Soft Matter

Role of complementary shape in protein dimerization

Journal:	*Soft Matter*
Manuscript ID	SM-ART-03-2021-000468.R1
Article Type:	Paper
Date Submitted by the Author:	03-Jul-2021
Complete List of Authors:	Gao, Fengyi; University of Michigan, Chemical Engineering Glaser, Jens; University of Michigan, Chemical Engineering Glotzer, Sharon; University of Michigan, Chemical Engineering

Soft Matter

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The role of complementary shape in protein dimerization

Fengyi Gao,[a] Jens Glaser,[‡a] and Sharon C. Glotzer[*ab]

Shape guides colloidal nanoparticles to form complex assemblies, but its role in defining interfaces in biomolecular complexes is less clear. In this work, we isolate the role of shape in protein complexes by studying the reversible binding processes of 46 protein dimer pairs, and investigate when entropic effects from shape complementarity alone are sufficient to predict the native protein binding interface. We employ depletants using a generic, implicit depletion model to amplify the magnitude of the entropic forces arising from lock-and-key binding and isolate the effect of shape complementarity in protein dimerization. For 13% of the complexes studied here, protein shape is sufficient to predict native complexes as equilibrium assemblies. We elucidate the results by analyzing the importance of competing binding configurations and how it affects the assembly. A machine learning classifier, with a precision of 89.14% and a recall of 77.11%, is able to identify the cases where shape alone predicts the native protein interface.

1 Introduction

When proteins associate with other proteins, they form complexes with biological function, including signal transduction[1–3], immune response[4,5], DNA binding[6–8], and enzyme activation[9–11]. Predicting the structure of these complexes and understanding their assembly mechanisms are of fundamental importance for design of protein assemblies[12–24] and rational drug design. The heuristic nature of currently available models to predict the structure of a protein complex based on steric and/or physicochemical complementarity at the protein-protein interface[25,26] illustrates our limited understanding of *in-vitro* protein-protein interactions. Conversely, simulation approaches at atomistic resolution come at a significant computational cost, which, in practice, limits their ability to study assembly processes and predict protein complexes[27].

The significance of shape complementarity has been reported since the earliest days of protein structure determination[28,29]. Tightly packed interfaces are observed in co-crystallized complexes in the Protein Data Bank (PDB)[25], motivating studies on the statistics of protein shape complementary[30,31] and development of geometry-based models of protein affinity[32–34]. In these studies, geometric match at the protein interfaces was reported in different functional classes including antibodyâĂŞantigen pairs, enzymeâĂŞinhibitor/substrate and other complexes, suggesting its important role in biomolecular recognition. Protein shape complementarity has also been incorporated in several machine learning models for binding interface prediction[35–37]. However, to our knowledge, shape complementarity has yet to be applied to directly simulate and predict protein association except for idealized shape models[38], which prevents the one-to-one mapping to biomolecules.

Simplified colloidal systems allow testing models of protein-protein association because their physical chemistry can be precisely controlled in experiment, and because colloidal interactions are well understood theoretically[39]. Motivated by Fischer's lock-and-key principle[40], colloids with complementary shape have been studied computationally and experimentally to exploit entropic depletion forces for assembly, regardless of their composition and surface chemistry[41–47]. Specifically, it has been conjectured that shape-allophilic, entropic interactions contribute significantly to proteins' lock-and-key binding[31,38,45]. In addition, shape complementarity has been exploited experimentally to enable the hierarchical assembly of DNA duplexes into higher-order structures[48,49], suggesting a universality beyond inorganic systems. For proteins, coarse-grained patchy representations accelerate simulations and capture their phase behavior in some cases[50–54]. However, constructing these models requires knowing the crystallized structures, which can obscure the main driving force. Isolating and investigating the role of shape complementarity in protein dimerization will enable us to better understand the mechanism of protein assembly and engineer protein interfaces for nanomaterials and therapeutics. Here, we aim to answer the question: is shape alone sufficient to assemble protein native complexes? If so, another challenge remains: can we identify these cases and use shape complementarity to predict the native interface? We address the first question by simulating protein interfaces with atomic-level resolution of molecular

[a] Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA. E-mail: sglotzer@umich.edu
[b] Biointerfaces Institute, University of Michigan, Ann Arbor, Michigan 48109, USA
‡ Present address: Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.

shape and a generic depletion interaction[55,56], and compare the assembled configurations with those known from experiment. We elucidate our results by analyzing the importance of competing binding configurations and how it affects the assembly. Based on the analysis, a machine learning classifier is built to answer the second question by identifying the shape binders and showing that shape complementarity can predict the native configuration correctly for strong and selective binders.

2 Model and methods

In depletion[57,58], overlap between the solute excluded volumes increases the volume available to the cosolutes (hereafter called "depletants"), and thus their entropy. To isolate shape contributions, we consider only hard interactions between proteins, *i.e.*, overlapping particle configurations are forbidden ($U = +\infty$). The system is therefore purely entropic [$\beta = (k_B T)^{-1} \equiv 1$]. We are interested in the idealized case where the depletant particles form an ideal gas, *i.e.*, they are mutually penetrable but hard with respect to the solute[56,57,59,60]. These depletants exchange with the grand-canonical system of interest through a reservoir with number density n_R. Then, the Boltzmann weight of a configuration $\vec{X}$ of solutes is given by

$$P(\vec{X}) \propto \exp\left[-H(\vec{X}) + n_R V_f(\vec{X})\right] \tag{1}$$

Here, H is the Hamiltonian, and V_f is the free volume available to the depletants, which for N solutes decomposes into contributions from single particle excluded volume V_{ex}^i, system volume V and overlap volume ΔV between solutes as $V_f = V - \sum_{i=1}^{N} V_{ex}^i + \Delta V$ (Fig. 1). It is therefore sufficient to consider only ΔV to determine the free energy change ΔF due to depletion interaction, having $\Delta F = -n_R \Delta V$. The dependence on overlap volume can be generalized to include the four Minkowski measures[61–63], but here we are interested solely in demonstrating the usefulness of the most idealized approach to protein assembly that only includes volume terms.

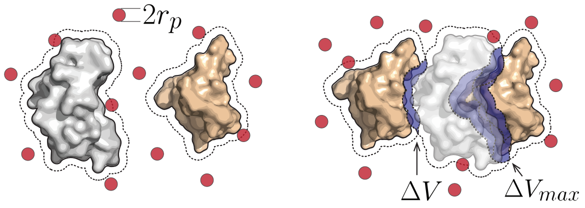

Fig. 1 Mechanism of depletion interactions, mediating shape complementarity at a protein dimer interface through the maximization of overlap volume ΔV, and hence entropy.

To investigate the shape effect on protein self-assembly, we simulated 46 dimer pairs in the Dockground database[64]. Four of the first 50 dimers in the database are excluded because inspection of the native configuration revealed that the path to assembly is topologically forbidden without reconfiguration due to their intertwined structure. Fig. 2 shows their corresponding native binding configurations. To reduce the computational effort, we use interface templates, which keep only the regions extracted from the full structures at a distance of 12Å from the interface, as shown in the highlighted region in Fig. 2. For each dimer pair, we study the binding process driven by shape complementarity, determine the binding interface, and compare the predicted interface with the native one. We simulate the binding processes of the protein dimers across a range of depletant sizes and densities with HOOMD-blue[65–67] using its Hard Particle Monte Carlo (HPMC) simulation method[55,56] in μ_dNVT thermodynamic ensemble. This method stores information only about the colloidal particles and accounts for depletion interactions implicitly by sampling the free volume change in the local environment of a colloidal particle during a trial move. Treating the depletants implicitly rather than explicitly along with consideration of the protein interface rather than the entire protein makes these simulations computationally possible[56].

As illustrated in Fig. 2, we model each protein monomer interface as a rigid sphere union with a single bead representing a heavy atom to capture the molecular shape of proteins. The corresponding atom position and radius of each atom bead are generated using the MSMS software[68]. For each dimer pair, we initialized the system in a random unbound configuration containing one receptor (grey monomers in Fig. 2) and $N = 45$ non-interacting ligands (gold monomers in Fig. 2). We fix the position and orientation of the receptor and only perform trail moves on the ligands. The ligands are non-interacting in the sense that they can penetrate each other and only have depletion interactions with the receptor. This setup enables modeling binding processes of multiple independent dimer pairs simultaneously, and prevents the system from forming larger scale aggregates. To resolve the geometric features of the protein surface, we choose depletant radii r_p of 0.20 and 0.25nm, slightly larger than a water molecule, for all the assembly simulations. The depletant size is determined to capture the surface geometry of the protein and to represent a generic depletion attraction range in an aqueous environment. We also vary the depletant reservoir volume fraction ($\phi = n_R * \frac{4}{3}\pi r_p^3$) ranging from 0.54 to 0.70. Each simulation runs for at least 3×10^7 HPMC steps. This choice of runtime was determined by observing yield remains approximately constant (fluctuates less than 5%) for a million HPMC steps. We ran three replicas at each statepoint.

To compare the assembled structure to the experimentally determined native binding configuration, we align the receptors in the assembled configuration to the native one, and calculate the root mean squared deviation (RMSD) between the ligands. Yield of a dimer is defined as the fraction of dimer pairs with RMSD less than 1nm among the multiple binding processes. We choose the RMSD tolerance following the Critical Assessment of Prediction of Interactions (CAPRI) criterion[69]. According to CAPRI, the ligand RMSD, calculated over the ligand residues after a structural superposition of the receptor, is an important quantity to determine the quality of a predicted model. A prediction with less than 1nm is classified to have acceptable quality, while beyond that the prediction is incorrect. We report equilibrium yield $\langle Y \rangle$ by averaging the transient yield Y over the final one million HPMC steps. The yield Y used in our study can be mapped to association constant K_a for a typical protein binding reaction $[L] + [R] \longrightarrow [LR]_{native}$, where $K_a = (V * Y)/(N * (1 - Y)^2)$ in which V is the system vol-

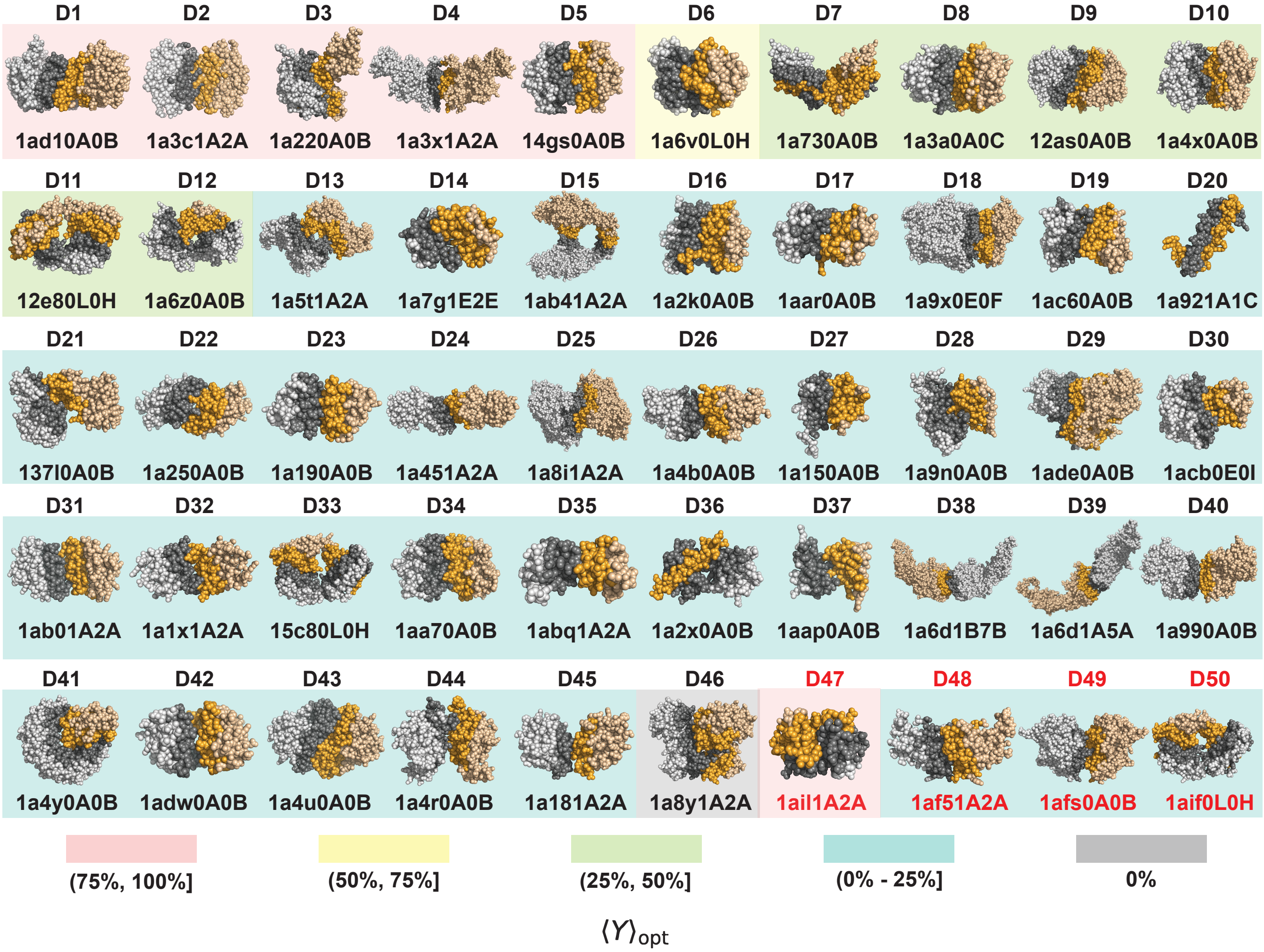

Fig. 2 Protein dimer structures in the Dockground database[64]. We summarize 50 studied protein dimer pairs ranked by their optimal yield $\langle Y \rangle_{opt}$. Interfaces are highlighted as the darkened regions. The background color shows the optimal yield $\langle Y \rangle_{opt}$ of each dimer. 46 protein dimers (labelled in black) are used in the original study, and four pairs (labelled in red) are used as the test cases in the shape binder classification.

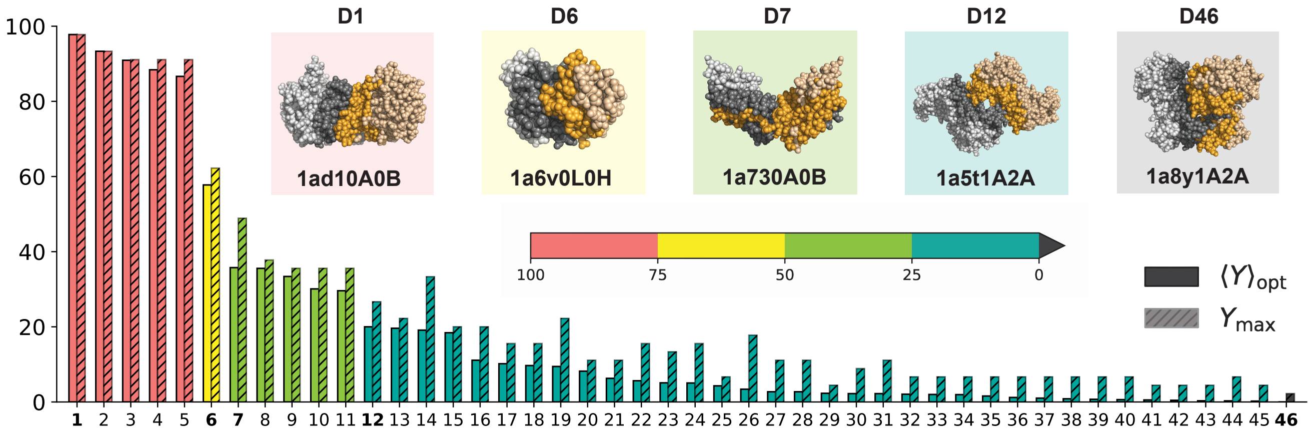

Fig. 3 Overall performance of dimer assembly. Equilibrium yield $\langle Y \rangle_{\rm opt}$ (solid color bars) and maximum yield $Y_{\rm max}$ (hashed) of 46 dimer pairs, showing significant variation across complexes, as also indicated by the discrete color scale. The finite value of $\langle Y \rangle_{\rm opt}$ for 45/46 dimers indicates protein dimers can form due solely to complementary shape. Non-zero values of $Y_{\rm max}$ reveal that the model can capture the native binding configurations. Insets Typical native binding configurations of protein dimers, as highlighted in the axis label.

ume, and N is the number of available reactant ligands.

We calculate the potential of mean force (PMF)[70,71] with the freud software toolbox[72] on the bound dimer pairs over the final one million HPMC steps. These calculations allow us to visualize the distribution of ligand binding configurations, and to quantify the free energy associated with the ligands binding in some positions relative to the fixed receptor. We also employ the support vector machines (SVM) classifier, using the metrics generated from PMF calculation, to identify the cases where shape alone is sufficient produce the native interface.

3 Results and discussion

All 46 dimers assembled in simulation, suggesting that our method samples the native complexes of most protein dimers. We evaluate the performance of our method on each protein dimer interface in terms of the yield of native assemblies found during simulation. In Fig. 3, we plot the maximum yield $Y_{\max}$ and the optimized equilibrium yield $\langle Y\rangle_{\text{opt}}$ for each protein pair. $Y_{\max}$ is defined as the maximum transient yield observed for simulation trajectories of a given protein dimer, while $\langle Y\rangle_{\text{opt}}$ is the equilibrium yield optimized across depletant parameters. Snapshots of the typical native binding configurations with various yield are shown in the insets. The equilibrium yield $\langle Y\rangle_{\text{opt}}$ is lower than $Y_{\max}$ because of transient binding, especially for the pairs in the tail of the distribution (Fig. 3, main plot). Notably, six protein dimers achieved over 50% yield, indicating that complementary shape on its own is sufficient for predicting their dimerization interfaces.

We further find that $\langle Y\rangle_{\text{opt}}$ varies significantly ($0-98\%$) across different dimer pairs. To investigate this behavior, we compare the optimal simulation trajectories of three representative examples with different $\langle Y\rangle_{\text{opt}}$, including D1 ($\langle Y\rangle_{\text{opt}} = 98\%$), D7 ($\langle Y\rangle_{\text{opt}} = 35\%$) and D21 ($\langle Y\rangle_{\text{opt}} = 7\%$) as shown in Fig. 4. We plot the evolution of the average RMSD and yield over all independent replicas in the left panel with insets showing the native contact. To further understand the distribution of the binding sites, we examine the three-dimensional PMF $W(\vec{r}) = -\ln g(\vec{r})$[70,71], which allows us to visualize the free energy map of pairwise interactions between receptor and the bound ligands (right panel of Fig. 4). Here, $g(\vec{r})$ is the pair correlation function. In all three systems, yield increases and average RMSD decreases over the course of the simulation, suggesting that the near-native protein binding interfaces are equilibrium configurations stabilized by complementary shape. The native interface of D1 is planar with complementary locks and keys on both sides. It is not surprising that the native contact is favored because increasing the size of such a "facet" strongly increases the protein overlap volume, hence the depletant entropy. Consistently, the PMF of D1 exhibits a deep well around the native binding site with few completing complexes. For D7, the native interface resembles a tadpole with two tails binding the pocket of the other, which is not consistent with the largest surface alignment. Yet, the depletion model still achieves 35% yield. This demonstrates that the maximization of overlapping volume in depletion model does not necessarily favor the largest surface alignment, and the addition of depletants with atomic length scale is able to capture the local geometric information on protein surface. The PMF becomes more disperse as more competing complexes with comparable shape entropy appear in the system, resulting in a lower yield. Finally, for D21, the ground state is no longer discernible due to a plethora of completing transient configurations. From these three examples, we infer that some protein dimers can be assembled solely with complementary shape, but they also beg the question, why do different dimers achieve different optimal $\langle Y\rangle_{\text{opt}}$? And how does $\langle Y\rangle$ depend on the depletant size and reservoir concentration?

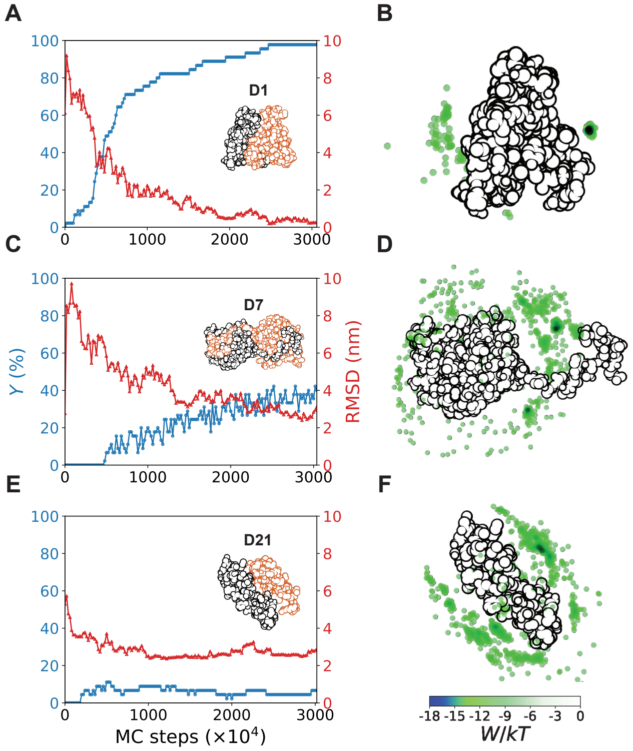

Fig. 4 Optimal assembly trajectories of D1, D7 and D21. (A, C, E) Yield increases and average RMSD decreases for all three systems as the binding processes evolve. Insets show the experimentally determined binding configurations with receptor colored in black and ligand colored in gold. (B, D, F) PMF of the complexes sampled by individual association events. The bound complexes are aligned to the receptor and positioned the same as the experimentally determined complex. A single snapshot of the receptor (black sphere unions) is shown for reference, together with the center of mass of ligands in competing configurations colored by the PMF.

We hypothesize three prerequisites to achieve a significant yield for the native configuration solely from the contribution of complementary shape. (1) Bind: The binding of two proteins should lower the depletant free energy. (2) Predict: Multiple different dimer configurations are sampled in a single trajectory and in the different replicas. To achieve high yield, the system needs to select a particular complex as its equilibrium configuration. Degenerate binding configurations, on the other hand, limit predictivity. The above two criteria can be quantified by analyzing the assembly trajectories without knowing the native binding interface. However, the prediction can result in a non-native interface even though these two conditions are satisfied. Thus a third prerequisite is needed, (3) Match: The dominant configuration, i.e. the prediction, coincides with the experimentally determined complex. For the systems studied here, the true interface is known. In the following, we quantify the above criteria for all complexes studied to evaluate the performance of the depletion method, extrapolate it to a situation where the native binding interface is unknown, and determine whether it binds due to complementary shape.

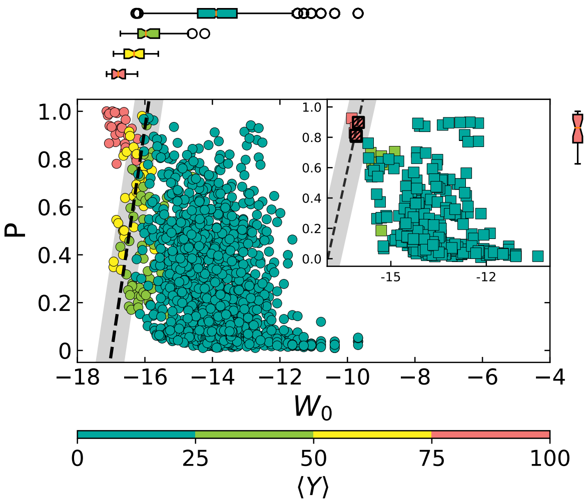

Fig. 5 Factors that determine the model performance. Correlations between assembly properties and yield. We show the distribution of two assembly properties colored by corresponding yield $\langle Y\rangle$ for all studied systems. The performance of our model is correlated to two assembly properties: binding free energy of prediction W_0 and binding selectivity P. In general, shape binders have low W_0 and high P (close to the upper left corner). The dashed line shows the SVM decision boundary between shape and non-shape binders. The SVM margin is depicted by the shaded area. Inset Evaluation of the SVM classifier on test cases (D47-D50 in Fig. 2), resulting in prediction accuracy of 98.96%. The classifier is generalizable and able to predict if a given protein pair is likely to be a shape binder. Three hashed square markers inside the SVM margin correspond to misclassified cases.

We analyze the first two criteria for each statepoint and investigate if they correlate with the yield. To quantify the tendency for two proteins to bind, we map their three-dimensional PMF as in Fig. 4, but only report the ground state free energy W_0 among all the bound complexes for a given pair. The bound complexes are defined as those in which the ligand has a closest distance less than 5.5Å to the receptor. To evaluate how predictive the simulation is, we first select bound states within $\Delta W = 6k_BT$ of W_0 as competing configurations, and cluster the PMF based on spatial distance with cutoff distance of 3Å. We then obtain several spatially disconnected competing clusters with comparable free energy, where cluster i contains N_i points with an average PMF of W_i. We define the selectivity P as the probability of forming the predicted complex among competing complexes, $P := N_0 \exp(-W_0)/\sum_{cluster:i} N_i \exp(-W_i)$, in which N_0 and W_0 corresponds to the prediction. We perform the analysis for all the statepoints and investigate how yield correlates with the assembly properties as shown in Fig. 5. Consistent with our hypothesis, W_0 decreases and P increases as yield increases. Overall, systems that successfully assemble the native complex, or high yield, are found at the upper left corner with low W_0 and high P, satisfying the first two prerequisites, strong and predictive binding. In contrast, systems with low yield (less than 20%) are typically binders with weaker binding strength, and they are frustrated by multiple competing configurations. We call shape binders the molecules that bind exclusively due to complementary shape with depletion, having $(\langle Y\rangle \geq 50\%)$. Additionally, for a given protein dimer pair, depletant parameters optimize yield by decreasing W_0 and increasing P, as shown in Fig. 6. We find that the optimal depletant parameters optimize $\langle Y\rangle$ by strengthening W_0 and increasing P, consistent with the first two prerequisite: bind and select. For D1, r_p and ϕ optimizes $\langle Y\rangle$ by achieving strong and selective binding at $r_p = 0.25nm$ and $\phi = 0.57$. For D7 and D21, the weak W_0 or low P limits the yield for the whole range of depletant parameters studied.

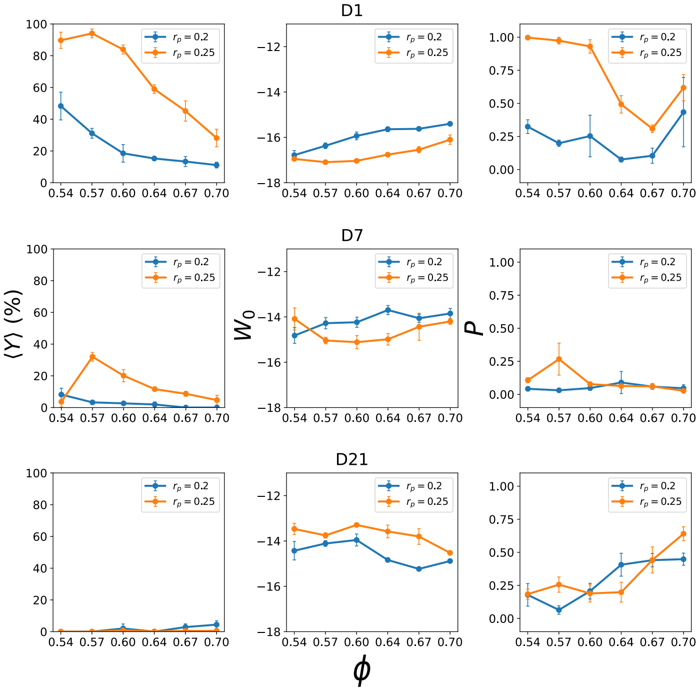

Fig. 6 Depletion affects yield by binding strength and selectivity. Depletant radii r_p and depletant reservoir volume fraction ϕ affect the equilibrium yield $\langle Y\rangle$ (Left Column) by tuning the ground state binding free energy W_0 (Middle Column) and the selectivity P (Right Column). Top, middle and bottom rows correspond to D1, D7 and D21 respectively.

We train a SVM model to classify shape binders $(\langle Y\rangle \geq 50\%)$ and non-shape binders $(\langle Y\rangle < 50\%)$ based on W_0 and P. The decision boundary of the classifier, given by $-2.49W_0 + 2.63P - 42.32 = 0$, is shown as the dashed line in Fig. 5. The model has a precision of 89.14 % and a recall of 77.11 % on the modeled systems with 10-fold cross validation, where precision $:=$ $\sum$shape binders found$/\sum$shape binders predicted and recall $:=$ $\sum$shape binders found$/\sum$all shape binders. We selected the next four dimers pairs in the Dockground database (Fig. 2, labelled in red) as test cases. The classifier has 98.96 % prediction accuracy on the test set, misclassifying only three shape binders inside the SVM margin (Fig. 5 Inset). Here the prediction accuracy is defined as the fraction of correctly identified shape binders and non-shape binders. This demonstrates the interface prediction ability of this model, i.e., depletion identifies the shape binder and predicts the native configuration correctly if the binders are strong and selective.

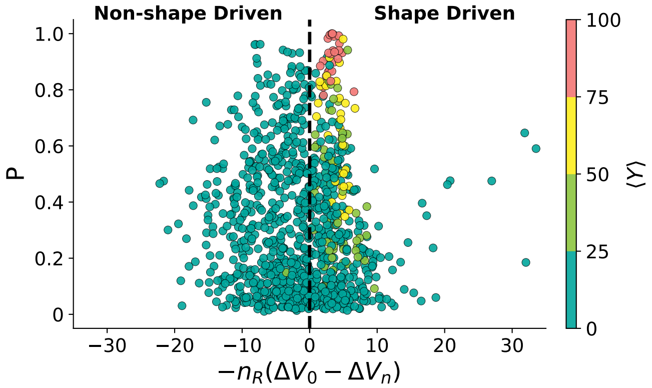

Fig. 7 Overlap volume analysis. The system selectivity (P) versus the depletant free energy difference between the predicted and native configuration ($-n_R(\Delta V_0 - \Delta V_n)$) colored by $\langle Y \rangle$.

To understand the circumstances under which the model makes a wrong prediction, we revisit the depletion free energy given by $\Delta F = -n_R \Delta V$ where n_R is the depletant number density, and ΔV is the overlap volume. We measure the difference of ΔV between the predicted equilibrium (ΔV_0) and the native one (ΔV_n) to categorize the systems into shape-driven states ($\Delta V_0 \leq \Delta V_n$) and non-shape driven states ($\Delta V_0 > \Delta V_n$) (Fig. 7). Shape driven systems can lower the free energy to form the equilibrium configuration, whereas for non-shape driven interfaces, at least one competing complex has lower free energy than the native one. In the latter case, depletion interaction alone cannot explain self-assembly of the native interface. Overall, 46% of systems in this study are shape driven. However, this property is predicted by the observed yield $\langle Y \rangle$. For low yield $\langle Y \rangle < 50\%$, 42.6% of the pairs are shape driven, while all the systems with $\langle Y \rangle > 50\%$ are shape driven. This finding validates the third prerequisite, that when native complexes form with depletion, they form reversibly and represent true equilibrium configurations. For non-shape driven systems, other forces such as electrostatic or hydrophobic interactions may drive the native interface formation. It is also possible that our study introduces different competing complexes with higher shape complementarity by only incorporating the interfacial parts of proteins. Taken together, however, there is strong evidence for the power of shape complementarity to predict native protein interfaces.

4 Conclusions

In this contribution, we isolated the effect of complementary shape on protein dimerization to evaluate if and how molecular shape affinity drives assembly. It is remarkable that shape alone can ever be sufficient to predict the native protein dimer interface. The shape-only model we introduced achieves a maximum yield of 98%, and samples the native configurations in all systems. Our results suggest that shape complementarity is more important for a subset of highly ranked proteins. We expect this knowledge to be useful for experimentalists who engineer protein-protein interfaces and interactions to design hierarchical protein structures.

Of course, biological systems such as enzymes and proteins rely not only on geometry and entropy, but also on intra- and intermolecular forces to guide and hold structures in place. Hence, we further outline the prerequisites for the model to be predictive and isolate the three contributions including the binding strength, selectivity and overlapping excluded volume, which all correlate with yield. The first two prerequisites provide guidance on the quality of the prediction even without prior knowledge of the native interface. Despite the simplicity of our model incorporating nothing more than protein shape and excluded volume interactions to elucidate the role of entropy in dimerization, it is predictive compared to patchy protein models that use existing structural information about the target protein complex that is only known a posteriori[73,74]. We expect more generally that for molecules with strong shape complementarity, the entropic nature of their excluded volume may be important for assembly. In future applications, *e.g.*, in drug discovery and for more flexible molecules, the shape-based method could be extended to include conformational ensembles due to its trivially parallel nature. Our results suggest that complementary shape not only serves as a conceptual model, *e.g.*, for enzyme kinetics, but also its predictive power is important for machine-learning approaches that may be otherwise model-agnostic.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This material is based upon work supported by the U.S. Army Research Laboratory and the U. S. Army Research Office under contract/Grant No. W911NF-18-1-0167. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725; INCITE Project MAT110: Nucleation and Growth of Colloidal Crystals on the Summit supercomputer. The computational workflow in general and data management in particular for this publication was primarily supported by the signac data management framework[75].

Notes and references

1 C. A. Stuermer, M. F. Langhorst, M. F. Wiechers, D. F. Legler, S. H. v. Hanwehr, A. H. Guse and H. Plattner, *The FASEB Journal*, 2004, **18**, 1731–1733.

2 S. E. Egan, B. W. Giddings, M. W. Brooks, L. Buday, A. M. Sizeland and R. A. Weinberg, *Nature*, 1993, **363**, 45–51.

3 J. D. Klemm, S. L. Schreiber and G. R. Crabtree, *Annual Review of Immunology*, 1998, **16**, 569–592.

4 M. L. Giglio, S. Ituarte, V. Milesi, M. S. Dreon, T. R. Brola, J. Caramelo, J. C. Ip, S. Maté, J.-W. Qiu, L. Otero and H. Heras, *Journal of Structural Biology*, 2020, **211**, 107531.

5 L. Premkumar, B. Segovia-Chumbez, R. Jadi, D. R. Martinez, R. Raut, A. J. Markmann, C. Cornaby, L. Bartelt, S. Weiss, Y. Park, C. E. Edwards, E. Weimer, E. M. Scherer, N. Rouphael,

S. Edupuganti, D. Weiskopf, L. V. Tse, Y. J. Hou, D. Margolis, A. Sette, M. H. Collins, J. Schmitz, R. S. Baric and A. M. de Silva, *Science Immunology*, 2020, **5**, eabc8413.

6 G. Arents and E. N. Moudrianakis, *Proceedings of the National Academy of Sciences*, 1993, **90**, 10489–10493.

7 H. Liang, S. Chen, P. Li, L. Wang, J. Li, J. Li, H.-H. Yang and W. Tan, *Journal of the American Chemical Society*, 2018, **140**, 4186–4190.

8 L. Andreeva, B. Hiller, D. Kostrewa, C. Lässig, C. C. de Oliveira Mann, D. Jan Drexler, A. Maiser, M. Gaidt, H. Leonhardt, V. Hornung and K.-P. Hopfner, *Nature*, 2017, **549**, 394–398.

9 M. Renatus, H. R. Stennicke, F. L. Scott, R. C. Liddington and G. S. Salvesen, *Proceedings of the National Academy of Sciences*, 2001, **98**, 14250–14255.

10 D. T. Dang, H. D. Nguyen, M. Merkx and L. Brunsveld, *Angewandte Chemie*, 2013, **125**, 2987–2991.

11 D. Cardinale, G. Guaitoli, D. Tondi, R. Luciani, S. Henrich, O. M. H. Salo-Ahen, S. Ferrari, G. Marverti, D. Guerrieri, A. Ligabue, C. Frassineti, C. Pozzi, S. Mangani, D. Fessas, R. Guerrini, G. Ponterini, R. C. Wade and M. P. Costi, *Proceedings of the National Academy of Sciences*, 2011, **108**, E542–E549.

12 J. E. Padilla, C. Colovos and T. O. Yeates, *Proceedings of the National Academy of Sciences*, 2001, **98**, 2217–2221.

13 S. Gonen, F. DiMaio, T. Gonen and D. Baker, *Science*, 2015, **348**, 1365–1368.

14 N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. André, T. Gonen, T. O. Yeates and D. Baker, *Science*, 2012, **336**, 1171–1174.

15 Y. Hsia, J. B. Bale, S. Gonen, D. Shi, W. Sheffler, K. K. Fong, U. Nattermann, C. Xu, P.-S. Huang, R. Ravichandran, S. Yi, T. N. Davis, T. Gonen, N. P. King and D. Baker, *Nature*, 2016, **535**, 136–139.

16 A. J. Simon, Y. Zhou, V. Ramasubramani, J. Glaser, A. Pothukuchy, J. Gollihar, J. C. Gerberich, J. C. Leggere, B. R. Morrow, C. Jung, S. C. Glotzer, D. W. Taylor and A. D. Ellington, *Nature Chemistry*, 2019, **11**, 204–212.

17 A. D. Malay, N. Miyazaki, A. Biela, S. Chakraborti, K. Majsterkiewicz, I. Stupka, C. S. Kaplan, A. Kowalczyk, B. M. A. G. Piette, G. K. A. Hochberg, D. Wu, T. P. Wrobel, A. Fineberg, M. S. Kushwah, M. Kelemen, P. Vavpetič, P. Pelicon, P. Kukura, J. L. P. Benesch, K. Iwasaki and J. G. Heddle, *Nature*, 2019, **569**, 438–442.

18 E. Golub, R. H. Subramanian, J. Esselborn, R. G. Alberstein, J. B. Bailey, J. A. Chiong, X. Yan, T. Booth, T. S. Baker and F. A. Tezcan, *Nature*, 2020, **578**, 172–176.

19 S. Zhang, R. G. Alberstein, J. J. De Yoreo and F. A. Tezcan, *Nature Communications*, 2020, **11**, 1–12.

20 J. I. Park, T. D. Nguyen, G. de Queirós Silveira, J. H. Bahng, S. Srivastava, G. Zhao, K. Sun, P. Zhang, S. C. Glotzer and N. A. Kotov, *Nature Communications*, 2014, **5**, 1–9.

21 G. de Q. Silveira, N. S. Ramesar, T. D. Nguyen, J. H. Bahng, S. C. Glotzer and N. A. Kotov, *Chemistry of Materials*, 2019, **31**, 7493–7500.

22 J. D. Brodin, E. Auyeung and C. A. Mirkin, *Proceedings of the National Academy of Sciences*, 2015, **112**, 4564–4569.

23 O. G. Hayes, J. R. McMillan, B. Lee and C. A. Mirkin, *Journal of the American Chemical Society*, 2018, **140**, 9269–9274.

24 J. R. McMillan, J. D. Brodin, J. A. Millan, B. Lee, M. Olvera de la Cruz and C. A. Mirkin, *Journal of the American Chemical Society*, 2017, **139**, 1754–1757.

25 I. A. Vakser, *Biophysical Journal*, 2014, **107**, 1785–1793.

26 L. C. Xue, D. Dobbs, A. M. Bonvin and V. Honavar, *FEBS Letters*, 2015, **589**, 3516–3526.

27 A. C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T. M. Weinreich and D. E. Shaw, *Proceedings of the National Academy of Sciences*, 2019, **116**, 4244–4249.

28 F. Crick, *Acta Crystallographica*, 1953, **6**, 689–697.

29 D. E. Koshland Jr, *Angewandte Chemie International Edition in English*, 1995, **33**, 2375–2378.

30 M. C. Lawrence and P. M. Colman, *Journal of Molecular Biology*, 1993, **234**, 946 – 950.

31 D. Kuroda and J. J. Gray, *Bioinformatics*, 2016, **32**, 2451–2456.

32 M. L. Connolly, *Biopolymers*, 1986, **25**, 1229–1247.

33 R. Norel, D. Petrey, H. J. Wolfson and R. Nussinov, *Proteins: Structure, Function, and Bioinformatics*, 1999, **36**, 307–317.

34 R. Chen and Z. Weng, *Proteins: Structure, Function, and Bioinformatics*, 2003, **51**, 397–408.

35 R. Townshend, R. Bedi, P. Suriana and R. Dror, Advances in Neural Information Processing Systems, 2019, pp. 15642–15651.

36 P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. Bronstein and B. Correia, *Nature Methods*, 2020, **17**, 184–192.

37 C. L. McCafferty, E. M. Marcotte and D. W. Taylor, *Proteins: Structure, Function, and Bioinformatics*, 2021, **89**, 348–360.

38 Y. Li, X. Zhang and D. Cao, *Scientific Reports*, 2013, **3**, 3271.

39 W. B. Russel, D. A. Saville and W. R. Schowalter, *Colloidal Dispersions*, Cambridge University Press, 1989.

40 E. Fischer, *Berichte der deutschen chemischen Gesellschaft*, 1894, **27**, 2985–2993.

41 M. Kinoshita and T. Oguni, *Chemical Physics Letters*, 2002, **351**, 79–84.

42 M. Kinoshita, *The Journal of Chemical Physics*, 2002, **116**, 3493–3501.

43 P.-M. König, R. Roth and S. Dietrich, *EPL (Europhysics Letters)*, 2009, **84**, 68006.

44 S. Sacanna, W. T. Irvine, P. M. Chaikin and D. J. Pine, *Nature*, 2010, **464**, 575–578.

45 E. S. Harper, R. L. Marson, J. A. Anderson, G. Van Anders and S. C. Glotzer, *Soft Matter*, 2015, **11**, 7250–7256.

46 L. Colón-Meléndez, D. J. Beltran-Villegas, G. Van Anders, J. Liu, M. Spellings, S. Sacanna, D. J. Pine, S. C. Glotzer, R. G. Larson and M. J. Solomon, *The Journal of Chemical Physics*, 2015, **142**, 174909.

47 A. M. Mihut, B. Stenqvist, M. Lund, P. Schurtenberger and

J. J. Crassous, *Science Advances*, 2017, **3**, e1700321.

48 S. Woo and P. W. Rothemund, *Nature Chemistry*, 2011, **3**, 620.

49 T. Gerling, K. F. Wagenbauer, A. M. Neuner and H. Dietz, *Science*, 2015, **347**, 1446–1452.

50 A. Stradner and P. Schurtenberger, *Soft Matter*, 2020, **16**, 307–323.

51 D. Fusco and P. Charbonneau, *Physical Review E*, 2013, **88**, 012721.

52 S. Whitelam, *Physical Review Letters*, 2010, **105**, 088102.

53 N. Dorsaz, L. Filion, F. Smallenburg and D. Frenkel, *Faraday Discussions*, 2012, **159**, 9–21.

54 I. Staneva and D. Frenkel, *The Journal of Chemical Physics*, 2015, **143**, 194511.

55 A. S. Karas, J. Glaser and S. C. Glotzer, *Soft Matter*, 2016, **12**, 5199–5204.

56 J. Glaser, A. S. Karas and S. C. Glotzer, *Journal of Chemical Physics*, 2015, **143**, 184110.

57 S. Asakura and F. Oosawa, *The Journal of Chemical Physics*, 1954, **22**, 1255–1256.

58 H. N. Lekkerkerker and R. Tuinier, *Colloids and the Depletion Interaction*, Springer, 2011.

59 B. Widom and J. S. Rowlinson, *The Journal of Chemical Physics*, 1970, **52**, 1670–1684.

60 M. Dijkstra, R. van Roij, R. Roth and A. Fortini, *Physical Review E*, 2006, **73**, 041404.

61 M. Oettel, H. Hansen-Goos, P. Bryk and R. Roth, *EPL (Europhysics Letters)*, 2009, **85**, 36003.

62 R. Roth, Y. Harano and M. Kinoshita, *Physical Review Letters*, 2006, **97**, 078101.

63 J. F. Robinson, F. Turci, R. Roth and C. P. Royall, *Physical Review Letters*, 2019, **122**, 068004.

64 P. J. Kundrotas, I. Anishchenko, T. Dauzhenka, I. Kotthoff, D. Mnevets, M. M. Copeland and I. A. Vakser, *Protein Science*, 2018, **27**, 172–181.

65 J. A. Anderson, C. D. Lorenz and A. Travesset, *Journal of Computational Physics*, 2008, **227**, 5342–5359.

66 J. A. Anderson, M. Eric Irrgang and S. C. Glotzer, *Computer Physics Communications*, 2016, **204**, 21–30.

67 J. A. Anderson, J. Glaser and S. C. Glotzer, *Computational Materials Science*, 2020, **173**, 109363.

68 M. F. Sanner, A. J. Olson and J.-C. Spehner, *Biopolymers*, 1996, **38**, 305–320.

69 M. F. Lensink, S. Velankar, A. Kryshtafovych, S.-Y. Huang, D. Schneidman-Duhovny, A. Sali, J. Segura, N. Fernandez-Fuentes, S. Viswanath, R. Elber *et al.*, *Proteins: Structure, Function, and Bioinformatics*, 2016, **84**, 323–348.

70 G. van Anders, N. K. Ahmed, R. Smith, M. Engel and S. C. Glotzer, *ACS Nano*, 2014, **8**, 931–940.

71 G. van Anders, D. Klotsa, N. K. Ahmed, M. Engel and S. C. Glotzer, *Proceedings of the National Academy of Sciences*, 2014, **111**, E4812–E4821.

72 V. Ramasubramani, B. D. Dice, E. S. Harper, M. P. Spellings, J. A. Anderson and S. C. Glotzer, *Computer Physics Communications*, 2020, **254**, 107275.

73 D. Fusco, J. J. Headd, A. De Simone, J. Wang and P. Charbonneau, *Soft Matter*, 2014, **10**, 290–302.

74 D. Fusco and P. Charbonneau, *Colloids and Surfaces B: Biointerfaces*, 2016, **137**, 22–31.

75 C. S. Adorf, P. M. Dodd, V. Ramasubramani and S. C. Glotzer, *Comput. Mater. Sci.*, 2018, **146**, 220–229.