2RAD (with two Rs), 10Y3R (with three Rs), and 10Y5R (with five Rs), the simulated mean radii of gyration are, respectively, 17.6, 15.6, and 12.1Å. In contrast, the simulated mean radii of gyration for 10Y3D and 8YAD (with no R) are 21.7 and 21.1Å, respectively, which are practically identical to the \(~21\text{Å} \) mentioned above for the \( 4Y_n-10Y_n \) sequences in Fig. 1.

Matching chain simulation results with experimental data. The chain simulation model described above was applied to analyze the experimental activity measurements reported in the main text. In addition, the model was also used to re-analyze earlier experimental activity data [1] (Fig. S3), to address the similarities and differences in the binding properties of monomer and dimer EADs (Fig. S4), to motivate our analytical model (discussion below and Figs. S5–S7), and to study possible association of EAD with IDP targets (Fig. S8) including the RGG3 sequence in the Ewing’s sarcoma RNA-binding domain [27,28]. Although how precisely real EAD binding triggers specific functional events is not known, we recognize that the energetics of oncogenic EAD processes is unlikely to comprise solely of the interactions included in our model. For instance, a certain energetic contribution can be associated with a specific functional event, similar to the contact energy \( E_b \) in the polyelectrostatics model [29]. In the approach adopted here, we assume as a first approximation that this unknown energy (let us denoted it as \( E_{b,EAD} \)) is a constant irrespective of the bound EAD conformation. Because we did not include such an energy in our model, the binding free energy \( \Delta G_b = -k_BT\ln\{P_b/(1-P_b)\} \) in our model (Fig. 2D and Fig. S4) was computed without regard to \( E_{b,EAD} \). In view of this limitation, the \( \Delta G_b \) values in the present work were used to address only relative, but not absolute, activities of real EAD, because the true binding free energy could have been \( \Delta G_b + E_{b,EAD} \).

Rationale and constructional details of the analytical model

Overall goal, strengths and limitations. The analytical model in this work [Eq. (1) in the main text] was developed as a complement to the chain simulation. Our goal in developing the analytical model is to provide further insights into the trends observed in simulations and experiments. The analytical model addresses multi-site EAD binding by considering the balance among the energetic contributions from cation–π contacts, translational and conformational entropies of the EAD as well as intra-EAD and EAD-target excluded-volume effects. For simplicity, hydrophobic and electrostatic interactions are not incorporated in the present analytical model. As a model for chain behavior, the analytical model
lacks an explicit representation of the polymer chain and thus is less accurate than the chain simulation model. Despite its reliance on approximations, the analytical model is valuable because it offers conceptual clarity and computational efficiency. Its tractability allows for efficient exploration of model parameters and, therefore, a more comprehensive assessment of the robustness of the model’s predictions.

Conformational entropic effects of EAD binding. Multisite binding of EAD entails large changes in its conformational ensemble upon binding. Therefore, unlike the mean-field poly-electrostatic model for the Sic1-Cdc4 system [29] that assumes no significant changes in the conformational shape of the IDP ligand upon binding, we now need to estimate the change in conformational entropy upon binding of an EAD to its target. As outlined in the main text, in order to afford a rudimentary account of excluded volume effects on conformational freedom, we adopted exact lattice enumeration to assist in the necessary entropy estimations. Exact enumeration of lattice conformations is a powerful and versatile technique that has contributed to fundamental advances in polymer physics and in the study of protein structure and stability [30–32].

A simple model of the EAD sequence. All ingredients of the analytical model have been introduced in the main text. Again, for simplicity, the present formulation of the analytical model considers only EAD sequences with \( N \) equally spaced aromatic residues (\( k \) bonds between two successive aromatics), as illustrated by Fig. S5A. If necessary, this restriction can be relaxed as the model can readily be generalized to tackle EAD sequences with any set of aromatic positions. In our analytical model, the aromatic positions along the EAD define a set of possible loops of EAD chain segments by having two or more EAD aromatics contacting the cations on the target. The conformational entropy of such a loop is determined by the length \( l \) of the loop and the distance \( R \) between the two cation-aromatic contacts. We used the geometry of the generic target in the chain simulation model to determine the distribution \( n_{t}(R) \) for \( R \) (Fig. S5B).

Lattice estimation of conformational entropy. Exact enumeration of conformations in the presence of an impenetrable infinite planar surface [33] was utilized to provide a general approximation of the loop conformational entropy that we can readily apply to targets with different geometries. We first obtained the numbers of conformations (self-avoiding flights) configured on a simple cubic lattice that are (i) subject to no constraint, i.e., it is free to configure on an infinite lattice subject only to the condition that it cannot intersect itself, (ii) constrained to have one end of the chain contacting the impenetrable plane, as illustrated in Fig. S5C, and (iii) constrained to have the middle of the chain contacting the impenetrable plane, as illustrated in Fig. S5D. We denote the number of such conformations, as a function of chain length \( n \), by \( \Omega(n) \), \( \Omega_{\text{L}}(n) \), and \( \Omega_{\text{m}}(n) \), respectively. These conformational counts are given in Table S1 for \( n = 4 \) to \( n = 17 \). The quantity \( \Omega(n) \) has been studied extensively before. The \( \Omega_{\text{L}}(n) \) counts in Table S1 are consistent with an early \( n \leq 16 \) enumeration by Sykes [34] and a more recent enumeration by Clisby et al [35]. More specifically, our \( \Omega_{\text{L}}(n) \) is equivalent to the coefficient for \( x^{n-1} \) in Eq. 2.1 of [34]. \( \Omega_{\text{L}}(17) \) was also provided in this reference but the coefficient 100,117,875,366 for \( x^{16} \) was incorrect. Our \( \Omega_{\text{L}}(n) \) corresponds to the variable \( c_{n-1} \) in Table A5 of [35], which provides \( \Omega_{\text{L}}(n) \) for \( n \leq 31 \). Note, however, that the variable \( n \) in Clisby et al [35] is the number of bonds and thus is equivalent to our \( n - 1 \).

The quantity \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) represents the change in conformational entropy, in units of \( k_B \), upon bringing a free, unconstrained chain to the vicinity of the impenetrable surface and making a first contact with the surface at a chain end. Likewise, \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) represents a similar entropy change but the first contact with the plane is made at mid-chain. For random flights, the corresponding conformational entropy change scales as \(- (\ln n)/2\) irrespective of which point along the chain makes the contact with the plane [36]. However, for self-avoiding flights, the difference between \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) and \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) is significant and increases with \( n \) (Fig. S5E). In the present formulation of the analytical model, we used \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) to provide a general approximation for the conformational entropy change upon the formation of the first chain-plane contact [see Eq. (1) of main text]. The reason for this choice is that for the chain lengths we studied, a chain-plane contact is more likely to be sufficiently far away from the chain ends to be better represented by a mid-chain contact rather than a chain-end contact. Accordingly, in binding free energy calculations using Eq. (1) of main text, \( \ln[\Omega_{\text{L}}(n)/\Omega_{\text{L}}(n)] \) was determined using the data in Table
S1 for $n \leq 17$ and estimated for $n > 17$ by extrapolating the fitting equation for $\ln[\Omega^m_a(n)/\Omega_b(n)]$ provided in the caption for Fig. S5.

After the first EAD-target contact has been made, EAD loops can form on the target surface, leading to further reduction in conformational entropy. We estimated such entropy reduction by enumerating $\Omega(l, R_j|n)$, which is the number of conformations with one chain end anchored to the impenetrable surface while a loop of length $l$ is formed by a second contact at a distance $R_j$ from the anchored chain end (Fig. S6, top left drawing). Examples of such conformational counts are provided in Table S2; and a complete listing of $\Omega(l = n - 1, R_j|n)$ counts for $n = 17$ is included in Table S3 to illustrate our method. Values for $\ln[\Omega(l, R_j|n)/\Omega^m_a(n)]$ from $n = 4$ through $n = 17$ were then grouped by loop length $l$ and plotted in Fig. S6. Recognizing that $\ln[\Omega(l, R_j|n)/\Omega^m_a(n)]$ for a given $l$ is not very sensitive to chain length $n$, we obtained quadratic fits for $\ln[\Omega(l, R_j)/\Omega^m_a]$ in the form of $-a(l)[R_j - b(l)]^2 + c(l)$ for $l \leq 16$ (caption of Fig. S6). We then extrapolated the fitting parameters $a(l), b(l)$, and $c(l)$ for $l > 16$ (Fig. S7A,B,C) to obtain the necessary $\Omega(\mathbf{k}l_i, R_j|n)/\Omega^m_a(n)$ values (now approximated as independent of $n$) that enter Eq. (1). It should be noted here that $R_j$ was measured in units of lattice bond length in the enumeration data, and that the lattice bond length is taken to be equivalent to the $C_\alpha-C_\alpha$ virtual bond length $b = 3.8\AA$ in our analysis. It follows that the $R_j$ values in the $n_\alpha(R_j)$ distributions in Fig. S5B have to be converted to lattice units ($R_j \rightarrow R_j/b = R_j/3.8\AA$) before they enter Eq. (1).

**Robustness of the predicted binding free energies.** For the EAD sequences studied in this work, we found that the binding free energy is not very sensitive to the functional form of $\ln[\Omega(l, R_j)/\Omega^m_a]$ for large $l$. To evaluate this sensitivity, we have compared binding free energies calculated using the above procedure and one that used exact enumeration $\ln[\Omega(l, R_j)/\Omega^m_a]$ for $l \leq 16$ but substituted $\ln[\Omega(l, R_j)/\Omega^m_a]$ with the random-flight expression $\ln[\Omega(l, R_j|n)/\Omega_b(n)] = (3/2) \ln(3/2m) - \ln(l - R_j^2/l)$ for $l > 16$. [The latter expression follows from the random-flight probability $(3/2\pi r)^{3/2} \exp(-3R_j^2/2l)$ for a chain of length $l$ starting from the origin and ending at a position that is at a distance $R_j$ from the origin.] Despite the appreciable difference between the two entropy expressions for large $l$ (Fig. S7D), the calculated binding free energies using the two different schemes are nearly identical for the set of sequences tested in Fig. S7E.

**Energetic and entropic components of the binding free energy.** $\Delta G_b$ can readily be expressed as a sum of an energy and an entropy, with the binding energy (enthalpy) given by

$$
\Delta E_b = \frac{E_{\text{ent}}}{Q_b} \left\{ N_k + \sum_{n_{\text{loop}} + 1} n_{\text{loop}} + 1 \right. \\
\times \prod_{l} e^{-E_{\text{ent}}/k_B T} \sum_{j} n_c(R_j) \\
\times \left[ \frac{\Omega(\mathbf{k}l_i, R_j|n)}{\Omega^m_a(n)} \right] \right\}
$$

(S3)

where $n_{\text{loop}} = \sum_i 1$ is the number of loops and $n_{\text{loop}} + 1$ is the number of cation-π contacts, and

$$
Q_b = N_k + \sum_{l} \prod_{l} e^{-E_{\text{ent}}/k_B T} \\
\times \sum_{j} n_c(R_j) \left[ \frac{\Omega(\mathbf{k}l_i, R_j|n)}{\Omega^m_a(n)} \right] .
$$

(S4)

After $\Delta E_b$ has been determined, the binding entropy $\Delta S_b$ can be calculated using the standard relation

$$
T \Delta S_b = \Delta E_b - \Delta G_b ,
$$

(S5)

where $\Delta G_b$ is given by Eq. (1) in the main text. These expressions were used to compute the $\Delta E_b$ and $T \Delta S_b$ values in the inset of Fig. 2A.

**Possible interference among multiple Ys interacting with same cation**

Our present simulation model is seen to overestimate the affinity of 5YP in Fig. 4 of the main
text. The experimental activity of 5YP is approximately the same as that of 10YN, but the simulated \( P_b \) for 5YP is more than double that of 10YN. To address this mismatch, we note that in the present formulation of our model, two sequentially adjacent Ys are assumed to be able to interact strongly and simultaneously with the same cation as if the two Ys were far apart along the sequence and interacting with different cations. But in reality, the two adjacent Ys would most likely interfere with each other, resulting in weakened individual interactions with the same cation, as cation-\( \pi \) interactions are strongly orientation dependent [37]. This issue did not arise for the other EAD sequences we have simulated in this work because the individual Ys are well separated in those sequences. As a first attempt to explore this issue, we have performed additional simulations using a modified model in which the well depth for an individual cation-Y interaction is reduced from the full strength of 3.58 kcal/mol (Fig. 1B) when the given cation is interacting with two or more Ys. By considering a class of such models we found that if the reduced individual well depth is 2.86 kcal/mol for multiple Y contacting the same cation (representing a \( \approx 20\% \) reduction), the simulated binding probability for 5YP (\( P_b = 0.46 \)) would be similar to that for 10YN (\( P_b = 0.43 \)). To account for the behavior of the 5YP sequence in the analytical model, we considered a model with \( N_s = 5 \) and \( k = 12 \), but with \( E_{\text{cen}} \) replaced by an energy \( E_{\text{cen}}^{(2)} \) for the combined cation-\( \pi \) interaction energy when a pair of Y’s contact a cation simultaneously. If we take \( E_{\text{cen}}^{(2)} = 2E_{\text{cen}} = 2(-3.5k_B T) = -7.0 \) \( k_B T \), 5YP is predicted by our analytical model to bind much more tightly (\( \Delta G_b = -11.7k_BT \)) than 10YN (\( \Delta G_b = -3.2k_BT \)). However, if \( E_{\text{cen}}^{(2)} = -5.2 \) \( k_B T \) (\( \approx 74\% \) of \( 2E_{\text{cen}} \)), the corresponding binding free energy \( \Delta G_b \approx -3.1k_BT \) for 5YP is similar to that for 10YN as observed in the activity experiments. \( E_{\text{cen}}^{(2)} \) is slightly weaker for the open circle data point that was included in Fig. 4 of the main text as an example (\( E_{\text{cen}}^{(2)} = -5.1 \) \( k_B T \), \( \approx 73\% \) of \( 2E_{\text{cen}} \)), resulting in slightly weaker binding with \( \Delta G_b \approx -2.6k_BT \). These considerations indicate that an interference effect between two aromatics contacting the same cation that amounts to a \( \approx 20-30\% \) moderate reduction in the individual cation-\( \pi \) interaction strengths would be sufficient to provide a quantitative account for the experimental activity of 5YP in the context of a polycation-\( \pi \) model.

**Supporting References**


Supporting Figures

Figure S1. Proteins and EAD sequences used in the present study. Transcriptional activator proteins (Top) contain the experimental sequences related to the N-terminal 66 residues of EAD1-66 (box with purple Ys), the region of ATF1 protein (ΔATF1) present in the EWS/ATF1 oncogene and the DNA-binding domain of zta protein (ztaDBD). In (A)–(C), amino acid residues are denoted by the standard one-letter code. Sequences for Figs. 1, 3, and 4 in the main text are listed, respectively, under (A), (B) and (C).
**Figure S2. The chain simulation model.** (A) The generic EAD binding target (partner) is a sphere of radius $R_p = 16\text{Å}$ with essentially evenly distributed positive and negative charges (represented by blue and red beads respectively). (B) An EAD sequence is modeled as a $C_\alpha$ chain (beads on a string) that can engage in cation-$\pi$, electrostatic, hydrophobic, and excluded-volume interactions as specified in the main text and Text S1. In this figure and subsequent supporting figures, aromatic (Y in this drawing) and hydrophobic (hφ) residues are shown in magenta and orange, respectively, whereas positively and negatively charged residues are shown in blue and red respectively. All other residues are shown in grey. (C) The distribution of positively charged residues on the heterodimer of the Rpb4/Rpb7 subunits of human RNA polymerase II was used as a reference for the design of the charge density on the generic EAD binding target. The histogram here shows the shortest distance from each of the 32 positively charged amino acid residues (R or K) on Rpb4/Rpb7 (16 each along the Rpb4 and Rpb7 chains) from another positively charged residue, based on the X-ray crystal structure (PDB ID: 2C35) determined by Meka et al. (ref. [10] of Text S1). The distances are measured between the atoms that have the positive charges. The red dashed horizontal line marks the average shortest distance which is $\approx 9.4\text{Å}$. (D) EAD-target binding is defined in the model as having at least one EAD aromatic residue (magenta circle) within a capture radius $R_c = 6\text{Å}$ from a positive charge (blue circle) on the target. One such cation-$\pi$ contact between an EAD sequence (brown string connecting magenta circles) and the target (large shaded circle with embedded blue circles) is shown in this schematic.
(E, F) Energetic components of the interaction potential, the horizontal variable \( r \) here corresponds to \( r_{ij} \) in Eq. (S1) or \( r_{i\nu} \) in Eq. (S2). (E) Model cation-\( \pi \) interaction potentials in the form of
\[ \varepsilon_{ij}^{\alpha} \left[ \left( \frac{\sigma_{\alpha i}/r_{ij}}{\sigma_{\alpha j}/r_{ij}} \right)^{12} - \left( \sigma_{\alpha i}/r_{ij} \right)^6 \right] \] or \[ \varepsilon_{i\nu}^{\alpha} \left[ \left( \frac{\sigma_{\alpha i}/r_{ij}}{\sigma_{\alpha \nu}/r_{i\nu}} \right)^{12} - \left( \sigma_{\alpha i}/r_{ij} \right)^6 \right] \] in Eqs. (S1) and (S2) respectively [i.e., equivalent to Fig. 1B in the main text minus the \( \varepsilon_{ex}(r_{rep,i}/r_{ij})^{12} \) term]. The green and blue curves show the potentials for cation-W and cation-Y, respectively, as in Fig. 1B, whereas the red curve corresponds to the weakest among the model cation-F interactions considered in Fig. 1B. (F) Total interaction potential between hydrophobic residues and between charged residues in the simulation chain model, including their respective excluded-volume interactions. Solid curves show potential functions used for all simulation results presented in this work except specifically noted otherwise. Dashed curves show alternative potential functions that we have used for the control simulations reported in Text S1. The potential functions used for hydrophobic interaction are shown in magenta. The solid curve is for hydrophobic interaction strength \( \varepsilon_{\text{hp}} = -3.0 \text{ kT} \) [Eq. (S1)] whereas the dashed curve is for \( \varepsilon_{\text{hp}} = -7.0 \text{ kT} \). The potential functions for electrostatic interactions between like charges and between opposite charges are shown, respectively, in red and blue. The solid curves are for \( \varepsilon_{d} = 40 \) whereas the dashed curves are for \( \varepsilon_{d} = 20 \).

Figure S3. Evidence for the polycation-\( \pi \) hypothesis from a re-analysis of early experiments on 33-residue EAD sequences. Sequences and experimental data were taken from ref. [1] of Text S1. Simulations were conducted using the same chain model as described in Text S1 and the main text in a (600Å)\(^3\) simulation box. (A) The sequences are defined in the above reference. The experimental relative activities and the simulated relative binding probabilities are represented by the black and grey bars respectively. (B) The sequences in (A) are grouped according to their Y number \( n_Y \). Plotted are the simulated binding probability (solid squares) and the relative experimental activity (open circles) averaged over sequences belonging to each given \( n_Y \). For the simulation results, the averages are over all possible permutations of Y positions for a given \( n_Y \), including those not studied by experiments. Note that both Y number and Y density are varied among this set of sequences (unlike the set in Fig. 1 that varies only the Y number while keeping Y density constant). Error bars show variation among sequences with the same \( n_Y \). Lines joining the solid squares are merely a guide for the eye.
Figure S4. Simulated binding probabilities of monomer and dimer EAD sequences follow similar trends. Similar dependences on \( n_Y \) are observed for cis-duplication of small EAD elements in a single dimer. The monomer sequences used in the present simulations are the same 33-residue sequences based on the construction by Feng and Lee (ref. [1] of Text S1) studied in Fig. S3. As for the simulations in Fig. S3, all possible permutations of Y positions are considered. Each dimer was constructed by joining the C-terminus of a given monomer sequence to the C-terminus of another copy of the same monomer sequence by a linker chain. The linker consists of six residues that are neither charged nor hydrophobic; all reference bond angles within the linker are equal to 165° with a stiff bond-angle force constant equal to 10.0\( k_B T \). Thus, in this figure, a dimer sequence with \( Y \) number 2\( n_Y \) is equivalent to two identical monomer sequences with \( Y \) number \( n_Y \) connected by such a linker. (A) A snapshot of an \( n_Y = 5 \) monomer bound to the target. (B) A snapshot of the corresponding \( n_Y = 10 \) dimer bound to the target. The EAD chains are depicted in a tube representation with the color code for different residue types specified in Fig. S2B. (C) Free energies of binding were computed under the same conditions as those used for Fig. S3. \( \Delta G_b \) values averaging over sequences with the same \( n_Y \) are plotted.
Figure S5. Components of the analytical model. (A) Schematic of cation-π contacts along the IDP. Here we only consider IDP chains with evenly spaced aromatics that are $k$ residues apart; thus the contour length between two cation-contacting aromatics is always in the form of $kl$, where $l$ is a positive integer. Three example contact patterns are shown, wherein the aromatics and cations are depicted as magenta and blue circles respectively. (B) Distribution of cation-cation distance $R_j$ on the target. Each $R_j$ value is the distance in Å from a given cation to a different cation, measured on the spherical surface of the model target (left drawing). The distribution $n_c(R_j)$ is shown (histograms) for three different targets of the same size but different cation densities. As for the target with $N_c = 32$ cations in most of our simulations, the cations are essentially evenly distributed on the surface for the $N_c = 8$ and $N_c = 96$ targets. The approximately even distribution of charges on the target sphere was achieved by a numerical algorithm (see Text S1). As can be seen from the histograms, only a few of the $R_j$ values are exactly identical. (C) An example conformation configured in the simple cubic lattice with one end of the chain touching a plane. The number of such conformations is referred to as $\Omega_{a0}(n)$ in this work. (D) An example simple cubic lattice conformation with two of its mid-chain sites in contact with a plane. We denote the number of such conformations as $\Omega_{am}(n)$. (E) Change in conformational entropy (in units of the Boltzmann constant $k_B$) upon bringing a free lattice conformation to form a contact at a chain end (squares) or at mid-chain (circles) with an infinite impenetrable plane that imposes excluded volume on the other side of the plane (the space underneath the plane is not accessible to the chain). The data points (squares or circles) were computed using exact enumeration data in Table S1. The curves through the data points were generated by fitting the assumed relation $\gamma = \ln[A\exp(-\omega n) + B\exp(-\sigma n)]$. The fitting parameters here are $A = 0.5365, B = 0.53139, \omega = 0.02786$, and $\sigma = 0.33604$ for $\gamma = \ln[\Omega_{a0}^0(n)/\Omega_0(n)]$; and $A = 0.40915, B = 1.12627, \omega = 0.05373$, and $\sigma = 0.39353$ for $\gamma = \ln[\Omega_{am}^0(n)/\Omega_0(n)]$. 
Figure S6. **Conformational entropy loss upon loop formation.** The quantity $\Omega(l, R_j|n)$ is the number of simple cubic lattice conformations of length $n$ ($n$ is the total number of beads along the chain) that have one chain end (bead number 1) touching an excluded-volume plane at a given point (as in Fig. S5C) and, at the same time, bead number $l + 1$ also making a contact with a given point on the plane at a distance $R_j$ from where bead number 1 touches the plane, thus forming a loop of length $l$ that spans a distance $R_j$ on the plane (top left drawing). Note that conformations that form other chain-plane contact(s) in addition to these two are included in the $\Omega(l, R_j|n)$ count. As discussed in the main text and in Text S1, the vertical variable $\ln\left[\Omega\left(l, R_j|n\right)/\Omega_{a}^{m}(n)\right]$ for the plots in this figure corresponds approximately to the conformational entropy change, in units of $k_B$, upon making an additional chain-plane contact to form a loop of length $l$ along a chain that has already made at least one contact with the plane. Each of the plotting panels provides the conformational entropy change upon forming a loop of a given length $l$ as a function of $R_j$. Both $l$ and $R_j$ are shown in units of the lattice bond length (nearest distance between two beads on the simple cubic lattice). Data points (open circles) in the plotting panels were computed by exact enumeration of lattice conformations with chain lengths from $n = 4$ through $n = 17$ (see Text S1 and Tables S2 and S3). Multiple data points for the same $R_j$ value represent results from different $n$ values. The continuous curves are quadratic fits in the form of $\ln\left[\Omega\left(l, R_j|n\right)/\Omega_{a}^{m}(n)\right] = -a(l)\left[R_j - b(l)\right]^2 + c(l)$. The $l$-dependent fitting parameters $a(l), b(l),$ and $c(l)$ are provided in Fig. S7. In view of the clustering of data points from different $n$ values, we have made an approximation in the analytical model that $\ln\left[\Omega\left(l, R_j|n\right)/\Omega_{a}^{m}(n)\right]$ is independent of $n.$
Figure S7. Applying the lattice conformational entropy estimates to the analytical model. (A–C) The fitting parameters $a(l)$, $b(l)$, and $c(l)$ for the conformational entropy changes shown in Fig. S6 are provided as data points in (A), (B), and (C), respectively. The continuous fitting curves are given by (A) $a(l) = A + B \exp(-C l)$, where $A = 0.13748$, $B = 7.04181$, and $C = 0.52115$; (B) $b(l) = A + B \ln(C l)$, where $A = 0.97499$, $B = 0.93564$, and $C = 0.97495$; and (C) $c(l) = A + B \exp[-C(l - D)]$, where $A = -5.19530$, $B = 2.98286$, $C = 0.31975$, and $D = 2.79004$. These expressions were used to estimate $\ln[\Omega(l, R_j | n) / \Omega_m(n)]$ for $l > 16$ by extrapolation. (D) The extrapolated $\ln[\Omega(l, R_j | n) / \Omega_m(n)]$ function (black curve) is compared against the corresponding random-flight expression $\ln[(3/2\pi l)^{3/2} \exp(-3R_j^2/2l)]$ (red dashed curve) for $l = 60$. (E) Two methods for estimating the entropic cost of loop formation in the analytical model are compared. Plotted are the binding free energies of the model EAD chains in Fig. 1 for $E_{c\pi} = -3.5 k_B T$. The black data points (circles) were computed by using entropy estimates from exact enumeration for $l \leq 16$ and extrapolated estimates for $l > 16$, whereas the red data points (triangles) were obtained by using entropy estimates from exact enumeration for $l \leq 16$ but random-flight estimates for $l > 16$. The plot here shows that the predicted $\Delta G_b$ values based on the two different loop entropy estimates are very similar.
Figure S8. Exploring other EAD-target binding scenarios. The EAD sequences are the same as those in Fig. 1. (A) Simulated EAD binding probability $P_b$ with a hypothetical target in which the surface charges are not evenly distributed but confined to a patch. Two such hypothetical patch partners were considered, both with 12 cations localized on a patch with the same local cation density as the generic target with 32 cations (Fig. S2A) that we have used for most of the simulations. One of the targets (referred to as the positive patch target) contains 12 cations and no anions on the patch whereas the other (referred to as the neutral patch target) contains 12 cations and 12 anions. Plotted here are the simulated binding probabilities for the positive (squares) and neutral (circles) patch targets in either a simulation box of size of $(300\text{Å})^3$ (black symbols) or $(600\text{Å})^3$ (blue symbols). (B) A snapshot of an $n_Y = 10$ EAD sequence (tube representation) bound to the neutral patch target. (C) Simulated EAD binding probability $P_b$ with hypothetical disordered (IDP) partners. The EAD sequences and simulation conditions are the same as those in Fig. 1B,C, using a simulation box of size $(600\text{Å})^3$. During the binding simulations, both the EAD and the hypothetical IDP target were allowed to sample all accessible conformations while the center of mass of the IDP target was kept at a fixed position in the center of the simulation box. We considered a class of such targets, each of which is a chain consisting of 64 alternating cations and anions (32 cations and 32 anions). The adjacent cation and anion are connected by a 5Å virtual bond with a stiff bond-angle force constant equal to $10.0k_BT$. Shown here are binding probabilities for four different such IDP targets with equilibrium bond angles that equal, respectively, to $105^\circ$ (crosses), $120^\circ$ (diamonds), $135^\circ$ (squares) and $150^\circ$ (circles). A general trend of increasing binding with increasing $n_Y$ is observed for all four hypothetical IDP targets. Not surprisingly, the quantitative details of this trend are sensitive to the persistence length of the IDP target. Binding increases with the flexibility of the IDP target. Also included for comparison (blue triangles) are the simulated probabilities of EAD binding with the RGG3 sequence in the Ewing’s sarcoma RNA-binding domain GGDGRGGPFGGMRGRGGGLMDRGPGMGFRGGRGGGDRGGRGGRGMDRGPGFMFRGGRGGGDRGGRGGRGGGG (refs. [27,28] in Text S1). Here the RGG3 sequence was modeled as a $C_n$ chain using the same modeling scheme as that for the EAD sequences. (D) A snapshot of an $n_Y = 10$ EAD sequence (tube representation) bound to a hypothetical IDP target (red and blue beads) with $150^\circ$ bond angles.
Supporting Tables

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Omega_0(n)$</th>
<th>$\Omega_0^0(n)$</th>
<th>$\Omega_0^m(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>150</td>
<td>93</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>726</td>
<td>409</td>
<td>337</td>
</tr>
<tr>
<td>6</td>
<td>3,534</td>
<td>1,853</td>
<td>1,433</td>
</tr>
<tr>
<td>7</td>
<td>16,926</td>
<td>8,333</td>
<td>5,937</td>
</tr>
<tr>
<td>8</td>
<td>81,390</td>
<td>37,965</td>
<td>25,809</td>
</tr>
<tr>
<td>9</td>
<td>387,966</td>
<td>172,265</td>
<td>110,369</td>
</tr>
<tr>
<td>10</td>
<td>1,853,886</td>
<td>787,557</td>
<td>486,049</td>
</tr>
<tr>
<td>11</td>
<td>8,809,878</td>
<td>3,593,465</td>
<td>2,118,369</td>
</tr>
<tr>
<td>12</td>
<td>41,934,150</td>
<td>16,477,845</td>
<td>9,427,777</td>
</tr>
<tr>
<td>13</td>
<td>198,842,742</td>
<td>75,481,105</td>
<td>41,662,809</td>
</tr>
<tr>
<td>14</td>
<td>943,974,510</td>
<td>346,960,613</td>
<td>186,303,561</td>
</tr>
<tr>
<td>15</td>
<td>4,468,911,678</td>
<td>1,593,924,045</td>
<td>828,799,641</td>
</tr>
<tr>
<td>16</td>
<td>21,175,146,054</td>
<td>7,341,070,889</td>
<td>3,725,715,541</td>
</tr>
<tr>
<td>17</td>
<td>100,121,875,974</td>
<td>33,798,930,541</td>
<td>16,682,103,329</td>
</tr>
</tbody>
</table>

Table S1. Numbers of conformations, or self-avoiding flights, on the simple cubic lattice. Conformational counts as functions of chain length (number of beads) $n$ are obtained by exact enumeration. A chain with $n$ beads has $n - 1$ bonds. Here, $\Omega_0$ is the number of unconstrained conformations; $\Omega_0^0$ is the number of conformations that have one chain end anchored onto an impenetrable plane (Fig. S5C); and $\Omega_0^m$ is the number of conformations that have the mid-chain bead [(n/2)th bead if $n$ is even, ((n+1)/2)th bead if $n$ is odd] making a contact with an impenetrable plane (Fig. S5D).
Table S2. Loop probabilities determined by exact lattice conformational enumeration. Tabulated here are examples (not a complete list) of conformational counts \( \Omega(l, R_j|n) \) used in Fig. S6. Here one chain end is always in contact with the origin (0,0) of a two-dimensional coordinate system for the impenetrable plane. In this table, the positions on the impenetrable plane where another contact with the chain existed are indicated by the \((x, y)\) coordinates. In the present treatment of our analytical model, \( R_j \) values from all combinations of \( x, y \) (where \( x < y \)) that have nonzero \( \Omega(l, R_j|n) \) counts for \( n \leq 17 \) were used to estimate the conformational entropic cost of loop formation (Figs. S6 and S7).

<table>
<thead>
<tr>
<th>( x, y )</th>
<th>( R_j )</th>
<th>( l = 3 )</th>
<th>( l = 4 )</th>
<th>( l = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 4 )</td>
<td>( n = 5 )</td>
<td>( n = 6 )</td>
<td>( n = 5 )</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>38</td>
</tr>
<tr>
<td>0.3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>1.2</td>
<td>( \sqrt{3} )</td>
<td>6</td>
<td>24</td>
<td>98</td>
</tr>
<tr>
<td>0.2</td>
<td>2</td>
<td></td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>0.4</td>
<td>4</td>
<td></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>( \sqrt{10} )</td>
<td></td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>0.3</td>
<td>3</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>0.5</td>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>( \sqrt{5} )</td>
<td></td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>1.4</td>
<td>( \sqrt{17} )</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>( \sqrt{13} )</td>
<td></td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

Table S3. Exact lattice enumeration data for loop formation probability. Tabulated here as examples are the exact \( \Omega(l, R_j|n) \) counts for \( l = 16 \) and \( n = 17 \). The horizontal and vertical labels correspond, respectively, to the \( x \) and \( y \) coordinates of the positions on the impenetrable plane. One end of the chain (first bead) is always anchored at the origin (0,0). In this table, the entry at a given position \((x, y)\) is the number of conformations that have the chain’s last \((n^{th})\) bead contacting the given position and thus making a loop with \( R_j = \sqrt{x^2 + y^2} \). Data are shown only for \( x \leq y \) because of the obvious rotational symmetry.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62,008,495</td>
<td>0</td>
<td>56,803,594</td>
<td>0</td>
<td>27,780,947</td>
<td>0</td>
<td>6,593,054</td>
<td>0</td>
<td>639,838</td>
<td>0</td>
<td>25,060</td>
<td>0</td>
<td>315</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>13,944,500</td>
<td>0</td>
<td>128,368,232</td>
<td>0</td>
<td>83,248,596</td>
<td>0</td>
<td>28,370,488</td>
<td>0</td>
<td>4,339,452</td>
<td>0</td>
<td>261,698</td>
<td>0</td>
<td>5,492</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>131,632,780</td>
<td>0</td>
<td>102,566,344</td>
<td>0</td>
<td>46,804,240</td>
<td>0</td>
<td>10,178,096</td>
<td>0</td>
<td>911,232</td>
<td>0</td>
<td>28,658</td>
<td>0</td>
<td>240</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>108,903,312</td>
<td>0</td>
<td>60,835,620</td>
<td>0</td>
<td>18,085,904</td>
<td>0</td>
<td>2,278,472</td>
<td>0</td>
<td>101,192</td>
<td>0</td>
<td>1,120</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>66,318,904</td>
<td>0</td>
<td>24,958,578</td>
<td>0</td>
<td>4,273,524</td>
<td>0</td>
<td>257,894</td>
<td>0</td>
<td>3,640</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>27,752,872</td>
<td>0</td>
<td>6,179,844</td>
<td>0</td>
<td>692,092</td>
<td>0</td>
<td>8,736</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>6,976,902</td>
<td>0</td>
<td>720,434</td>
<td>0</td>
<td>16,016</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>816,816</td>
<td>0</td>
<td>22,880</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>25,740</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>