nature computational science

ARTICLES

https://doi.org/10.1038/s43588-022-00229-w

Check for updates

Unifying structural descriptors for biological and bioinspired nanoscale complexes

Minjeong Cha[1,2,11], Emine Sumeyra Turali Emre[2,3,11], Xiongye Xiao[4], Ji-Young Kim[2,3], Paul Bogdan[4], J. Scott VanEpps[2,5,6,7,8], Angela Violi[3,9,10] and Nicholas A. Kotov[1,2,3,5,6] ✉

Biomimetic nanoparticles are known to serve as nanoscale adjuvants, enzyme mimics and amyloid fibrillation inhibitors. Their further development requires better understanding of their interactions with proteins. The abundant knowledge about protein–protein interactions can serve as a guide for designing protein–nanoparticle assemblies, but the chemical and biological inputs used in computational packages for protein–protein interactions are not applicable to inorganic nanoparticles. Analysing chemical, geometrical and graph-theoretical descriptors for protein complexes, we found that geometrical and graph-theoretical descriptors are uniformly applicable to biological and inorganic nanostructures and can predict interaction sites in protein pairs with accuracy >80% and classification probability ~90%. We extended the machine-learning algorithms trained on protein–protein interactions to inorganic nanoparticles and found a nearly exact match between experimental and predicted interaction sites with proteins. These findings can be extended to other organic and inorganic nanoparticles to predict their assemblies with biomolecules and other chemical structures forming lock-and-key complexes.

Interactions between proteins are conceptually described as lock-and-key complexes[1], reflected in multiple successful protein–protein interactions (PPI) algorithms, such as PRISM, PSIVER and MaSIF[2–4]. These and other computational packages predict protein complex formation and interaction sites by assessing the pairwise similarity of a potential 'key' with many other 'keys'. A similar concept can be applied to nanoparticle (NP)–protein interactions, but its realization requires a massive library of X-ray diffraction data for NP–protein pairs comparable to the Protein Data Bank (PDB), which is currently unavailable. Other PPI algorithms, such as SPPIDER and Pre-PPI, combine the geometrical description of docking molecules with structural relations at the organism level, exemplified by protein networks from evolutionary homology and genomics[5,6]. Importantly, these PPI software packages[7–11] also assume that the interacting molecules are linear polymers from amino acids (AAs). Such descriptors are natural for proteins but make it impossible to extend these algorithms to bioinspired inorganic NPs, even though they may carry some AAs as surface ligands[11–13]. The simplified molecular-input line-entry system can annotate the structure of nonpeptide biomolecules[14] but is, again, inapplicable to biomimetic NPs, even those based on carbon atoms, while many NPs exhibiting strong specific biological activity are entirely inorganic[15,16]. Unifying structural description of proteins and NPs is possible at the atomistic molecular dynamics (MD) level that represents the state of the art in predictions of NP–protein interactions[17–19]. However, the interaction time probed by typical atomistic MD methods is mainly limited to hundreds of nanoseconds[17–21]. Even with the dedicated Anton2 supercomputer, the interaction time can only reach up to 2 μs (ref. [22]), while the time required for the formation of protein–protein and NP–protein complexes may exceed minutes or sometimes hours[23,24]. While being significant for complexes between macromolecules, the weak multicentre interactions exemplified by dipole–dipole forces and collective hydrogen bonds are difficult to implement without drastic time restrictions. The complexity of the energy landscape for nanoscale interactions may also lead to entrapment of MD simulations in metastable states before the formation of a fully equilibrated complex.

Here, we analyse the role of different structural features contributing to the formation of protein–protein complexes with the goal of identifying structural descriptors that could be uniformly applicable to complexes between proteins and NPs. Identifying such descriptors would enable one to extend the knowledge gained from the vast PPI datasets and existing algorithms to NP–protein pairs encountered in diverse biomedical contexts, from drug delivery to the environmental effects of NPs.

Results

Distance matrices of protein complexes. A protein complex (Fig. 1a) can be represented as a distance matrix D_{AB} ($d_{i,k}$) where $1 < i < N_A$ and $1 < k < N_B$ with a set of matrix elements $d_{i,k}$ representing the distance in angstroms between pairs of α-carbon (C_α) in AA residues from proteins A and B (Fig. 1b)[25]. The darkest areas of the matrix (yellow boxes) indicate the AAs in macromolecules A and B that are the closest to each other. The level of proximity of A_i and B_k in D_{AB} ($d_{i,k}$) will be used to distinguish interacting and noninteracting residue pairs in machine-learning (ML) algorithms (Supplementary Fig. 1). The proteins are less likely to form a lock-and-key complex when the predicted probabilities of interacting AA residue pairs within 7 Å from each other are low (<0.5). If this mathematical approach is successful for protein complex, it can, perhaps, be

[1]Department of Materials Science and Engineering, University of Michigan, Ann Arbor, MI, USA. [2]Biointerfaces Institute, University of Michigan, Ann Arbor, MI, USA. [3]Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, USA. [4]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. [5]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA. [6]Program in Macromolecular Science and Engineering, University of Michigan, Ann Arbor, MI, USA. [7]Department of Emergency Medicine, University of Michigan, Ann Arbor, MI, USA. [8]Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, MI, USA. [9]Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA. [10]Biophysics Program, University of Michigan, Ann Arbor, MI, USA. [11]These authors contributed equally: Minjeong Cha, Emine Sumeyra Turali Emre. ✉e-mail: kotov@umich.edu

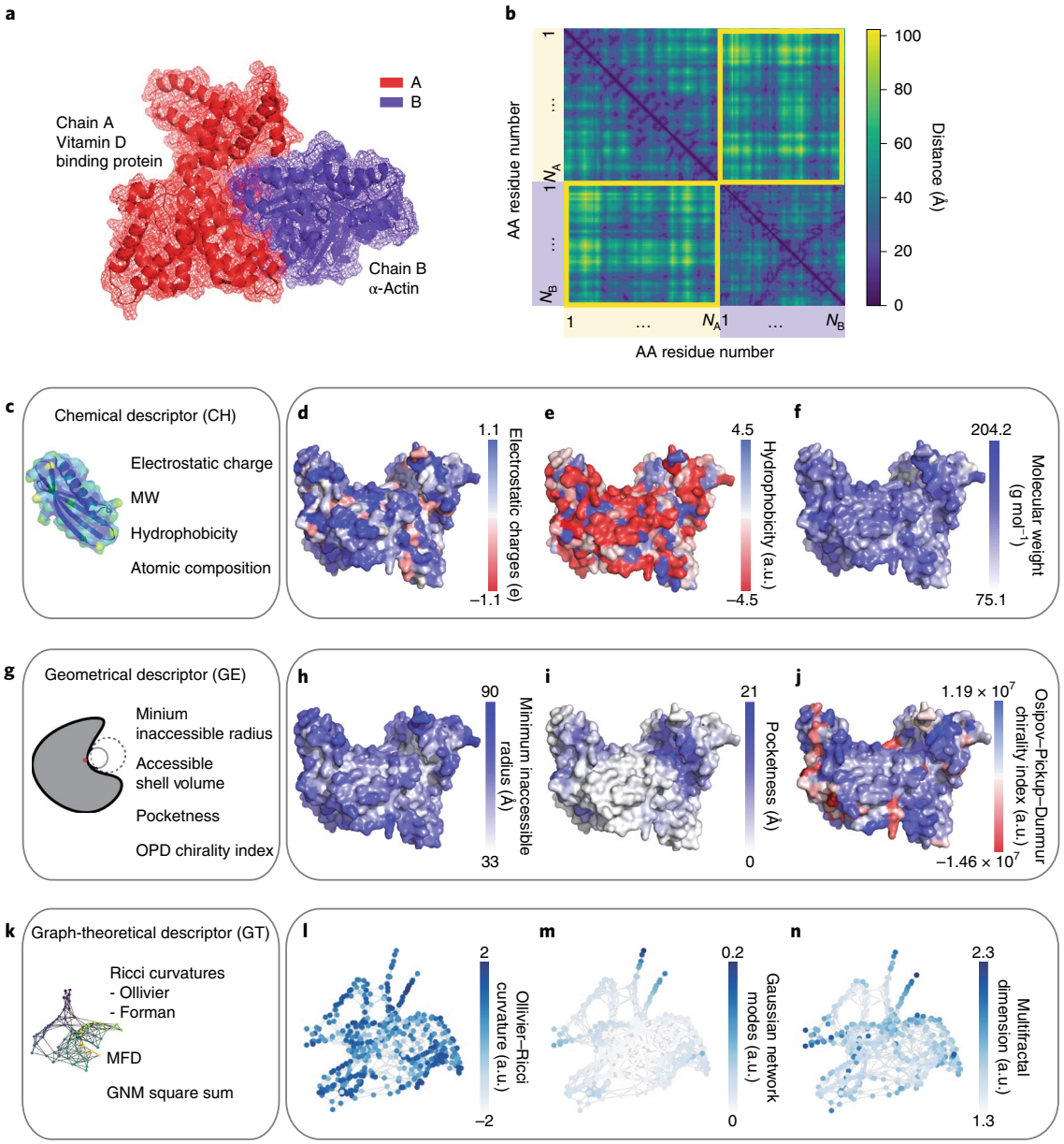

Fig. 1 | The concept of the distance matrix of a protein complex and the introduction of descriptors. a, An example of two interacting proteins, chain A and B of PDB ID 1MA9 (vitamin D binding protein and α-actin). **b**, The distance matrix (in Å) of a protein complex (PDB ID 1MA9), where the yellow box represents the interaction fingerprints between two different proteins. The darkest areas of the matrix (yellow boxes) indicate the AAs in macromolecules A and B that are the closest to each other. **c**, Feature list of chemical (CH) descriptors. **d**–**f**, Example feature visualization of electrostatic charge of the carbon atom (**d**), hydrophobicity (**e**) and molecular weight (**f**). **g**, Feature list of geometrical (GE) descriptors. **h**–**j**, Example feature visualization of minimum inaccessible radius (R_{inacc}) (**h**), pocketness (Pocket) (**i**) and Osipov–Pickup–Dunmur (OPD) chirality index (**j**). **k**, Feature list of graph-theoretical (GT) descriptors. **l**–**n**, Example feature visualization of Ollivier–Ricci curvature (ORC) (**l**), Gaussian network models (GNM) modes (**m**) and multifractal dimension (MFD) (**n**).

extended to nanoscale assemblies from abiological nanostructures because it relies on structural coding based on the three-dimensional (3D) geometry of the macromolecules.

Contributing descriptors. The chemical (CH), geometrical (GE) and graph-theoretical (GT) descriptors are computed and embedded into each of A_i and B_k to form characteristic feature matrices that comprehensively characterize the interacting macromolecules from different physicochemical perspectives. The CH descriptors include the electrostatic charge (C-charges)[26], hydrophobicity (Hp), molecular weight (MW), polarity and atomic compositions (C-count) of the biomolecules (Fig. 1c–f and Supplementary Figs. 8–10).

The GE descriptors include Cartesian (local distances and shapes), topological (global organization) and asymmetry (chirality) characteristics of the interacting subunits at the nanoscale. The GE descriptors also include the minimum inaccessible radius (R_{inacc}), the accessible shell volume (Shell) and the pocketness (Pocket)[27] (Fig. 1g–i and Supplementary Fig. 11). Chirality is calculated for the vicinity of each AA residue as the Osipov–Pickup–Dunmur (OPD, Fig. 1j) indices[28]. These GE measures assesses the compatibility of the protein geometries to each other at nanometre and subnanometre scales.

We found that the areas of positive OPD values in proteins are distinctly associated with α-helices (Supplementary Fig. 12).

The proteins and their assemblies can also be represented as a graph, $G(n, e)$, constructed by taking individual AA residues as nodes (n) while the edges (e) between the nodes are assigned depending on the distance matrix for a single-folded protein, D_A ($d_{i,j}$). GT descriptors enable structural encoding of protein complexes without reliance on the AA sequence in the macromolecule. Furthermore, GT descriptors add classifiers that depict the shape complexity, chemical connectedness and molecular deformability of these structures (Fig. 1k–n). GT descriptors utilize well-developed applied-mathematics methods that enable acceleration of the computations while reducing the computational resources required[29,30]. For GT descriptors, three parameters were calculated: (1) Gaussian network models (GNMs) representing macromolecules as elastic networks to describe their flexibility[31,32] (Fig. 1m and Supplementary Fig. 15), (2) the Ollivier–Ricci curvature (ORC) and Forman–Ricci curvature (FRC) describing the macromolecules in terms of Riemannian geometry to identify the segments subject to conformational changes (Fig. 1l and Supplementary Fig. 14) and (3) the node-based multifractal dimension (MFD) describing the molecules as fractals to account for the hierarchical organization of macromolecules essential for their interactions (Fig. 1n and Supplementary Fig. 13).

Descriptor correlations in protein–protein complexes. We analysed cross-correlations between the CH, GE and GT descriptors to understand their (1) independent inputs into the formation of protein complexes and (2) enumeration of suitability of non-proteinaceous macromolecules for descriptors of similar complexes with NPs. Apart from some a priori expected correlations between FRC and ORC, Shell versus R_{inacc} and MW and C-count, the correlation between different components of the descriptors is small (Fig. 2a,b). Notably, the correlations between the descriptors for the molecules overall (Fig. 2a) and interfaces are quite different (Fig. 2b). This fact highlights (1) the mutual structural adaptation of the macromolecule at the interface and (2) the significance of descriptors such as R_{inacc}, Shell, Pocket, OPD, ORC and MFD characterizing the geometry and dynamics of the interfaces. In view of the lock-and-key concept, one can also ask whether a specific value of descriptors in one protein requires a particular value of the same feature on the counterpart forming the complex, which will be informative in predictions of preferred interaction sites[33,34]. The contour plots in Fig. 2c–e present the distribution of R_{inacc}, ORC and OPD for AA pairs at different distances from each other. The plots for distances of <7 Å and 7–10 Å represent AAs located in close proximity of PPI. The distinct maxima on these plots, especially for AAs located at distances of 7–10 Å, vividly indicate that all these structural descriptors indeed require specific values when the macromolecules try to fit each other. As such, the local chirality in the neighbourhood of AAs located directly across the interface (<7 Å) tends to be small, indicating that highly mirror-asymmetric 'holes' have greater difficulty in finding a fitting 'key', which shifts to mutually positive OPD values of about 0.3×10^7 for AAs separated by 7–10 Å. The three maxima observed in the contour plots for R_{inacc} for distances of 10–20 Å clearly indicate the long-range correlation between GE features required for complex formation (see Supplementary Figs. 16 and 17 for additional descriptors).

ML algorithms for protein–protein complexes. CH, GE and GT descriptors calculated for each AA residue of the constituents in the protein complexes served as inputs for ML algorithms, while the distance matrix D_{AB} ($d_{i,k}$) was the output (Figs. 1 and 3a and Supplementary Sections 1.1–1.4). We trained different ML algorithms, namely logistic regression, Gaussian naïve Bayes, support vector machine (SVM), random forest (RF), XGBoost (XGB) and deep neural network (DNN), using seven independent datasets: all descriptors, CH only, GE only, GT only, CH + GE, CH + GT and GE + GT. Among the sets with a single descriptor type, the GT-only set performed best versus the CH-only or GE-only sets, with an area under the receiver operating characteristic curve (ROC-AUC) of 87.7 ± 0.7%, accuracy as high as 80.8 ± 1% and an F1 score of 80.0 ± 1.2% with the tenfold cross-validated DNN model. The same characteristics were 83.8 ± 1.3%, 77.0 ± 1.2% and 75.6 ± 1.2% for the GE descriptors and 59.0 ± 1.1%, 57.4 ± 0.8% and 54.8 ± 2.8% for the CH descriptors, respectively. It was unexpected that, when adding the CH descriptors to the GE descriptors for training the high-performance DNN model, the ROC-AUC and accuracy increased by only 1.9% and 2.1%, respectively. With the addition of the CH descriptors to the GT descriptors, the ROC-AUC and accuracy scores decreased by 0.1% and 0.2%, respectively (Fig. 3b,c and Table 1). The other ML algorithms, such as RF and XGB (Table 1 and Supplementary Fig. 18), performed similarly to the DNN and maintained the same trends in terms of ROC-AUC, accuracy and F1 score (Supplementary Section 3.2.1).

The feature ablation study suggests that the GT descriptors make the most significant contribution to the predictive power of ML algorithms. Despite the absence of a strong direct correlation between the CH and other descriptors, it is apparent that GT and GE descriptors contain adequate information to predict the formation of protein complexes. This finding is quite surprising because electrostatic, van der Waals and hydrogen-bonding interactions are expected to be most relevant to PPI, being dependent on C-charges, MW, C-count and Hp. However, CH descriptors are strongly correlated amongst themselves (Fig. 2a), which indicates that training of ML algorithms using all of them as classifiers increases the laboriousness of the process but not the accuracy of the predictions. The unexpectedly low impact on the ML algorithm performance of adding the CH to the GT or GE descriptors is observed because GE and topological parameters of the macromolecules emerge from the multiplicity of weak and strong chemical interactions, which creates a path for partial embedding of chemical information into GE and GT features. Another important factor is the scale of GE and GT descriptors, which matches the dimensions of the protein–protein interface covering the area of several square nanometres, while the scale of CH features is commensurate with the size of single AA residues.

ML predictions for protein–protein complexes. The predictive capability of the ML algorithm based on different sets of structural descriptors was compared for several proteins that were not included in the training set. Furthermore, they were nonhomologous to those present in the database to test the true 'learning' rather than 'memorization' capabilities of the algorithms applied. The tested protein complexes included chains A and B of (1) severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) nucleocapsid (PDB ID 6WZO)[35], (2) fluorescent protein Dronpa (PDB ID 6NQN)[36] and (3) bacterial tryptophan synthase (PDB ID 1C29)[37] (Fig. 4, Supplementary Figs. 19–23 and Supplementary Tables 4–6). The formation of a complex was based on the AA residues from each macromolecules being within a distance of 7 Å. These residues defined interfaces between the interacting molecules. For a fair comparison, the ground-truth data were calculated with the same assumptions as the ML algorithms (Fig. 4a,c,e and Supplementary Section 3.3.1). Comparing the outcomes of ML using different descriptor sets, we confirmed that the GE + GT descriptors predicted the interaction sites of each protein complex (Fig. 4b,d,f) with ~80% accuracy with only a few false negatives (Supplementary Tables 4–6). The false positives were predominantly located in the vicinity of the true interface sites (Supplementary Table 3), which reflects the connectivity of the protein globules, perhaps overestimated by GT parameters. When comparing the models trained on

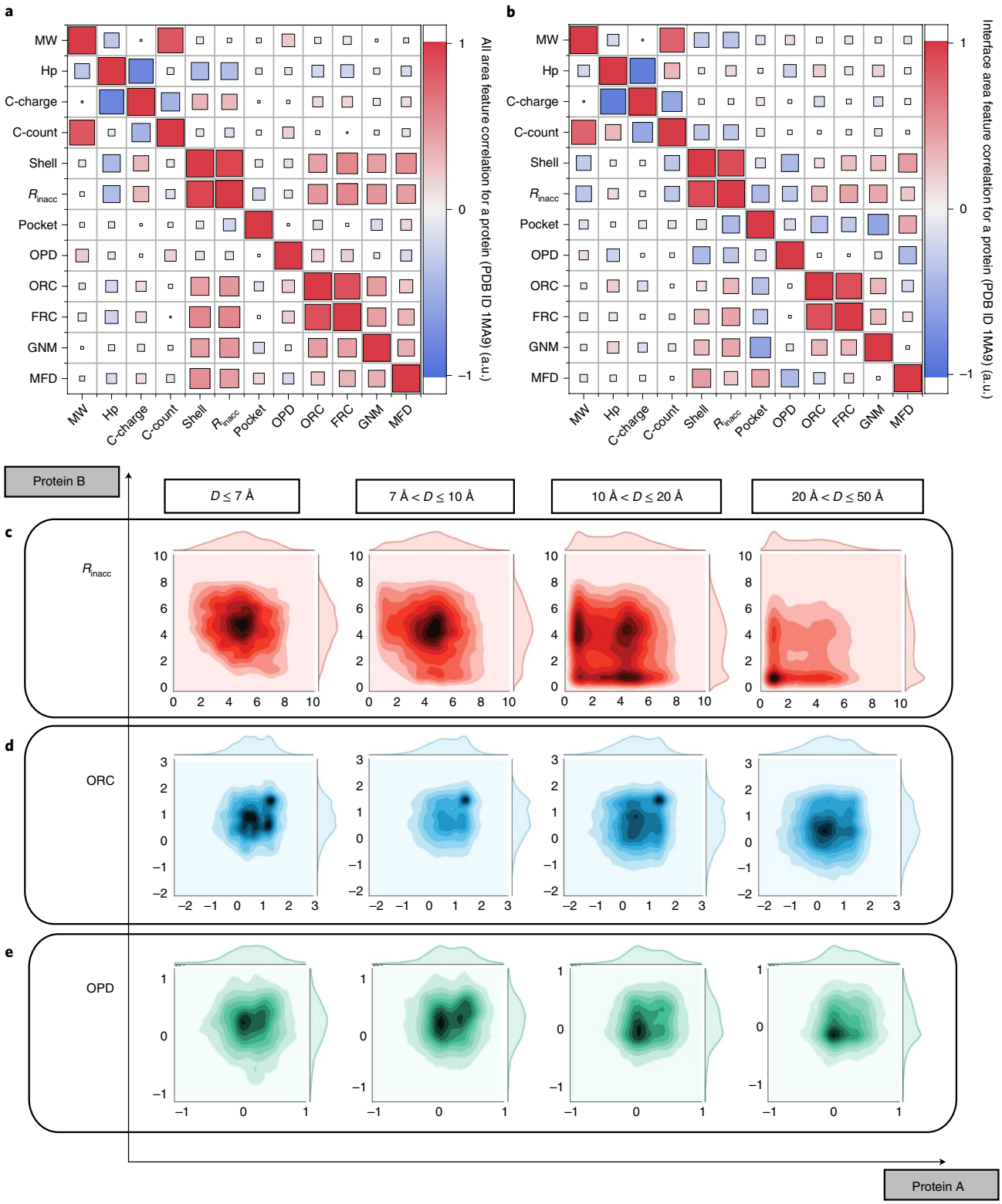

Fig. 2 | Analysis of descriptors. a, Descriptor correlations for all AA residues in a protein complex (PDB ID 1MA9). **b**, Descriptor correlation for interface AA residues (contact distance <7 Å) in a protein complex (PDB ID 1MA9) shown as correlation matrices where the size of the square depicts the correlation strength and its colour indicates its sign. Pocketness (Pocket) features are expected to correlate with accessible shell volume (Shell) and minimum inaccessible radius (R_{inacc}) because geometries with bumps and protrusions are more accessible than others. There are also positive correlations between ORC, FRC, GNM modes and MFD with R_{inacc} and Shell, because these GE and GT descriptors tend to have higher values in the convex part of the molecular structures. **c**–**e**, The correlation distribution of R_{inacc} (**c**), ORC (**d**) and OPD indices (**e**) values from each protein forming a complex depending on the distance between AA residues. In each contour plot, the *x* and *y* axes indicate the descriptor values of protein A and B, respectively. Four distance classes describe the physical distance between AA residues in the protein–protein complex. The 7 Å and 10 Å classes describe the immediate vicinity of protein–protein interfaces. The values for the OPD indices in **e** are scaled by division fraction of 1×10^{7}.

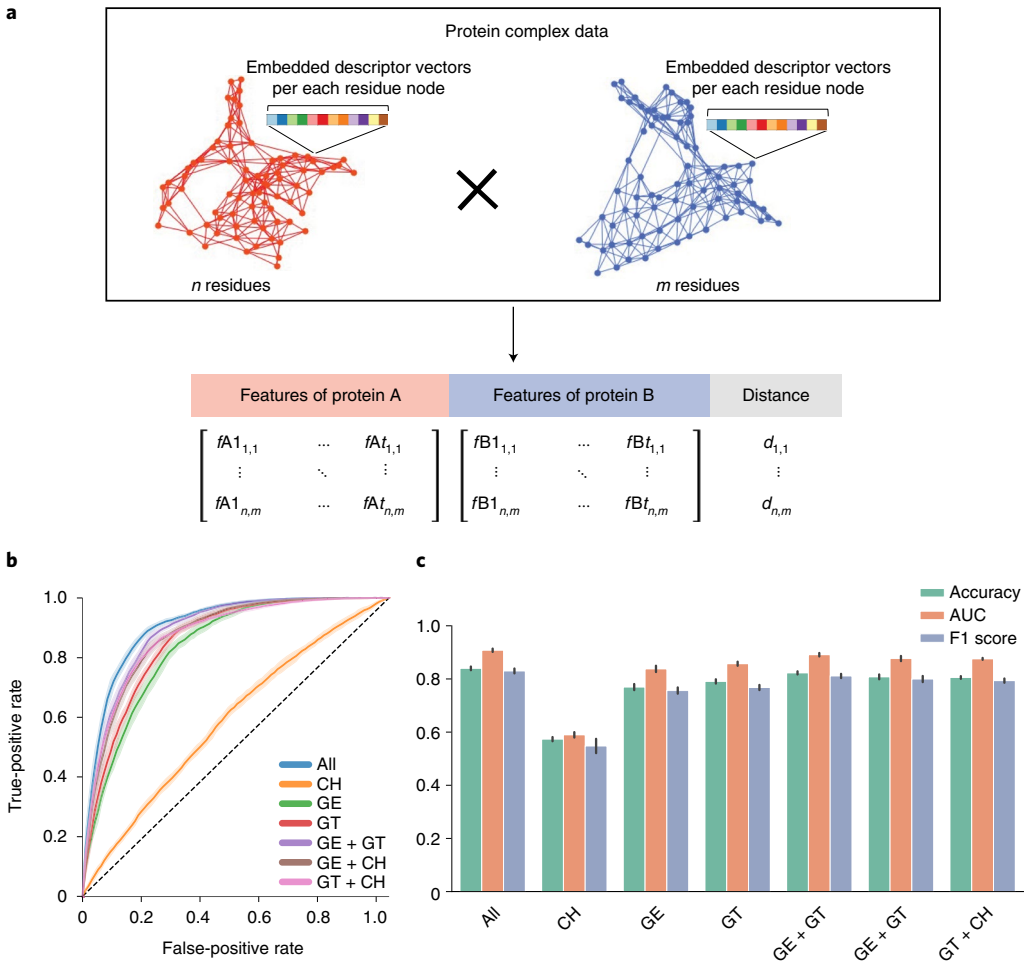

Fig. 3 | Construction of distance-based feature matrices for prediction of protein complexes. a, Schematic explanation of feature matrix extraction from protein complexes. The descriptor vectors are embedded per each AA residue, and the pairwise descriptor sets form the final feature matrix to train the ML algorithms. **b**,**c**, Comparison of the ROC curve (**b**) and model performance metrics (**c**) depending on the descriptor subsets when using the tenfold cross-validated DNN.

the GE + GT versus all the descriptors, both the precision and recall tended to increase for the former, that is, smaller, set of descriptors. Such an effect is related to the 'curse of dimensionality' when selection of the most important uncorrelated features from a larger pool of descriptors increases the accuracy of ML algorithms.

ML predictions for complexes of proteins with NPs. The exclusion of CH components enables the direct application of the ML algorithms trained on PPI to biomimetic NPs whose structure can be enumerated by the same parameters. The atomic structure of NPs is coarse-grained to match the scale of AA residues in proteins. Considering the number of atoms per AA (10 in glycine, 28 in tryptophan) and the relative abundance of different residues, 13 atoms are grouped into a single 'residue' in the coarse-grained representation of carbon nanostructures to obtain D_{NP} ($d_{i,j}$) (Supplementary Section 3.3.5). The $G(n, e)$ of NPs are constructed by taking coarse-grained groups of atoms as nodes and connecting an edge when $d_{i,j}$ is less than 7 Å. Utilizing the GE and GT descriptors, we tested the performance of the ML algorithms to predict protein–NP complexes. We focused primarily on carbon-based nanomaterials, namely graphene quantum dots (GQDs), spherical carbon NPs and single-walled carbon nanotubes (SWNTs) (Supplementary Section 3.3.5) but other types of NP can be coarse-grained in the similar way. Our focus on nanocarbons was also based on the rapid development of this diverse class of biocompatible nanomaterials for biomedical applications, such as drug carriers[38], antibacterials[39], antivirals[40] and high-sensitivity bioanalysis[41,42]. However, research progress in this area is slowed down by difficulties in predicting the enzymatic degradation of nanocarbons[43], the protein coronas[44–46] around them and other types of biochemical processes in the complex milieu of biomolecules.

The ML predictions were compared against experimental and supporting MD simulations from literature[19,47,48]. For the simplest example, we tested the docking of carboxylated GQDs to phenol-soluble modulin-α (PSM-α) peptide (PDB ID 5KHB). The prediction results match well with the MD simulation by displaying GQD docking near the *N*-terminus of the peptides[19] (Fig. 5a). Also, we found that predicted interaction sites in the complex between hydroxylated GQDs and a monomer of human islet amyloid polypeptide (hIAPP, PDB idID 2L86) (Fig. 5b) match the experimental observations established independently from (1) comprehensive liquid chromatography with tandem mass spectrometry (LC–MS/MS) evaluation by Faridi et al.[47] and (2) quenching of the nanostructure autofluorescence in a dose-dependent manner by Wang et al.[48]. However, the interaction sites change drastically when hIAPP fibril is formed (PDB IS 6ZRF) (Fig. 5c). Then, GQDs are bound onto the amyloid's surface, which was again accurately identified in the DNN models using only GT + GE descriptors (Fig. 5c). ML predictions for the complex between short-cut carbon nanotube (CNT, 17 Å length) and human myeloperoxidase (hMPO, PDB ID 1CXP)

Table 1 | Comparison of ML algorithms depending on the training of different descriptors subsets

		All	CH	GE	GT	GE + GT	CH + GE	CH + GT
Logistic regression	Accuracy	0.553 ± 0.005	0.551 ± 0.009	0.553 ± 0.009	0.671 ± 0.009	0.551 ± 0.010	0.551 ± 0.010	0.684 ± 0.008
	AUC	0.565 ± 0.010	0.573 ± 0.009	0.563 ± 0.011	0.730 ± 0.008	0.562 ± 0.013	0.562 ± 0.010	0.743 ± 0.006
	F1 score	0.592 ± 0.007	0.553 ± 0.011	0.592 ± 0.009	0.648 ± 0.009	0.587 ± 0.009	0.589 ± 0.012	0.667 ± 0.009
Gaussian naïve Bayes	Accuracy	0.578 ± 0.010	0.537 ± 0.007	0.574 ± 0.008	0.615 ± 0.010	0.576 ± 0.007	0.575 ± 0.008	0.652 ± 0.009
	AUC	0.633 ± 0.010	0.559 ± 0.008	0.628 ± 0.012	0.711 ± 0.007	0.630 ± 0.005	0.627 ± 0.010	0.709 ± 0.006
	F1 score	0.668 ± 0.010	0.480 ± 0.014	0.665 ± 0.007	0.670 ± 0.006	0.666 ± 0.007	0.666 ± 0.006	0.668 ± 0.010
SVM	Accuracy	0.607 ± 0.012	0.552 ± 0.008	0.606 ± 0.013	0.663 ± 0.014	0.606 ± 0.011	0.605 ± 0.012	0.674 ± 0.010
	AUC	0.656 ± 0.012	0.576 ± 0.013	0.652 ± 0.015	0.719 ± 0.015	0.652 ± 0.009	0.650 ± 0.014	0.741 ± 0.007
	F1 score	0.655 ± 0.010	0.523 ± 0.012	0.655 ± 0.010	0.589 ± 0.016	0.657 ± 0.008	0.656 ± 0.008	0.607 ± 0.012
RF	Accuracy	0.822 ± 0.009	0.561 ± 0.009	0.769 ± 0.010	0.804 ± 0.005	0.810 ± 0.008	0.802 ± 0.008	0.812 ± 0.006
	AUC	0.920 ± 0.006	0.583 ± 0.012	0.869 ± 0.008	0.889 ± 0.004	0.907 ± 0.006	0.899 ± 0.005	0.905 ± 0.004
	F1 score	0.803 ± 0.011	0.545 ± 0.011	0.737 ± 0.012	0.789 ± 0.005	0.789 ± 0.010	0.778 ± 0.010	0.797 ± 0.006
XGB	Accuracy	0.829 ± 0.009	0.559 ± 0.010	0.780 ± 0.011	0.804 ± 0.007	0.815 ± 0.010	0.798 ± 0.010	0.815 ± 0.006
	AUC	0.913 ± 0.006	0.583 ± 0.013	0.862 ± 0.007	0.879 ± 0.006	0.899 ± 0.006	0.885 ± 0.007	0.895 ± 0.005
	F1 score	0.817 ± 0.010	0.542 ± 0.011	0.763 ± 0.014	0.794 ± 0.007	0.803 ± 0.012	0.784 ± 0.014	0.804 ± 0.006
DNN	Accuracy	0.840 ± 0.007	0.574 ± 0.008	0.770 ± 0.012	0.808 ± 0.010	0.823 ± 0.006	0.791 ± 0.009	0.806 ± 0.005
	AUC	0.908 ± 0.007	0.590 ± 0.011	0.838 ± 0.013	0.877 ± 0.007	0.891 ± 0.007	0.857 ± 0.008	0.876 ± 0.004
	F1 score	0.830 ± 0.010	0.548 ± 0.028	0.756 ± 0.012	0.800 ± 0.012	0.811 ± 0.008	0.768 ± 0.010	0794 ± 0.008

The accuracy, AUC and F1 score are obtained from the mean of tenfold cross-validated logistic regression, Gaussian naïve Bayes, SVM, RF, XGB and DNN model. The s.d. is measured as an error. The columns indicate types of descriptor subsets. 'All' stands for the combined CH, GE and GT descriptor sets, while '+' indicates the combination of two sets of descriptors.

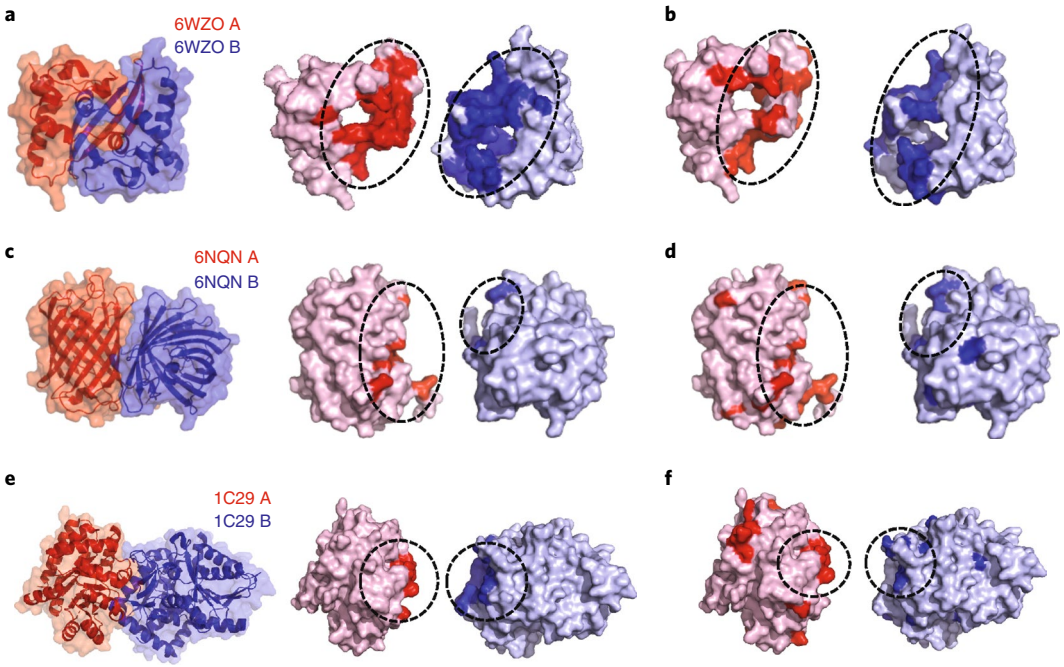

Fig. 4 | Prediction of protein–protein complexes with the DNN model trained on the GE + GT descriptors. a–**f**, The ground-truth interface (**a**,**c**,**e**) and the complex interface predicted by the DNN model trained on GE + GT (**b**,**d**,**f**) of the SARS-CoV-2 dimer protein (PDB ID 6WZO, AB) (**a**,**b**), fluorescent protein Dronpa with the β-barrel structures (PDB ID 6NQN, AB) (**c**,**d**) and bacterial tryptophan synthase having triose-phosphate isomerase (TIM) barrel structure (PDB ID 1C29, AB) (**e**,**f**), with A and B highlighted in red and blue, respectively. The dashed ellipses indicate that the main interaction interfaces of protein chain A and B for both ground truths and predicted ones.

pointed to two interaction sites that were nearly identical to those established by Kagan et al.[43] using MD simulations. Specifically, the tyrosine residues (nos. 293 and 313) and arginine residues (nos. 307 and 294) correctly emerged (Fig. 5d,e) as specific AAs in hMPO closely interacting with CNTs.

The contributions from specific groups and interactions that are stronger than others for particular pairs of NPs and proteins can also be found. These interactions are determined by both the chemical composition and the nanoscale geometry of the interacting species, which is captured well by the combination of GE and

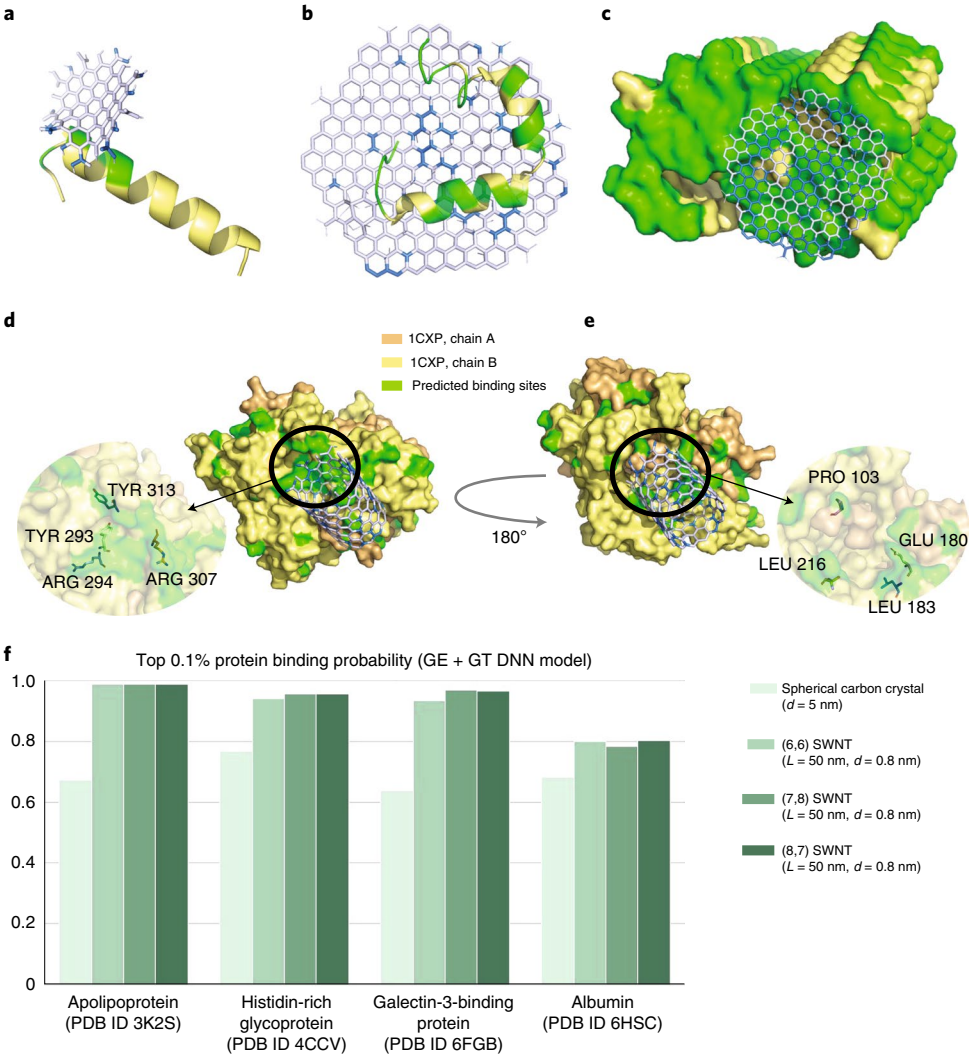

Fig. 5 | The prediction of complexes between protein and nanocarbons. a, Carboxylated GQD and PSM-α assembly. **b**, Hydroxylated GQD and hIAPP assembly. **c**, Hydroxylated GQD and hIAPP fibril assembly. **d**,**e**, The carboxylated CNT and hMPO assembly (**d**) and a view rotated by 180° (**e**), showing the second binding sites. The blue and green highlights in **a**–**e** indicate the predicted interaction sites in NPs and proteins, respectively. Yellow and gray surfaces are non-interacting sites. **f**, Probability of protein binding sites on the different carbon macromolecules, spherical carbon crystal and three different SWNTs.

GT descriptors that account for attractive interactions while minimizing the frustration from molecular reconfiguration. Taking a specific example of the GQD and hIAPP complex, the edges and surface of GQD are hydroxylated, which results in local curvature that can provide a specific fit to the geometry of the protein, capable of forming hydrogen bonds with –OH groups. The GE+GT structural descriptors point to the sites on the GQD that form hydrogen bonds with hIAPP. A similar mechanism can also be traced in the interactions between CNT and hMPO.

While localization of potential interaction sites is essential for NP adjuvants, enzyme mimics, amyloid fibrillation inhibitors and antibiotic and antiviral agents[15,19,39,40], the assessment of the relative propensity of several proteins to interact with NPs of different shapes is also important for nanoscale contrast agents and drug delivery agents. Thus, we also tested the ability of DNN to predict the relative protein abundance in the protein corona around SWNTs as studied by Pinals et al.[44] and spherical carbon NPs as studied by Monopoli et al.[45] and Visalakshan et al.[46]. We found that albumin displays a lower tendency to adsorb on SWNTs than apolipoprotein, histidine-rich glycoprotein or galectin-3-binding protein (Supplementary Figs. 24 and 25). Also, the spherical carbon NPs showed a lower probability of forming a nanoscale complex with all four proteins than three different types of SWNTs. Both findings match recent experimental results for similar nanocarbons established using multiple centrifugation cycles followed by mass spectroscopy, small-angle X-ray scattering and isothermal titration calorimetry (Fig. 5f)[44–46].

These findings becomes particularly useful for the analysis of protein interactions with entirely or predominantly inorganic NPs made from gold[49], ZnO[16] and silica[46] that can acquire a variety of shapes[46]. Thus, we tested ML with pyramidal ZnO NPs (3 nm in the base, 3 nm height) that are known to form a reversible one-on-one complex with β-galactosidase (Supplementary Fig. 27)[16]. The coarse-graining protocol for these NPs was based on the crystal surface atoms. Taking into account the mean size of AAs (3.5–4 Å) and the ionic bond length of ZnO (1.89 Å), two atoms were grouped to produce distance matrices and $G(n, e)$ (Supplementary Section 3.3.5). As a result, we found that the interaction sites responsible for the formation of the complex between ZnO NPs and β-galactosidase are located at the apex and edge of the nanopyramids (Supplementary Fig. 28), which coincides perfectly with experimental data[16].

Discussion

Despite the complexity of intermolecular interactions between nanoscale structures, GE+GT descriptors adequately predict the

formation of complexes and interaction sites for proteins. The same descriptors can be applied directly to NPs. The fact that ML algorithms trained on protein–protein complexes accurately predict the structure of protein–NP complexes provides direct and incontrovertible evidence of the biomimetic nature of water-soluble inorganic NPs known to display a variety of biological functions[16,19,43–46,48].

The chemistry of nanocarbons and other inorganic NPs can be very different from that of proteins. While the dynamics of their complexes with proteins tends to be challenging to model, the developed ML algorithms can streamline their molecular design for specific biomedical or biomanufacturing applications. The nanoscale species' rigidity level can be described by the GNM parameters (Supplementary Fig. 26), which can be calculated rapidly and accurately. Analysis of GNM modes can be instrumental for (1) engineering of molecular rigidity across the spectrum of different nanoscale species and (2) predicting interaction sites at different temperatures. Both tasks can be accomplished by adding physics-based descriptions of thermal motion for various chemical bonds using Boltzmann distributions, which will provide complementary descriptors for biological and abiological nanoscale species.

From a fundamental perspective, these findings extend the boundaries of understanding of the structural requirements for forming lock-and-key interfaces between nanoscale entities and integrate the concepts of topology, Riemannian geometry and multifractality to establish commonalities between them. From a practical perspective, these findings offer a toolbox for the rapid design of abiological nanostructures with specific shapes and surface chemistries for biomedical and other applications.

While the traditional CH descriptors provide limited input to the accuracy of the prediction of protein–protein and protein–NP complexes, we expect that subsequent development of unified CH descriptors inclusive of nonadditivity and collective effects between proteins[50,51] and NPs[52] using, for instance, MD or density functional theory calculations calculated locally will also improve the accuracy of such predictions of interaction sites and affinity constants.

Methods

The atomic coordinates of proteins were acquired from the RCSB protein data bank (https://www.rcsb.org/), and the coordinates for the NPs were modelled by using BIOVIA Materials Studio.

Training database formation. The curated database formed the input of the training dataset, while distance matrices formed the output (Supplementary Fig. 1). The final PPI training set comprised 464 uniquely interacting protein pairs (Supplementary Fig. 4) and 27,859,297 pairs of AA residues in total.

Computation of descriptors. The following descriptors were computed and embedded into each A_i and B_k to form characteristic feature matrices (Supplementary Section 2).

CH descriptors. The probability of AA residues in proteins interacting can be related to their electrostatic charge, hydrophobicity, molecular weight, polarity and atomic composition (Supplementary Figs. 8–10). These chemical parameters determine the repulsive/attractive forces between the macromolecules and are used in nearly all current PPI algorithms[4,53]. The atomic contribution of continuum electrostatic charges per residue is computed using the Chemistry at Harvard Macromolecular Mechanics (CHARMM) force field[26]. In addition to the previously used coarse description of residue charge as −1, 0 or 1, we used a more accurate representation of electrostatic interactions wherein the charge contribution per atom in the residue was considered. Hydrophobicity indices of residue are measured by the Kyte–Doolittle scale[54]. We note that CH descriptors do not account for nonadditivity and interdependence of electrostatic, hydrophobic or van der Waals interactions on the surface of biomolecules[50,51].

GE descriptors. The R_{inacc} descriptor measures the shallowness of the protein surface by calculating the minimum inaccessible radius of the circle at AA residue point A_i and B_k, in C_α coordinates. The Shell descriptor characterizes the depth of a specific AA in the folded protein chain, obtained by quantifying an accessible volume of a particular residue point, A_i and B_k. The Pocket descriptor enumerates the depth and size of a concavity on the surface of the protein globule. This value is inversely proportional to Shell but directly proportional to the pocket radius (Supplementary Section 2.2.1). OPD chirality indices[28] were calculated per each A_i and B_k, considering different distances from AA residue points. Left/right-handed geometries correspond to negative/positive values of OPD. Being conscious of the computational problems emerging for chiral objects with high dimensionality, we restrict OPD calculation to a limited group of N-neighbour residues around a particular residue (Supplementary Section 2.2.2).

GT descriptors. To produce $G(n, e)$, AA residues were connected with an edge when $d_{i,j}$ was less than 7 Å (refs. [55,56]). This cutoff value was chosen because it is larger than the average distance between the AA residues in the single protein (3.8 Å)[57] and corresponds to the segments interacting by supramolecular interactions[58,59] (Supplementary Section 2.3 and Supplementary Table 1). In GNM, the Gaussian modes are decomposed by eigenvalue and eigenvector from the Kirchhoff matrix calculated based on $G(n, e)$ (refs. [31,32]). To obtain the dominant modes, only the square sum of the first tenth of the modes is considered[2] (Supplementary Section 2.3.3 and Supplementary Fig. 15). The ORC evaluates the transport characteristics of the network and, therefore, the stress transfer and reconfigurability in relation to the centre of the molecule, while FRC describes the same characteristics of the periphery of the graph[60,61]. Also, we defined the node-based scalar Ricci curvature as the sum of all edge curvature values on that node[60] (Supplementary Section 2.3.2 and Supplementary Fig. 14). The MFD values are estimated by investigating the power-law behaviour between the partition function, associated with the qth powers of the node-based probability measure of covering the graph with boxes of a specific radius, and the box sizes employed to cover the graph[62,63] (Supplementary Section 2.3.1 and Supplementary Fig. 13).

ML algorithms. The pairwise AA residue descriptor data are used to independently train ML algorithms for seven subsets of {CH, GE, GT} to compare the contributions to the prediction scores. The DNN model consists of three fully connected layers with 512 neurons, with the rectified linear unit (ReLU) function used for activation. After these layers, a dropout layer is added with a rate of 0.5 to prevent overfitting. Lastly, the SoftMax layer is implemented to compute the probability of each class and the model is trained by optimization of the categorical cross-entropy loss function. While training, the best model having the minimum loss and highest accuracy is saved by the callback function. The performance of each ML algorithm is evaluated using tenfold cross-validation. The logistic regression, Gaussian naïve Bayes, SVM and RF approaches are implemented using the scikit-learn Python package[64], while the XGB approach uses the XGBoost Python package[65].

When we test unknown protein–protein complexes with any ML algorithms, the probability of forming a close contact (<7 Å) is computed for every molecule's AA residue pair. We take residue pairs in the top 0.1% by probability as interface residues (Supplementary Section 3 and Supplementary Fig. 19). An identical criterion was applied for protein–NP complexes when assessing D_A ($d_{i,j}$) and to obtain D_{NP} ($d_{i,j}$).

Data availability

Our Code Ocean Capsule[66] contains all the associated data for PPI training and PPI/protein–NP interaction testing. The source data for Figs. 1–5 and Table 1 are provided with this paper.

Code availability

All Python codes associated with this study are deposited in the Code Ocean capsule[66] at https://doi.org/10.24433/CO.7800040.v1.

Received: 23 September 2021; Accepted: 17 March 2022;
Published online: 28 April 2022

References

1. Morrison, J. L., Breitling, R., Higham, D. J. & Gilbert, D. R. A lock-and-key model for protein–protein interactions. *Bioinformatics* **22**, 2012–2019 (2006).
2. Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. & Gursoy, A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Res.* **42**, W285 (2014).
3. Murakami, Y. & Mizuguchi, K. Applying the naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* **26**, 1841–1848 (2010).
4. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
5. Montoya, M. A PrePPI way to make predictions. *Nat. Struct. Mol. Biol.* **19**, 1067 (2012).
6. Northey, T. C., Bareši, A. & Martin, A. C. R. IntPred: a structure-based predictor of protein–protein interaction sites. *Bioinformatics* **34**, 223–229 (2018).
7. Baranwal, M. et al. Struct2Graph: a graph attention network for structure based predictions of protein–protein interactions. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.17.301200 (2020).
8. Chen, K.-H., Wang, T.-F. & Hu, Y.-J. Protein–protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. *BMC Bioinformatics* **20**, 308 (2019).

9. Sarkar, D. & Saha, S. Machine-learning techniques for the prediction of protein–protein interactions. *J. Biosci.* **44**, 104 (2019).
10. Wang, Y. et al. Predicting protein interactions using a deep learning method-stacked sparse autoencoder combined with a probabilistic classification vector machine. *Complexity* **2018**, 4216813 (2018).
11. Kotov, N. A. Inorganic nanoparticles as protein mimics. *Science* **330**, 188–189 (2010).
12. Pinals, R. L., Chio, L., Ledesma, F. & Landry, M. P. Engineering at the nano–bio interface: harnessing the protein corona towards nanoparticle design and function. *Analyst* **145**, 5090–5112 (2020).
13. Govan, J. & Gun'ko, Y. K. Recent progress in chiral inorganic nanostructures. *Nanoscience* **3**, 1–30 (2016).
14. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
15. Xu, L. et al. Enantiomer-dependent immunological response to chiral nanoparticles. *Nature* **601**, 366–373 (2022).
16. Cha, S.-H. et al. Shape-dependent biomimetic inhibition of enzyme by nanoparticles and their antibacterial activity. *ACS Nano* **9**, 9097–9105 (2015).
17. Ravikumar, K. M., Huang, W. & Yang, S. Coarse-grained simulations of protein–protein association: an energy landscape perspective. *Biophys. J.* **103**, 837–845 (2012).
18. Kmiecik, S. et al. Coarse-grained protein models and their applications. *Chem. Rev.* **116**, 7898–7936 (2016).
19. Wang, Y. et al. Anti-biofilm activity of graphene quantum dots via self-assembly with bacterial amyloid proteins. *ACS Nano* **13**, 4278–4289 (2019).
20. Acosta-Tapia, N., Galindo, J. F. & Baldiris, R. Insights into the effect of Lowe syndrome-causing mutation p.Asn591Lys of OCRL-1 through protein–protein interaction networks and molecular dynamics simulations. *J. Chem. Inf. Model.* **60**, 1019–1027 (2020).
21. Verma, M. K. & Shakya, S. LRP-1 mediated endocytosis of EFE across the blood–brain barrier; protein–protein interaction and molecular dynamics analysis. *Int. J. Pept. Res. Ther.* **27**, 71–81 (2021).
22. Li, Z. L. & Buck, M. Modified potential functions result in enhanced predictions of a protein complex by all-atom molecular dynamics simulations, confirming a stepwise association process for native protein–protein interactions. *J. Chem. Theory Comput.* **15**, 4318–4331 (2019).
23. Liu, Y. et al. A compact biosensor for binding kinetics analysis of protein–protein interaction. *IEEE Sens. J.* **19**, 11955–11960 (2019).
24. Moscetti, I., Cannistraro, S. & Bizzarri, A. R. Surface plasmon resonance sensing of biorecognition interactions within the tumor suppressor P53 network. *Sensors* https://doi.org/10.3390/s17112680 (2017).
25. Verboven, C. et al. Actin-DBP: the perfect structural fit? *Acta Crystallogr. D* **59**, 263–273 (2003).
26. Dolinsky, T. J. et al. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, 522–525 (2007).
27. Kawabata, T. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **78**, 1195–1211 (2010).
28. Osipov, M. A., Pickup, B. T. & Dunmur, D. A. A new twist to molecular chirality: intrinsic chirality indices. *Mol. Phys.* **84**, 1193–1206 (1995).
29. May, A. et al. Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics* **30**, 326–334 (2014).
30. Vishveshwara, S., Brinda, K. V. & Kannan, N. Protein structure: insights from graph theory. *J. Theor. Comput. Chem.* **1**, 187–211 (2002).
31. Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181 (1997).
32. Haliloglu, T., Bahar, I. & Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79**, 3090–3093 (1997).
33. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, 1395–1406 (2006).
34. Gavin, A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
35. Ye, Q., West, A. M. V., Silletti, S. & Corbett, K. D. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci.* **29**, 1890–1901 (2020).
36. Romei, M. G., Lin, C., Mathews, I. I. & Boxer, S. G. Electrostatic control of photoisomerization pathways in proteins. *Science* **367**, 76–79 (2020).
37. Sachpatzidis, A. et al. Crystallographic studies of phosphonate-based α-reaction transition-state analogues complexed to tryptophan synthase. *Biochemistry* **38**, 12665–12674 (1999).
38. Ju, J., Regmi, S., Fu, A., Lim, S. & Liu, Q. Graphene quantum dot based charge-reversal nanomaterial for nucleus-targeted drug delivery and efficiency controllable photodynamic therapy. *J. Biophoton.* **12**, e201800367 (2019).
39. Ahmed, K. B. A., Raman, T. & Veerappan, A. Future prospects of antibacterial metal nanoparticles as enzyme inhibitor. *Mater. Sci. Eng. C* **68**, 939–947 (2016).
40. Unal, M. A. et al. Graphene oxide nanosheets interact and interfere with SARS-CoV-2 surface proteins and cell receptors to inhibit infectivity. *Small* **17**, 2101483 (2021).
41. Blanco-López, M. C. & Rivas, M. Nanoparticles for bioanalysis. *Anal. Bioanal. Chem.* **411**, 1789–1790 (2019).
42. Ma, W. et al. Attomolar DNA detection with chiral nanorod assemblies. *Nat. Commun.* **4**, 2689 (2013).
43. Kagan, V. E. et al. Carbon nanotubes degraded by neutrophil myeloperoxidase induce less pulmonary inflammation. *Nat. Nanotechnol.* **5**, 354–359 (2010).
44. Pinals, R. L. et al. Quantitative protein corona composition and dynamics on carbon nanotubes in biological environments. *Angew. Chem. Int. Ed.* **59**, 23668–23677 (2020).
45. Monopoli, M. P., Pitek, A. S., Lynch, I. & Dawson, K. A. Formation and characterization of the nanoparticle–protein corona. *Methods Mol. Biol.* **1025**, 137–155 (2013).
46. Madathiparambil Visalakshan, R. et al. The influence of nanoparticle shape on protein corona formation. *Small* https://doi.org/10.1002/smll.202000285 (2020).
47. Faridi, A. et al. Graphene quantum dots rescue protein dysregulation of pancreatic β-cells exposed to human islet amyloid polypeptide. *Nano Res.* **12**, 2827–2834 (2019).
48. Wang, M. et al. Graphene quantum dots against human IAPP aggregation and toxicity: in vivo. *Nanoscale* **10**, 19995–20006 (2018).
49. Lin, W. et al. Control of protein orientation on gold nanoparticles. *J. Phys. Chem. C* **119**, 21035–21043 (2015).
50. Ma, C. D., Wang, C., Acevedo-Vélez, C., Gellman, S. H. & Abbott, N. L. Modulation of hydrophobic interactions by proximally immobilized ions. *Nature* **517**, 347–350 (2015).
51. Horovitz, A. Non-additivity in protein–protein interactions. *J. Mol. Biol.* **196**, 733–735 (1987).
52. Batista, C. A. S. et al. Nonadditivity of nanoparticle interactions. *Science* **350**, https://doi.org/10.1126/science.1242477 (2015).
53. Qiao, Y., Xiong, Y., Gao, H., Zhu, X. & Chen, P. Protein–protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinformatics* **19**, 14 (2018).
54. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
55. Jumper, J. M., Faruk, N. F., Freed, K. F. & Sosnick, T. R. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS Comput. Biol.* **14**, e1006342 (2018).
56. Chakrabarty, B., Naganathan, V., Garg, K., Agarwal, Y. & Parekh, N. NAPS update: network analysis of molecular dynamics data and protein–nucleic acid complexes. *Nucleic Acids Res.* **47**, W462–W470 (2019).
57. Chakraborty, S., Venkatramani, R., Rao, B. J., Asgeirsson, B. & Dandekar, A. M. Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms. *F1000Res.* **2**, 1–12 (2013).
58. Brancolini, G. & Tozzini, V. Multiscale modeling of proteins interaction with functionalized nanoparticles. *Curr. Opin. Colloid Interface Sci.* **41**, 66–73 (2019).
59. Hazarika, Z. & Jha, A. N. Computational analysis of the silver nanoparticle–human serum albumin complex. *ACS Omega* **5**, 170–178 (2020).
60. Samal, A. et al. Comparative analysis of two siscretizations of Ricci curvature for complex networks. *Sci. Rep.* **8**, 8650 (2018).
61. Eidi, M. & Jost, J. Ollivier Ricci curvature of directed hypergraphs. *Sci. Rep.* **10**, 12466 (2020).
62. Yang, R. & Bogdan, P. Controlling the multifractal generating measures of complex networks. *Sci. Rep.* **10**, 5541 (2020).
63. Xiao, X., Chen, H. & Bogdan, P. Deciphering the generating rules and functionalities of complex networks. *Sci. Rep.* **11**, 22964 (2021).
64. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
65. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
66. Cha, M. et al. Unifying structural descriptors for biological and bioinspired nanoscale complexes [source code]. *Code Ocean* https://doi.org/10.24433/CO.7800040.V1 (2022).

Acknowledgements

We thank the University of Michigan College of Engineering for support through the BlueSky Initiative and the University of Michigan for access to the HPC resources of the Great Lakes Cluster. Support from the Vannevar Bush DoD Fellowship to N.A.K. ('Engineered Chiral Ceramics' ONR N000141812876, ONR COVID-19 Newton Award 'Pathways to Complexity with 'Imperfect' Nanoparticles' HQ00342010033 and AFOSR FA9550-20-1-0265 'Graph Theory Description of Network Material') is gratefully acknowledged. X.X. and P.B. gratefully acknowledge the support by the National Science Foundation Career award under grant number CPS/CNS-1453860, the NSF awards under grant numbers CCF-1837131, MCB-1936775, CNS-1932620 and CMMI-1936624, the Okawa Foundation research award, the Defense Advanced Research Projects Agency

(DARPA) Young Faculty Award and DARPA Director Award under grant number N66001-17-1-4044, a 2021 USC Stevens Center Technology Advancement Grant (TAG) award, an Intel faculty award and a Northrop Grumman grant. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied by the Defense Advanced Research Projects Agency, the Department of Defense or the National Science Foundation.

Author contributions

N.A.K conceived the project. M.C., E.S.T.E. and N.A.K. designed the descriptor sets and the workflow. E.S.T.E. collected and curated the protein complex dataset. M.C. analysed the protein complex data and computed the CH, GE and GT descriptors. J.-Y.K. contributed to OPD index calculation. X.X. and P.B. contributed to the computation of MFD in GT features. M.C., X.X. and P.B. designed and trained the DNN model and carried out comparative studies of different ML models. M.C. visualized the analysed data. M.C., E.S.T.E. and N.A.K co-wrote the paper. All authors contributed to data analysis, discussion and writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00229-w.

Correspondence and requests for materials should be addressed to Nicholas A. Kotov.

Peer review information *Nature Computational Science* thanks Ning Gu, Häkkinen Hannu and Açelya Yilmazer for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.