Predicting the Physicochemical Properties and Biological Activities of Monolayer-Protected Gold Nanoparticles using Simulation-Derived Descriptors

*Alex K. Chew[1], Joel A. Pedersen[2,3], and Reid C. Van Lehn[1]**

[1]Department of Chemical and Biological Engineering, University of Wisconsin – Madison, Madison, WI, 53706, United States.

[2]Department of Chemistry, University of Wisconsin – Madison, Madison, WI, 53706, United States.

[3]Department of Soil Science and Civil and Environmental Engineering, University of Wisconsin – Madison, Madison, WI, 53706, United States.

*send correspondence to: vanlehn@wisc.edu

ABSTRACT

Gold nanoparticles are versatile materials for biological applications because their properties can be modulated by assembling ligands on their surface to form monolayers. However, the physicochemical properties and behaviors of monolayer-protected nanoparticles in biological environments are difficult to anticipate because they emerge from the interplay of ligand–ligand and ligand–solvent interactions that cannot be readily inferred from ligand chemical structure alone. In this work, we demonstrate that quantitative nanostructure–activity relationship (QNAR) models can employ descriptors calculated from molecular dynamics simulations to predict nanoparticle properties and cellular uptake. We performed atomistic molecular dynamics simulations of 154 monolayer-protected gold nanoparticles and calculated a small library of simulation-derived descriptors that capture nanoparticle structural and chemical properties in aqueous solution. We then parameterized QNAR models using interpretable regression algorithms to predict experimental measurements of nanoparticle octanol–water partition coefficients, zeta potentials, and cellular uptake obtained from a curated database. These models reveal that simulation-derived descriptors can accurately predict experimental trends and provide physical insight into what descriptors are most important for obtaining desired nanoparticle properties or behaviors in biological environments. Finally, we demonstrate model generalizability by predicting cell uptake trends for 12 nanoparticles not included in the original data set. These results demonstrate that QNAR models parameterized with simulation-derived descriptors are accurate, generalizable computational tools that could be used to guide the design of monolayer-protected gold nanoparticles for biological applications without laborious trial-and-error experimentation.

KEYWORDS

molecular dynamics simulations, quantitative nanostructure–activity relationship, self-assembled monolayers, gold nanoparticles, machine learning

INTRODUCTION

Gold nanoparticles (GNPs) are promising materials for applications including drug delivery, biosensing, and photothermal therapy because their interactions with biological materials can be tailored by assembling organic ligands on the GNP surface to form a self-assembled monolayer (SAM).[1-3] Small GNPs with core diameters less than 10 nm are of particular interest because this size limits renal clearance from the body.[4] This size is also commensurate with that of typical biomolecules (*e.g.*, proteins) and the thickness of the cell membrane, enabling the study of nano–bio interactions between materials at comparable length scales.[5] Because such interactions can be modulated by tailoring the ligand composition of the protecting SAM, substantial effort has been devoted to understanding how ligand properties impact interactions at the nano–bio interface. For example, synthetic modification of ligand hydrophobicity has been shown to impact GNP immune response,[6] cellular uptake,[7] and binding affinities for proteins[8] and lipid bilayers[9]. Positively charged ligand end groups increase electrostatic interactions between small GNPs and negatively charged lipid bilayers, resulting in GNP adsorption to, disruption of, or insertion within the bilayer,[10-12] whereas negatively charged ligand end groups largely promote adsorption.[10, 12-14] Zwitterionic ligands have been shown to prevent strong protein adsorption onto the GNP surface, which could minimize the formation of the protein corona.[15] These studies highlight the structure–property relationships that have been identified for broad categorizations of ligand properties (*e.g.*, cationic vs. anionic vs. zwitterionic). However, many potential ligand chemical structures fit within these categories. Consequently, engineering small GNPs for targeted nano–bio interactions, such as selective protein binding or favorable cell uptake, remains challenging because subtle

variations to the gold core size and ligand properties can manifest as substantial changes to GNP behavior,[6, 10, 15-23] which are challenging to predict *a priori*.[24] Experimental exploration of the vast design space of possible GNP compositions is also time-consuming and difficult, in part because of the challenge in experimentally resolving molecular-level features of GNPs, such as the complex, non-planar geometries that arise from the interactions between ligands adsorbed to small GNPs and influence interactions with the surrounding environment.[25-26]

To complement experimental methods, computational modeling can be used to derive quantitative nanostructure–activity relationship (QNAR) models for the rational design of GNPs without extensive trial-and-error experimentation.[27-28] QNAR models use numerical parameters (descriptors) that capture characteristics of GNPs and relate them to the behavior of GNPs in biological environments. Descriptors typically consist of experimentally measured quantities (*e.g.,* size, shape, zeta potentials)[27] or single-molecule descriptors of organic ligands (*e.g.*, constitutional, topological, electrostatic) that are often used in drug discovery applications.[29] Descriptors are then related to relevant labels through a variety of machine learning algorithms (*e.g.*, multilinear regression, support vector machines).[28] A key challenge in the development of these models is determining an appropriate set of descriptors: experimental descriptors are challenging to measure for a large range of GNPs and single-molecule descriptors do not account for the collective properties of many organic and inorganic molecules contained within a SAM.[28] Recent studies have sought to address these issues by developing virtual GNP models.[30] These models are static atomistic representations of GNPs consisting of a gold core with chemically specific ligands randomly placed on the gold core with densities selected to mimic experimental measurements.[30] Virtual GNPs better capture properties such as the gold core size, ligand density, and surface chemistry than single-molecule descriptors while providing access to molecular-scale information

that is not easily resolved experimentally. Descriptors were developed to capture surface properties of virtual GNPs and used to develop QNAR models that could predict both biophysiochemical properties (*e.g.*, log *P*, zeta potentials) and behaviors emerging from nano–bio interactions (*e.g.*, cell uptake, GNP-enzyme binding).[30-33] Furthermore, deep learning methods, such as convolutional neural networks, have been used to analyze virtual GNPs without requiring descriptor calculations.[34] However, these static models do not account for changes to SAM structure that emerge from interactions between ligands and the surrounding solvent environment that could influence GNP behavior. Moreover, the descriptors developed for virtual GNPs are challenging to physically interpret and may not generalize to other classes of materials.

Molecular dynamics (MD) simulations are an alternative computational method to gain atomistic insight into both the structure and dynamics of GNPs in explicit solvent.[9, 26, 35-40] Unlike virtual GNPs, atomistic MD simulations of GNPs can model the formation of anisotropic structures that emerge from the interplay of ligand–ligand and ligand–solvent interactions. For example, GNPs protected by long, nonpolar alkanethiol ligands were experimentally[41] and computationally[35-40] found to form bundles, in which ligands align in the same direction due to preferred interactions between methylene moieties. These bundles dictate how GNPs bind with one another to minimize solvent-exposed hydrophobic surface area,[38, 42] how GNPs bend single-stranded nucleic acids,[43] and how spatially heterogenous surface properties arise from chemically homogeneous SAMs.[35] These simulations also permit analysis of GNPs in the presence of lipid bilayers or proteins,[9, 11, 16, 20-21, 44-47] allowing for in-depth mechanistic studies that are not possible with virtual GNPs. While these past studies, along with many other studies of SAM-protected GNPs,[16, 20-21, 24, 26, 48] have provided useful insights into the interplay of gold core and ligand selection on SAM properties that influence GNP behavior with other biomolecules, they have

focused primarily on mechanistic studies for a limited subset of GNPs and the integration of MD simulations with QNAR modeling has yet to be explored.

Based on these observations, we hypothesize that MD simulations can more accurately capture GNP properties than can static models, potentially improving QNAR predictions of experimental data. To test this hypothesis, we modeled 154 sub-10-nm GNPs in aqueous solution using atomistic MD simulations and developed a small library of 25 MD-derived descriptors that characterize GNP structural and chemical properties. We used two interpretable regression techniques — least absolute shrinkage and selection operator (LASSO) regression and random forest (RF) — to develop QNAR models that accept MD-derived descriptors as input and predict experimental log *P*, cell uptake, and zeta potentials from Ref. 32. The resulting regression models accurately predict experimental data based on only a small number of physically interpretable descriptors that are generalizable to different classes of GNPs or other nanomaterials. We then identify the most important descriptors that relate to these experimental observables to provide general guidelines for the design of GNPs. Finally, we show that the RF cell uptake model correctly generalizes to capture cell uptake trends in a separate dataset consisting of 12 GNPs.[17] These results demonstrate that QNAR models parameterized with simulation-derived descriptors are computationally efficient tools to predict GNP behavior *a priori*, thereby enabling the rational design of bioactive GNPs.

RESULTS AND DISCUSSION

Experimental datasets used to develop QNAR models

To develop QNAR models, we obtained data from a curated database of GNPs that are protected by either single- or multicomponent SAMs consisting of ligands with the general structure shown in Figure 1a.[32] Ligand backbones and end groups are both varied to provide

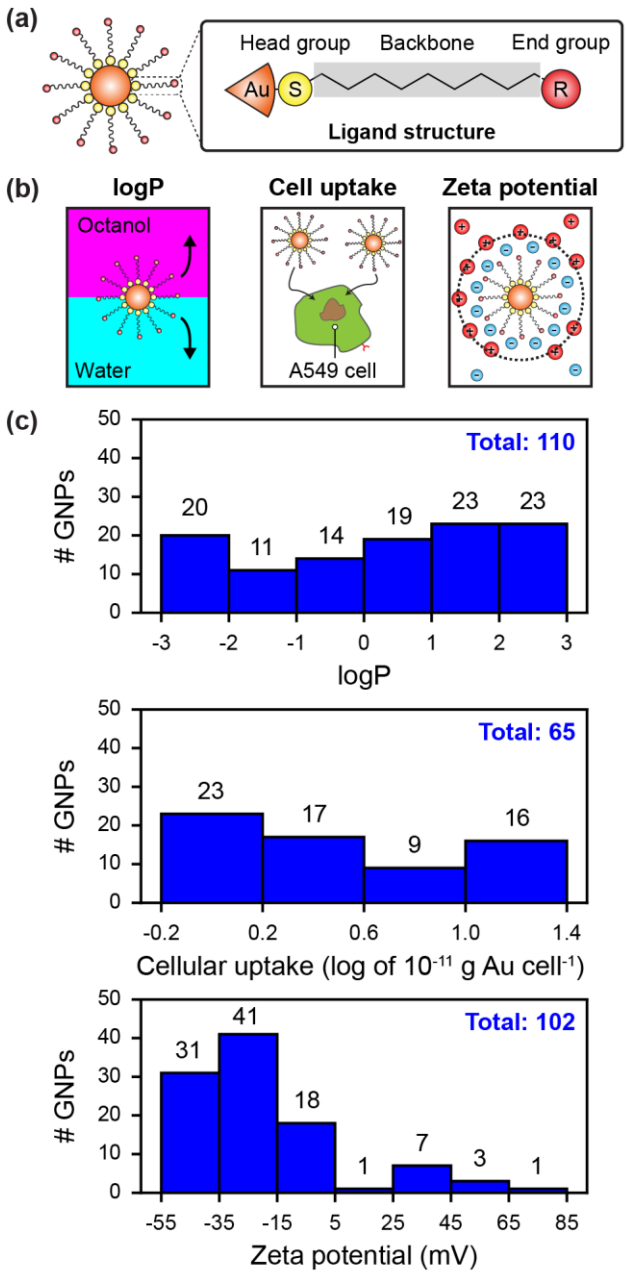

Figure 1. Experimental data used to develop QNAR models. (a) Typical ligand structure consisting of a sulfur head group, nonpolar backbone, and end group. One or two ligand components are assembled on the gold core to form a monolayer. **(b)** Schematics of the three experimental observables used to characterize gold nanoparticle (GNP) physicochemical properties and bioactivity. **(c)** Number distributions of GNPs labeled with each experimental observable. The total number of GNPs for each observable is written at the upper right. All experimental data were taken from Ref. 32.

structural and chemical diversity in the resulting SAMs. The database also includes experimental data for the core diameter, ligand structure, and ligand surface density of each GNP. We selected

spherical GNPs less than 10 nm in core diameter, resulting in 154 GNPs (including 96 single- and 58 multi-component SAMs) encompassing 105 distinct ligands and core diameters ranging between 2 and 8.5 nm.

To characterize GNP biophysiochemical properties and behavior in biological environments, the database includes experimental measurements of GNP octanol–water partition coefficient (log *P*) values, uptake into A549 cells, and zeta potentials, although all three values are not available for all GNPs.[32] These three GNP observables are schematically summarized in Figure 1b. Log *P* quantifies the partitioning of GNPs between *n*-octanol and water phases; a larger value of log *P* indicates that the GNP prefers the octanol phase more than the water phase, suggesting that the surface is hydrophobic. Log *P* thus provides information on lipophilicity which can influence GNP interactions with lipid bilayers.[9, 46, 49] Log *P* values were determined by measuring the partitioning of GNPs to the *n*-octanol and water phases using inductively coupled plasma mass spectrometry (ICP-MS) after 24 hours of shaking and 3 hours of relaxation (we note that it is possible that GNPs also partition to the interface between two phases, but this behavior is not readily captured in the log *P* measurement).[32] Uptake into A549 cells (a lung epithelial cancer cell line) measures nonspecific internalization after incubating 50 μg/mL of GNPs for 24 hours and is quantified by the mass of internalized gold per cell measured via ICP-MS.[32] Finally, the zeta potential in water measures the electric potential of the GNP at the shear plane as determined by electrophoresis; higher magnitude zeta potential values correspond to colloidally stable GNP suspensions.[50] The zeta potential provides information on electrostatic interactions in biological environments that can influence the adsorption of proteins, adsorption to the cell membrane, cell uptake, and cytotoxicity.[51-52] Zeta potentials were measured in water at pH 7 using a Malvern Zetasizer.[32] Figure 1c shows the number distribution of GNPs for the different experimental

values; in total, there are 110 GNPs with log *P* values, 65 GNPs with cell uptake values, and 102 GNPs with zeta potentials. Values for log *P* and cell uptake are well-distributed, whereas zeta potentials are skewed toward negative values. These experimental observables provide complementary insight into the behavior of GNPs within a biological environment and are thus suitable labels for developing QNAR models for GNPs intended for biological applications.

Computational workflow to compute MD-derived descriptors of GNP properties

To model the large set of GNPs, we modified our previously developed workflow for constructing atomistic models of GNPs with desired gold core shapes, sizes, and ligand selection.[36] Each GNP is specified by the diameter of its gold core (modeled as spherical), simplified molecular-input line-entry (SMILES) strings for each type of ligand in the protecting SAM, and the total number of ligands in the SAM. Figure 2 summarizes the workflow using GNP1 and GNP288 (GNP nomenclature follows that of Ref. 32) as representative examples that have distinct gold core sizes and ligand structures. Ligands within the GNP database[32] have one of two substructure patterns: (1) butanethiol (SMILES: SCCCC) and (2) 1,2-dithiolane (SMILES: C1CCSS1). Accordingly, we positioned ligands on the gold surface by placing an excess of either butanethiol or 1,2-dithiolane molecules around the gold core and permitting them to self-assemble onto the surface *via* a strong sulfur–gold Lennard-Jones (LJ) interaction.[36, 53] Adsorbed substructures were replaced with the desired ligand structures from the database. The resulting GNP was simulated in the presence of water molecules and sodium or chloride counterions (if necessary) for 50 ns in the *NPT* ensemble (*vide infra* for discussion of convergence within this timeframe). We used this workflow to

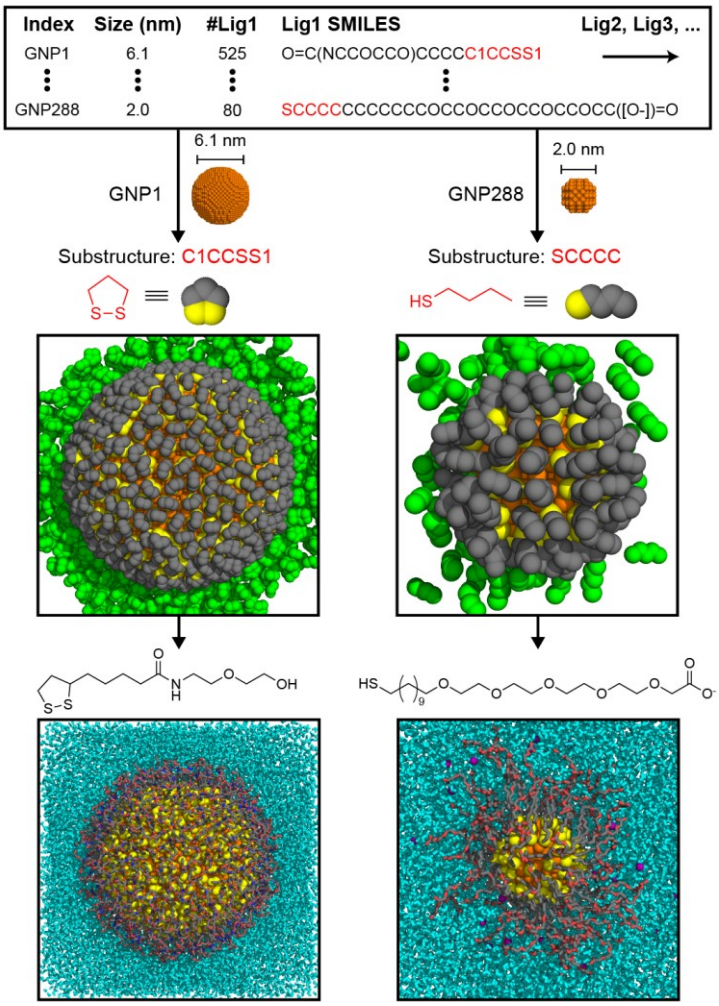

Figure 2. Workflow for modeling GNPs using MD simulations. Each GNP is selected from a database that specifies the GNP diameter, number of ligands adsorbed on the surface, and SMILES string for each type of ligand. The diameter and ligand substructure, either butanethiol (SCCCC) or 1,2-dithiolane (C1CCSS1), were used to initiate simulations in which model substructures adsorb to the gold core. Excess substructures (green) were removed and adsorbed substructures (gray) were replaced with the desired ligands. The GNP was then simulated in pure water (cyan) with sodium or chlorine counterions as needed to ensure charge-neutral systems. This schematic uses GNP1 and GNP288 as representative examples.

systematically model 154 GNPs and compute MD-derived descriptors. Additional details on the simulation workflow are included in the Methods and Supporting Information (Figures S1-S2).

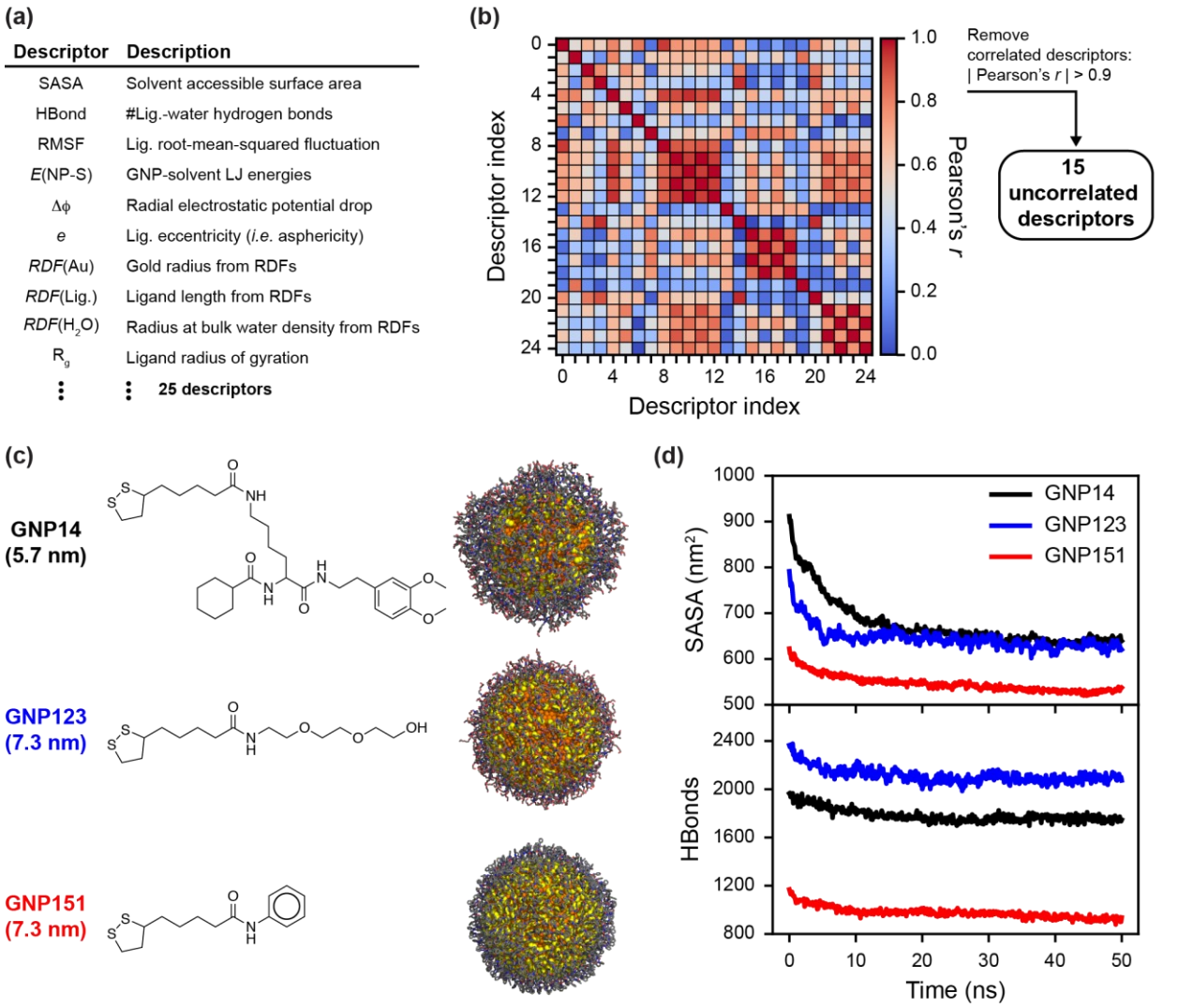

Figure 3. Summary of MD-derived descriptors. (a) List of selected MD-derived descriptors. All 25 descriptors are described in Table S1. **(b)** Matrix of Pearson's *r* values between all pairs of descriptors. Values are colored based on the absolute value of *r*; highly correlated pairs are colored in red, whereas uncorrelated descriptors are blue. Descriptor indexes correspond to Table S1. Fifteen uncorrelated descriptors were identified after removing one descriptor from each highly correlated pair (defined as $|r| > 0.90$). **(c)** Three example GNPs with varying core diameters (in parentheses) and ligand structures. Simulation snapshots are shown without water molecules. **(d)** Solvent-accessible-surface area (SASA) and number of ligand–water hydrogen bonds (HBonds) *versus* simulation time for the three GNPs.

We developed a library of 25 physically motivated descriptors that capture structural and chemical properties of GNPs based on previous studies.[30-31, 33, 36, 54-55] Figure 3a shows a truncated list of these descriptors; a full list is available in Table S1. Correlated descriptors were removed

by computing the Pearson's *r* correlation matrix between the 25 descriptors (Figure 3b). A value of $|r| \approx 1$ for a pair of descriptors indicates that they are highly correlated and provide redundant information. Consequently, one descriptor from each pair of descriptors for which $|r| > 0.90$ was removed, resulting in 15 uncorrelated descriptors (listed in Table S2) suitable for parameterizing QNAR models. These descriptors capture structural characteristics arising from the interplay of the gold core and ligand selection that may not be inferred from ligand structure or core diameter alone. To illustrate this, Figure 3c shows three example GNPs (GNP14, GNP123, and GNP151) with varying core diameters and ligand structures; the simulation snapshots illustrate the complex SAM geometries obtained in aqueous solution. Figure 3d plots two representative descriptors — the solvent-accessible-surface area (SASA) and the number of ligand–water hydrogen bonds (HBonds) — *versus* simulation time for each GNP. GNP14 has the smallest core diameter, but it has approximately the same SASA as GNP123 due to the longer ligand attached on GNP14. GNP123 has the largest number of ligand–water hydrogen bonds even though the ligand on GNP123 has fewer oxygen and nitrogen atoms compared to GNP14. These representative results highlight how MD-derived descriptors can capture non-obvious characteristics that emerge from the interplay of ligand–ligand and ligand–water interactions. Moreover, they confirm the convergence of descriptor calculations within reasonably rapid MD timescales (< 50 ns). Details on descriptor calculations and convergence are included in the Supporting Information and Figures S3-S5.

QNAR models using MD-derived descriptors accurately predict experimental trends

We sought to develop QNAR models that use the uncorrelated descriptors from Table S2 as input to predict the selected experimental labels (*i.e.*, log *P*, cell uptake, zeta potential). We compared LASSO and RF regression algorithms to probe the prediction capabilities of MD-

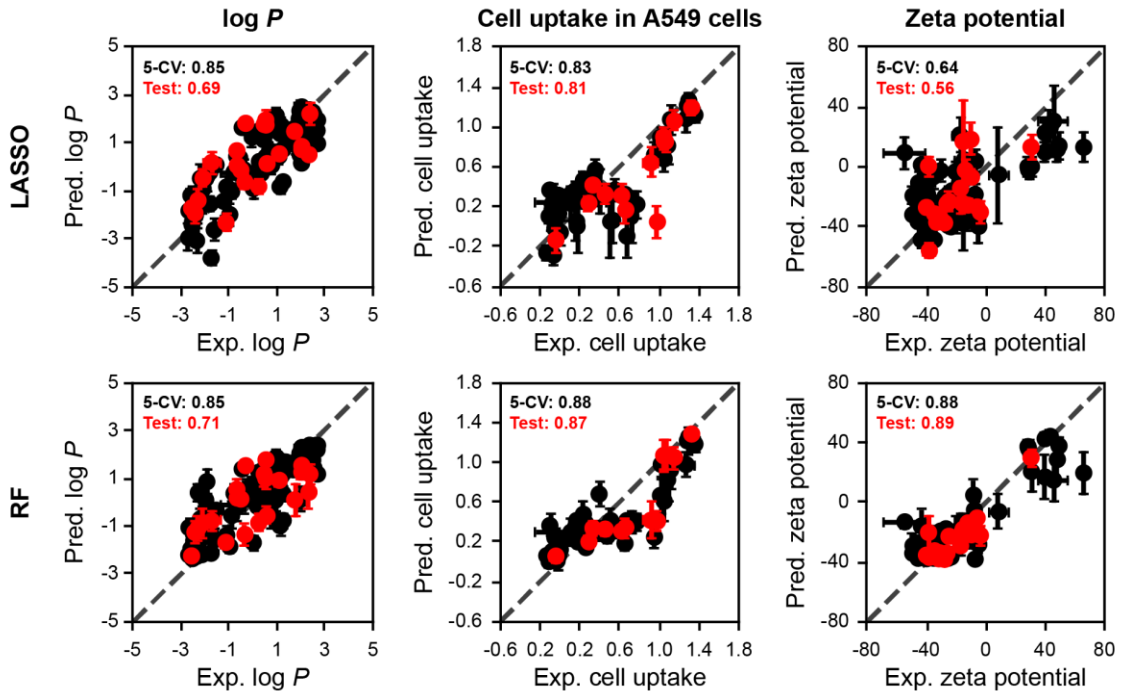

Figure 4. Prediction accuracy of QNAR models. Parity plots of predicted *versus* experimental log *P* (left), cell uptake (middle), and zeta potential (right) values are shown for LASSO (top) and random forest (bottom) models. 5-fold cross validation (5-CV) predictions are shown in black and test set predictions are shown in red. Pearson's *r* values that measure correlations between predicted and experimental values are shown in the upper left with the same color scheme. Units for all measurements are the same as Figure 1. Experimental data and error bars were taken from Ref. 32. Predicted values and error bars were estimated using a bagging approach with either 20 LASSO or RF models; the average of the predictions is reported, and the error is estimated by the standard deviation of the predictions.

derived descriptors. LASSO is a linear regression model that minimizes the residual sum squared and the sum of the absolute value of the regression weights. Compared to a typical multiple linear regression model, an advantage of LASSO is its ability to remove descriptors that do not significantly contribute to the prediction of the experimental observable, which is useful for identifying the most important descriptors that contribute to predictions. RF is a non-linear model consisting of an ensemble of decision trees that are each trained using different subsets of the training data; these trees then collectively vote on a predicted output value. To test the ability of the models to generalize to unseen data, we held out 20% of the experimental labels as a test set

and further performed 5-fold cross validation (5-CV) of the remaining training set as further described in the Methods.

Figure 4 compares predicted and experimental log *P*, cell uptake, and zeta potential values for QNAR models trained using LASSO and RF. Both 5-CV and test set predictions are shown and labeled with the computed value of Pearson's *r* as a measure of linear correlation between the predicted and experimental values. Values of Pearson's *r* close to 1 indicate that the QNAR model accurately predicts experimental trends. For log *P* and cell uptake data sets, the LASSO models perform well with $r > 0.8$ for the 5-CV data and slightly diminished test set performance with $r \geq 0.69$. However, the LASSO model performed poorly when predicting zeta potentials with $r = 0.64$ for 5-CV and $r = 0.56$ for the test set. Compared to the LASSO models, the nonlinear RF models improved prediction accuracy for all three data sets based upon comparison of Pearson's *r* for the 5-CV and test set data, with modest improvements obtained for the log *P* and cell uptake data sets but substantial improvement for the zeta potential data set ($r \geq 0.85$ for the 5-CV data and $r \geq 0.71$ for the test set data). Together, these results show that MD-derived descriptors can be used to predict experimentally determined GNP properties and cellular uptake across a diverse set of GNPs using a small set of descriptors (15) compared to the >600 used in prior virtual GNP studies.[32] The QNAR predictions also capture experimental trends even though the GNPs are modeled in aqueous solution rather than the environments corresponding to the experimental measurements (*e.g.*, simulations of GNPs interacting with the cell membrane are not required to predict cell uptake).

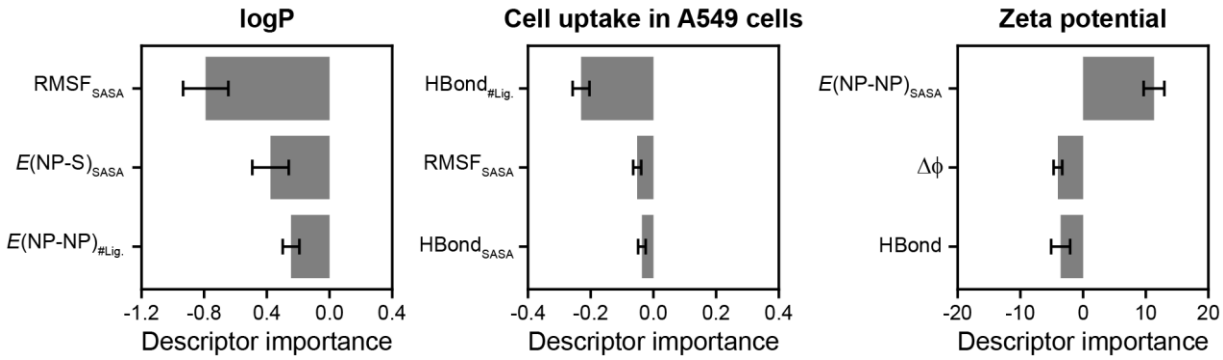

Figure 5. Descriptor importance for the QNAR models. Three most important descriptors for the log *P* (left), cell uptake (middle), and zeta potential (right) datasets identified for the random forest models (Figure 4). Descriptor abbreviations: RMSF is the ligand root-mean-squared fluctuation; *E*(NP–S) is the Lennard-Jones (LJ) interaction energy between GNP and solvent; *E*(NP–NP) is the LJ interaction energy between GNP ligands; HBond is the number of GNP-water hydrogen bonds; and Δϕ is the electrostatic potential difference between bulk water and the gold core. Descriptors with the subscript "SASA" are normalized by the solvent-accessible surface area of the GNP, and descriptors with the subscript "#Lig." are normalized by the total number of ligands on the gold core surface.

Analysis of important descriptors

We next sought to identify the descriptors that were most important to the QNAR model predictions of experimental trends. We quantified descriptor importance using the SHapley Additive exPlanation (SHAP) method, which was recently introduced as a model-agnostic method capable of quantifying descriptor importance even for "black box" models, such as deep neural networks.[56] The SHAP method assigns an importance value by comparing model predictions with and without a descriptor across all possible permutations of descriptor selections, then computing the Shapley value by averaging the marginal contribution of the descriptor.[57] The magnitude of the Shapley value estimates descriptor importance and the sign indicates if increasing the value of that descriptor increases (if positive) or decreases (if negative) the value of the model output.[56, 58] Additional details on these calculations are provided in the Methods. Given that the usefulness of descriptor importance analysis depends on model accuracy, we primarily focus on identifying

important descriptors using the RF models, which outperformed the LASSO models for all three experimental observables (Figure 4).

Figure 5 shows the three descriptors with the highest importance for the log *P*, cellular uptake, and zeta potential (right) datasets using the RF model. For each dataset, a single descriptor is identified as significantly more important than all other descriptors. These descriptors include the ligand root-mean-squared-fluctuations normalized by the SASA ($RMSF_{SASA}$) for the log *P* dataset, the number of GNP–water hydrogen bonds normalized by the number of ligands ($HBond_{\#Lig.}$) for the cell uptake dataset, and the GNP–GNP Lennard-Jones interaction energy normalized by the SASA ($E(NP–NP)_{SASA}$) for the zeta potential dataset. The same descriptors were also identified as most important upon performing a second trial of the entire computational workflow (see Methods and Figure S9). Moreover, the descriptors with the highest importance for the linear LASSO model are similar with those identified for the nonlinear RF model, subject to variations in normalization. Specifically, the most important descriptors for the LASSO model include the unnormalized ligand root-mean-squared-fluctuations (RMSF) for the log *P* dataset, the number of GNP–water hydrogen bonds normalized by the SASA ($HBond_{SASA}$) for the cell uptake dataset, and the GNP–GNP Lennard-Jones interaction energy normalized by the number of ligands ($E(NP–NP)_{\#Lig.}$) for the zeta potential dataset (Figure S6). The similarity of the most important descriptors for both the linear and nonlinear models in multiple simulation trials suggests the robustness of their importance.

An advantage of simulation-derived descriptors is that their physical significance can be readily interpreted to suggest GNP design guidelines or mechanistic hypotheses regarding their importance. We thus focus on understanding the most important descriptor for each dataset. For the log *P* dataset, larger values of log *P* indicate increased GNP partitioning to the hydrophobic

octanol phase. Because all simulations were performed in water, the negative importance of $RMSF_{SASA}$ indicates that larger ligand fluctuations are a signature of a more hydrophilic GNP, in agreement with our prior work that identified that decreasing ligand fluctuations correspond to increased surface hydrophobicity.[35] Ligand fluctuations are enhanced if water is a good solvent because a large number of energetically favorable conformations exist, whereas ligands will adopt a smaller number of conformations that minimize solvent contact if water is a poor solvent. For the cell uptake dataset, increasing the extent of GNP–water hydrogen bonding implies a decrease in cellular uptake. We can interpret a large value of this descriptor as reflecting the strong solvation of the GNP by water, suggesting that adsorption to the lipid bilayer prior to internalization is unfavorable. Recent work has similarly shown that increasing ligand hydrophobicity — thereby decreasing favorable interactions with water — promotes both adsorption to lipid bilayers and cellular internalization,[9, 15] agreeing with this interpretation. For the zeta potential dataset, it is expected that a descriptor of the GNP electrostatic potential ($\Delta\phi$, as defined in the Supporting Information) should be important since the zeta potential is typically used to estimate the effective charge of a GNP.[50, 59] Surprisingly, the most important descriptor for the zeta potential dataset quantifies Lennard-Jones interactions between ligands; increasing $E(\text{NP–NP})_{SASA}$ indicates that such interactions are less favorable. Weaker ligand–ligand interactions indicate less densely packed monolayers, either due to lower ligand surface densities or reduced ligand clustering, which could influence the zeta potential due to factors such as counterion condensation or changes to the titration state of charged ligands.[59] Recent simulations have similarly shown that molecular-scale features beyond charge density impact the zeta potential.[59] However, we caution that the importance of this descriptor may be biased by the imbalanced dataset which predominantly includes GNPs with negative zeta potentials in the dataset (Figure 1c).

Model generalizability to unseen datasets

We next tested the ability of the QNAR models to generalize to unseen datasets outside of the database used for model training. We focused on testing the generalizability of the cell uptake model because cell uptake is the most biologically relevant experimental measurement and predicting cell uptake is valuable for drug delivery applications. For this test, we utilized experimental measurements of cell uptake collected by Jiang *et al.* for 12 GNPs with core diameters of 2, 4, and 6 nm that were protected by ligands with four distinct end groups (shown in Figure 6a).[17] Trimethylammonium ($TTMA^+$) and carboxylate (COO^-) ligands are representative cationic and anionic ligands, respectively, while the $NS^{\pm}$ and $SN^{\pm}$ ligands are zwitterionic. Cell uptake measurements were performed after the GNPs were incubated with HeLa cells for 3 hours, and uptake was quantified using ICP-MS.[17] Figure 6b shows measurements of cell uptake for these 12 GNPs in human cervical carcinoma (HeLa) cells; all experimental data were taken from Ref. 17 and converted to the same units used for the cell uptake models in Figure 1c (see Table S3). In these units, GNPs with $TTMA^+$ ligands were found to have increased cell uptake compared to the other GNPs and increasing the GNP size from 2 nm to 6 nm generally increased cell uptake for all GNPs, although we note that the actual number of GNPs per cell decreases with increasing size for GNPs with COO^-, $NS^{\pm}$ and $SN^{\pm}$ ligands.[17] These trends emerge from changes in the endocytic pathways as a function of GNP size and surface properties.[17]

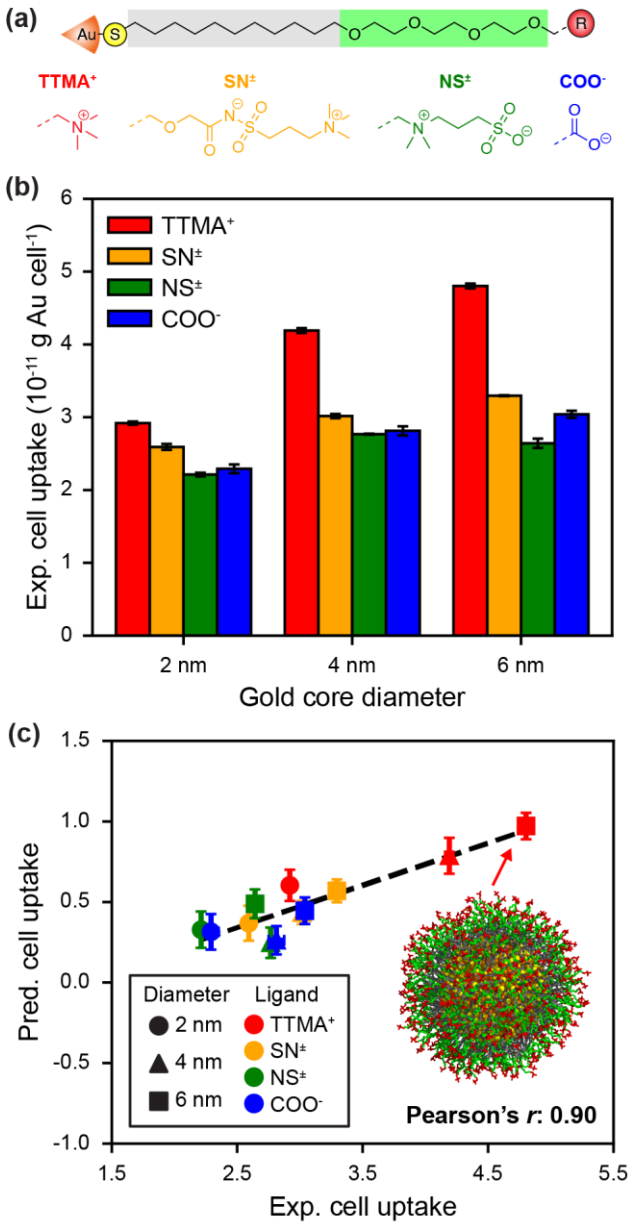

Figure 6. Generalizability of the cell uptake model to an unseen dataset. (a) Ligand structure with four end group chemistries. **(b)** Cell uptake of 12 GNPs with core diameters of 2, 4, and 6 nm and ligand structures from (a). Cell uptake values and errors were taken from experimental measurements in Ref. 17 and converted to the same units as Figure 1c (see Table S3). **(c)** Predicted *versus* experimental cell uptake values using a RF algorithm trained with 15 uncorrelated descriptors as input and 65 cell uptake labels from Figure 1c as output. Predicted values and error bars were estimated using a bagging approach with 20 RF models; the average of the predictions is reported, and the error is estimated by the standard deviation of the predictions. Pearson's *r* between predicted and experimental values using all 12 GNPs is shown in the lower right. The black dashed line shows the best fit line as a guide. The simulation snapshot illustrates a 6 nm GNP with $TTMA^+$ ligands.

Given that the RF model accurately related MD-derived descriptors to cell uptake in A549 cells (Figure 4), we trained a RF model on these 65 cell uptake measurements (Figure 1c). We then performed MD simulations of the 12 GNPs from Ref. 17 using the workflow in Figure 2, calculated the same uncorrelated descriptors required as input for the trained QNAR model, and utilized the QNAR model to predict cell uptake. We note that the experimental cell uptake measurements used for model training and testing utilized different cell lines (A549 cells in Ref. 32 *versus* HeLa cells in Ref. 17) and initial GNP concentrations (50 μg/mL in Ref. 32 *versus* ~1.2 μg/mL in Ref. 17). Hence, we do not expect the models trained with the data from Figure 1c to quantitatively predict cell uptake values from Figure 6b; rather, the trained RF model should capture qualitative cell uptake trends. Figure 6c shows the predicted *versus* experimental cell uptake for the 12 GNPs using the trained RF model. Pearson's *r* between predicted and experimental values is high (r = 0.90) and comparable to the value obtained for the original data set (Figure 4). The RF model also correctly predicts that the 6 nm GNP with $TTMA^+$ ligands (simulation snapshot in Figure 6c) has the highest cell uptake. The magnitude of the predicted cell uptake is lower than the experimental values, but this quantitative disagreement likely reflects differences in experimental conditions as noted above. We further tested the generalizability of the LASSO model to these 12 GNPs (Figure S7), which performed poorly with r = -0.06. The improved performance of the nonlinear RF model compared to the linear LASSO model is consistent with Figure 4. Together, these results indicate that the RF model generalizes well to a different cell uptake data set, suggesting that it may be a general tool for GNP design. Notably, the RF model can capture experimental trends despite differences in cell line and uptake mechanisms,[17] suggesting that GNP features related to cell uptake (*e.g.*, hydrogen bonding) are generally important for internalization into cells.

CONCLUSIONS

In this work, we modeled 154 SAM-protected GNPs in aqueous solution using atomistic MD simulations, computed a set of 15 uncorrelated simulation-derived descriptors, and parameterized QNAR models to relate these descriptors to experimentally measured log *P*, cell uptake, and zeta potential values. The MD simulation approach enables the calculation of descriptors that capture structural and chemical features of GNPs that emerge from the interplay of ligand–ligand, ligand–core, and ligand–water interactions in aqueous solution, thereby quantifying non-obvious, cooperative behaviors that may not be readily predicted based on single-molecule descriptors or descriptors based on static models of GNPs.[32, 34] We found that QNAR models trained with both LASSO and RF regression algorithms accurately related simulation-derived descriptors to experimental measurements. The small set of descriptors further permitted analysis of descriptor importance to reveal that ligand fluctuations and hydrogen bonding are key indicators of octanol partitioning and cellular uptake, suggesting general GNP design guidelines. We further showed that the RF model could predict the uptake of 12 additional GNPs in a different cell line, highlighting its generalizability.

Together, these results demonstrate that the combination of MD simulations and data-centric regression analysis can predict the impact of GNP composition on corresponding physicochemical properties and biological behavior. Similar QNAR models were also obtained from shorter simulations (Figure S8), even if descriptors were not fully converged, suggesting that the computational protocol could be utilized for high-throughput GNP screening. We also highlight that the descriptors employed in this work do not depend on GNP-specific information (*e.g.*, ligand chemical structure or core composition), but rather reflect structural and chemical properties that are materials agonistic (*e.g.*, hydrogen bonding) and thus could more broadly

generalize to other classes of nanomaterials. Another benefit of modeling GNPs with MD is that their interactions with biomolecules could be further interrogated by modeling GNPs in the presence of a lipid bilayer[9, 11, 16, 44, 47] or proteins,[45] enabling the calculation of additional environment-specific descriptors or complementary mechanistic investigations. Future work will also explore the ability of deep learning models to automatically extract features from the extensive data available within MD trajectories to improve prediction accuracy without predefining descriptors.[60-61]

METHODS

Nanoparticle simulation workflow

SAM-coated GNPs were constructed using a self-assembly approach described previously[36] and further detailed in the Supporting Information. Each gold core was modeled by cutting a spherical region from the bulk gold face-centered cubic lattice; faceting was ignored for simplicity and because prior studies have suggested that such facets do not substantially influence SAM structure.[36] Either butanethiol (SMILES: SCCCC) or 1,2-dithiolane (SMILES: C1CCSS1) ligand substructures were self-assembled onto the spherical gold core as shown in Figure 2. Adsorbed substructures were then replaced by desired ligands. Adsorbed ligands were randomly removed to match the number of ligands listed in the database; if the listed number of ligands was larger than the number of adsorbed ligands, then all adsorbed ligands were used. This approach led to reasonable ligand surface densities (see Figure S1). Atomic clashes were eliminated using a short 4 ps *NVT* simulation with harmonic restraints applied to extend the last heavy atom of each atom away from the gold surface with all interactions turned off. Van der Waals interactions were then slowly reintroduced in a series of energy minimization steps as illustrated in Figure S2. The system was then solvated with water, and, for charged systems, water molecules were replaced

with sodium or chlorine counterions to ensure charge neutrality. All subsequent simulations were performed at 300 K and 1 bar. A 1 ns *NPT* equilibration simulation was performed with the temperature controlled by the velocity-rescale thermostat and the pressure controlled by the Berendsen barostat, then a 50 ns *NPT* production simulation was performed with the same thermostat and the pressure controlled by the Parrinello-Rahman thermostat. The last 40 ns of simulation data were used to compute descriptors. This time was sufficient time for descriptors to converge (Figures S4 and S5).

Simulation parameters

In all simulations, Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential with a 1.2 nm cutoff that was smoothly shifted to zero between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the smooth particle mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and fourth-order interpolation. Bonds were constrained using the LINCS algorithm. Periodic boundary conditions were enabled in all directions. A rhombic dodecahedron simulation box geometry was used to maximize computational efficiency given the approximately spherical symmetry of the GNP systems. Ligand atoms were modeled with the CGenFF/CHARMM36 forcefield (July 2020 version),[62-64] gold atoms were modeled with the Interface force field,[65] and water molecules were modeled using the TIP3P model. All MD simulations were performed using Gromacs 2016[66] using the leapfrog integrator with a 2-fs timestep.

Computing MD-derived descriptors

All MD-derived descriptors were generated with a combination of in-house Python (MDTraj[67] and MDAnalysis[68]) and Gromacs analysis tools.[66] The full list of the 25 MD-derived

descriptors computed are tabulated in Table S1. The 15 uncorrelated descriptors shown in Table S2 were used to train the QNAR models.

QNAR model training

LASSO and RF algorithms were used to train QNAR models relating MD-derived descriptors to experimental labels. Each experimental dataset was randomly divided into training (80% of the original dataset) and test sets (20% of the original dataset). The training sets were used to train model parameters and the test sets were used to evaluate model accuracy. Five-fold cross validation was performed to evaluate the generalizability of the models to unseen data within the training set. In this approach, 80% of the training set was used to train the model and the remaining 20% (not used in model training) was used to validate model predictions. Validation set predictions are reported in Figure 4. This approach was repeated five times such that each GNP was included in the validation set once. When predicting the test set, the models were trained using all the training data. All descriptor values were standardized by subtracting the mean and dividing by the standard deviation so that they could be compared on the same magnitude. A single hyperparameter for each LASSO model was tuned using 5-fold cross validation as described in the Supporting Information (Figure S11). 500 trees were arbitrarily selected as a hyperparameter for the RF models. To estimate prediction errors, a bagging approach was implemented when training each QNAR model. In this approach, 20 LASSO or RF algorithms were trained by randomly sampling the training data with replacement following the procedure describe above. The average prediction of the twenty models is reported in the Results above, and the prediction error is estimated by the standard deviation of the predictions.

To test model robustness, the entire simulation workflow (starting from the self-assembly simulations) was repeated for a second trial. The second trial was performed with a shorter

simulation time of 20 ns per GNP with the last 10 ns used for MD-derived descriptor calculations. QNAR models trained using data from the second trial performed similarly to QNAR models trained using data from the longer 50 ns GNP–water simulations (reported in Figure 4), confirming robustness in the computational workflow and suggesting that a shorter simulation time could be used to obtain accurate predictions even if the MD-derived descriptors are not fully converged (Figure S8-S10).

Descriptor importance analysis

Descriptor importance was determined using the SHAP method.[57] For each descriptor, the average magnitude of the Shapley values across all instances is reported and the sign of the descriptor importance is determined by the sign of the Pearson's *r* value between the Shapley and descriptor values. To estimate the accuracy of the resulting importance values, we implemented a bootstrapping procedure by re-training the LASSO and RF models with 90% of the training set (randomly sampled without replacement) and computing corresponding importance values.[69] This procedure was iterated ten times to obtain the average and standard deviation of importance values across these trials.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2044997. This work used the computing resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science

Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

SUPPORTING INFORMATION

Detailed description of the computational workflow; description, convergence, and tabulated values of descriptors; table of uncorrelated descriptors; LASSO model predictions for 12 GNPs not within the training data; parity plots and descriptor importance for a shorter second trial of the computational workflow; values of all computed descriptors and experimental labels, along with literature references.

REFERENCES

1. Ghosh, P.; Han, G.; De, M.; Kim, C. K.; Rotello, V. M., Gold nanoparticles in delivery applications. *Advanced Drug Delivery Reviews* **2008,** *60* (11), 1307-1315.

2. Jans, H.; Huo, Q., Gold nanoparticle-enabled biological and chemical detection and analysis. *Chem Soc Rev* **2012,** *41* (7), 2849-2866.

3. Riley, R. S.; Day, E. S., Gold nanoparticle-mediated photothermal therapy: applications and opportunities for multimodal cancer treatment. *Wires Nanomed Nanobi* **2017,** *9* (4).

4. Longmire, M.; Choyke, P. L.; Kobayashi, H., Clearance properties of nano-sized particles and molecules as imaging agents: considerations and caveats. *Nanomedicine-Uk* **2008,** *3* (5), 703-717.

5. You, C. C.; De, M.; Rotello, V. M., Monolayer-protected nanoparticle-protein interactions. *Curr Opin Chem Biol* **2005,** *9* (6), 639-646.

6. Moyano, D. F.; Goldsmith, M.; Solfiell, D. J.; Landesman-Milo, D.; Miranda, O. R.; Peer, D.; Rotello, V. M., Nanoparticle Hydrophobicity Dictates Immune Response. *J. Am. Chem. Soc.* **2012,** *134* (9), 3965-3967.

7. Sun, S. S.; Huang, Y. Y.; Zhou, C.; Chen, S. S.; Yu, M. X.; Liu, J. B.; Zheng, J., Effect of Hydrophobicity on Nano-Bio Interactions of Zwitterionic Luminescent Gold Nanoparticles at the Cellular Level. *Bioconjugate Chem* **2018,** *29* (6), 1841-1846.

8. Chen, K. M.; Rana, S.; Moyano, D. F.; Xu, Y. S.; Guo, X. H.; Rotello, V. M., Optimizing the selective recognition of protein isoforms through tuning of nanoparticle hydrophobicity. *Nanoscale* **2014,** *6* (12), 6492-6495.

9. Lochbaum, C. A.; Chew, A. K.; Zhang, X. Z.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A., Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers. *Acs Nano* **2021,** *15* (4), 6562-6572.

10. Chong, G.; Foreman-Ortiz, I. U.; Wu, M.; Bautista, A.; Murphy, C. J.; Pedersen, J. A.; Hernandez, R., Defects in Self-Assembled Monolayers on Nanoparticles Prompt Phospholipid Extraction and Bilayer-Curvature-Dependent Deformations. *J Phys Chem C* **2019,** *123* (45), 27951-27958.

11. Lolicato, F.; Joly, L.; Martinez-Seara, H.; Fragneto, G.; Scoppola, E.; Bombelli, F. B.; Vattulainen, I.; Akola, J.; Maccarini, M., The Role of Temperature and Lipid Charge on Intake/Uptake of Cationic Gold Nanoparticles into Lipid Bilayers. *Small* **2019,** *15* (23).

12. Tatur, S.; Maccarini, M.; Barker, R.; Nelson, A.; Fragneto, G., Effect of Functionalized Gold Nanoparticles on Floating Lipid Bilayers. *Langmuir* **2013,** *29* (22), 6606-6614.

13. Heikkila, E.; Martinez-Seara, H.; Gurtovenko, A. A.; Javanainen, M.; Hakkinen, H.; Vattulainen, I.; Akola, J., Cationic Au Nanoparticle Binding with Plasma Membrane-like Lipid

Bilayers: Potential Mechanism for Spontaneous Permeation to Cells Revealed by Atomistic Simulations. *J. Phys. Chem. C* **2014,** *118* (20), 11131-11141.

14. Lin, J. Q.; Zhang, H. W.; Chen, Z.; Zheng, Y. G., Penetration of Lipid Membranes by Gold Nanoparticles: Insights into Cellular Uptake, Cytotoxicity, and Their Relationship. *Acs Nano* **2010,** *4* (9), 5421-5429.

15. Moyano, D. F.; Saha, K.; Prakash, G.; Yan, B.; Kong, H.; Yazdani, M.; Rotello, V. M., Fabrication of Corona-Free Nanoparticles with Tunable Hydrophobicity. *ACS Nano* **2014,** *8* (7), 6748-6755.

16. Contini, C.; Schneemilch, M.; Gaisford, S.; Quirke, N., Nanoparticle-membrane interactions. *J Exp Nanosci* **2018,** *13* (1), 62-81.

17. Jiang, Y.; Huo, S. D.; Mizuhara, T.; Das, R.; Lee, Y. W.; Hou, S.; Moyano, D. F.; Duncan, B.; Liang, X. J.; Rotello, V. M., The Interplay of Size and Surface Functionality on the Cellular Uptake of Sub-10 nm Gold Nanoparticles. *Acs Nano* **2015,** *9* (10), 9986-9993.

18. Li, X. N.; Robinson, S. M.; Gupta, A.; Saha, K.; Jiang, Z. W.; Moyano, D. F.; Sahar, A.; Riley, M. A.; Rotello, V. M., Functional Gold Nanoparticles as Potent Antimicrobial Agents against Multi-Drug-Resistant Bacteria. *Acs Nano* **2014,** *8* (10), 10682-10686.

19. Melby, E. S.; Lohse, S. E.; Park, J. E.; Vartanian, A. M.; Putans, R. A.; Abbott, H. B.; Hamers, R. J.; Murphy, C. J.; Pedersen, J. A., Cascading Effects of Nanoparticle Coatings: Surface Functionalization Dictates the Assemblage of Complexed Proteins and Subsequent Interaction with Model Cell Membranes. *ACS Nano* **2017,** *11* (6), 5489-5499.

20. Rossi, G.; Monticelli, L., Gold nanoparticles in model biological membranes: A computational perspective. *Bba-Biomembranes* **2016,** *1858* (10), 2380-2389.

21. Rossi, G.; Monticelli, L., Simulating the interaction of lipid membranes with polymer and ligand-coated nanoparticles. *Adv Phys-X* **2016,** *1* (2), 276-296.

22. Saha, K.; Rahimi, M.; Yazdani, M.; Kim, S. T.; Moyano, D. F.; Hou, S.; Das, R.; Mout, R.; Rezaee, F.; Mahmoudi, M.; Rotello, V. M., Regulation of Macrophage Recognition through the Interplay of Nanoparticle Surface Functionality and Protein Corona. *ACS Nano* **2016,** *10* (4), 4421-4430.

23. Yu, Q. H.; Zhao, L. X.; Guo, C. C.; Yan, B.; Su, G. X., Regulating Protein Corona Formation and Dynamic Protein Exchange by Controlling Nanoparticle Hydrophobicity. *Front Bioeng Biotech* **2020,** *8*.

24. Kim, S. T.; Saha, K.; Kim, C.; Rotello, V. M., The Role of Surface Functionality in Determining Nanoparticle Cytotoxicity. *Accounts Chem Res* **2013,** *46* (3), 681-691.

25. Bunker, A.; Magarkar, A.; Viitala, T., Rational design of liposomal drug delivery systems, a review: Combined experimental and computational studies of lipid membranes, liposomes and their PEGylation. *Bba-Biomembranes* **2016,** *1858* (10), 2334-2352.

26. Pengo, P.; Sologan, M.; Pasquato, L.; Guida, F.; Pacor, S.; Tossi, A.; Stellacci, F.; Marson, D.; Boccardo, S.; Pricl, S.; Posocco, P., Gold nanoparticles with patterned surface monolayers for nanomedicine: current perspectives. *Eur Biophys J Biophy* **2017,** *46* (8), 749-771.

27. Fourches, D.; Pu, D. Q. Y.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A., Quantitative Nanostructure-Activity Relationship Modeling. *ACS Nano* **2010,** *4* (10), 5703-5712.

28. Fourches, D.; Pu, D. Q. Y.; Tropsha, A., Exploring Quantitative Nanostructure-Activity Relationships (QNAR) Modeling as a Tool for Predicting Biological Effects of Manufactured Nanoparticles. *Combinatorial Chemistry & High Throughput Screening* **2011,** *14* (3), 217-225.

29. Singh, K. P.; Gupta, S., Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *Rsc Adv* **2014,** *4* (26), 13215-13230.

30. Wang, W. Y.; Sedykh, A.; Sun, H. N.; Zhao, L. L.; Russo, D. P.; Zhou, H. Y.; Yan, B.; Zhu, H., Predicting Nano-Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* **2017,** *11* (12), 12641-12649.

31. Wang, W. Y.; Yan, X. L.; Zhao, L. L.; Russo, D. P.; Wang, S. Q.; Liu, Y.; Sedykh, A.; Zhao, X. L.; Yan, B.; Zhu, H., Universal nanohydrophobicity predictions using virtual nanoparticle library. *J Cheminformatics* **2019,** *11*.

32. Yan, X. L.; Sedykh, A.; Wang, W. Y.; Yan, B.; Zhu, H., Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat Commun* **2020,** *11* (1).

33. Yan, X. L.; Sedykh, A.; Wang, W. Y.; Zhao, X. L.; Yan, B.; Zhu, H., In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* **2019,** *11* (17), 8352-8362.

34. Yan, X. L.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B., Prediction of Nano-Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images. *Acs Sustain Chem Eng* **2020,** *8* (51), 19096-19104.

35. Chew, A. K.; Dallin, B. C.; Van Lehn, R. C., The Interplay of Ligand Properties and Core Size Dictates the Hydrophobicity of Monolayer-Protected Gold Nanoparticles. *Acs Nano* **2021,** *15* (3), 4534-4545.

36. Chew, A. K.; Van Lehn, R. C., Effect of Core Morphology on the Structural Asymmetry of Alkanethiol Monolayer-Protected Gold Nanoparticles. *J Phys Chem C* **2018,** *122* (45), 26288-26297.

37. Ghorai, P. K.; Glotzer, S. C., Molecular dynamics simulation study of self-assembled monolayers of alkanethiol surfactants on spherical gold nanoparticles. *J. Phys. Chem. C* **2007,** *111* (43), 15857-15862.

38. Kister, T.; Monego, D.; Mulvaney, P.; Widmer-Cooper, A.; Kraus, T., Colloidal Stability of Apolar Nanoparticles: The Role of Particle Size and Ligand Shell Structure. *ACS Nano* **2018,** *12* (6), 5969-5977.

39. Lane, J. M. D.; Grest, G. S., Spontaneous Asymmetry of Coated Spherical Nanoparticles in Solution and at Liquid-Vapor Interfaces. *Phys. Rev. Lett.* **2010,** *104* (23).

40. Luedtke, W. D.; Landman, U., Structure, dynamics, and thermodynamics of passivated gold nanocrystallites and their assemblies. *J. Phys. Chem.* **1996,** *100* (32), 13323-13329.

41. Koch, A. H. R.; Leveque, G.; Harms, S.; Jaskiewicz, K.; Bernhardt, M.; Henkel, A.; Sonnichsen, C.; Landfester, K.; Fytas, G., Surface Asymmetry of Coated Spherical Nanoparticles. *Nano Lett.* **2014,** *14* (7), 4138-4144.

42. Lane, J. M. D.; Grest, G. S., Assembly of responsive-shape coated nanoparticles at water surfaces. *Nanoscale* **2014,** *6* (10), 5132-5137.

43. Nash, J. A.; Tucker, T. L.; Therriault, W.; Yingling, Y. G., Binding of single stranded nucleic acids to cationic ligand functionalized gold nanoparticles. *Biointerphases* **2016,** *11* (4).

44. Nakamura, H.; Sezawa, K.; Hata, M.; Ohsaki, S.; Watano, S., Direct translocation of nanoparticles across a model cell membrane by nanoparticle-induced local enhancement of membrane potential. *Phys Chem Chem Phys* **2019,** *21* (35), 18830-18838.

45. Simonelli, F.; Rossi, G.; Monticelli, L., Role of Ligand Conformation on Nanoparticle-Protein Interactions. *J Phys Chem B* **2019,** *123* (8), 1764-1769.

46. Van Lehn, R. C.; Atukorale, P. U.; Carney, R. P.; Yang, Y. S.; Stellacci, F.; Irvine, D. J.; Alexander-Katz, A., Effect of Particle Diameter and Surface Composition on the Spontaneous Fusion of Monolayer-Protected Gold Nanoparticles with Lipid Bilayers. *Nano Lett.* **2013,** *13* (9), 4060-4067.

47. Van Lehn, R. C.; Ricci, M.; Silva, P. H. J.; Andreozzi, P.; Reguera, J.; Voitchovsky, K.; Stellacci, F.; Alexander-Katz, A., Lipid tail protrusions mediate the insertion of nanoparticles into model cell membranes. *Nat Commun* **2014,** *5*.

48. Sridhar, D. B.; Gupta, R.; Rai, B., Effect of surface coverage and chemistry on self-assembly of monolayer protected gold nanoparticles: a molecular dynamics simulation study. *Phys Chem Chem Phys* **2018,** *20* (40), 25883-25891.

49. Olenick, L. L.; Troiano, J. M.; Vartanian, A.; Melby, E. S.; Mensch, A. C.; Zhang, L.; Hong, J.; Mesele, O.; Qiu, T.; Bozich, J., Lipid corona formation from nanoparticle interactions with bilayers. *Chem* **2018,** *4* (11), 2709-2723.

50. Predota, M.; Machesky, M. L.; Wesolowski, D. J., Molecular Origins of the Zeta Potential. *Langmuir* **2016,** *32* (40), 10189-10198.

51. Patil, S.; Sandberg, A.; Heckert, E.; Self, W.; Seal, S., Protein adsorption and cellular uptake of cerium oxide nanoparticles as a function of zeta potential. *Biomaterials* **2007,** *28* (31), 4600-4607.

52. Schwegmann, H.; Feitz, A. J.; Frimmel, F. H., Influence of the zeta potential on the sorption and toxicity of iron oxide nanoparticles on S. cerevisiae and E. coli. *J. Colloid Interface Sci.* **2010,** *347* (1), 43-48.

53. Djebaili, T.; Richardi, J.; Abel, S.; Marchi, M., Atomistic Simulations of the Surface Coverage of Large Gold Nanocrystals. *J. Phys. Chem. C* **2013,** *117* (34), 17791-17800.

54. Heikkila, E.; Gurtovenko, A. A.; Martinez-Seara, H.; Hakkinen, H.; Vattulainen, I.; Akola, J., Atomistic Simulations of Functional Au-144(SR)(60) Gold Nanoparticles in Aqueous Environment. *J Phys Chem C* **2012,** *116* (17), 9805-9815.

55. Nel, A. E.; Madler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M., Understanding biophysicochemical interactions at the nano-bio interface. *Nat Mater* **2009,** *8* (7), 543-557.

56. Rodriguez-Perez, R.; Bajorath, J., Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J Comput Aid Mol Des* **2020,** *34* (10), 1013-1026.

57. Lundberg, S. M.; Lee, S. I., A Unified Approach to Interpreting Model Predictions. *Adv Neur In* **2017,** *30*.

58. Boehmke, B.; Greenwell, B. M., *Hands-on machine learning with R*. CRC Press: Boca Raton, 2019; p volumes cm.

59. Liang, D. Y.; Dahal, U.; Zhang, Y. Q.; Lochbaum, C.; Ray, D.; Hamers, R. J.; Pedersen, J. A.; Cui, Q., Interfacial water and ion distribution determine zeta potential and binding affinity of nanoparticles to biomolecules. *Nanoscale* **2020,** *12* (35), 18106-18123.

60. Chew, A. K.; Jiang, S. L.; Zhang, W. Q.; Zavala, V. M.; Van Lehn, R. C., Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks. *Chem Sci* **2020,** *11* (46), 12464-12476.

61. Kelkar, A. S.; Dallin, B. C.; Van Lehn, R. C., Predicting Hydrophobicity by Learning Spatiotemporal Features of Interfacial Water Structure: Combining Molecular Dynamics Simulations with Convolutional Neural Networks. *J Phys Chem B* **2020,** *124* (41), 9103-9114.

62. Huang, J.; MacKerell, A. D., CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* **2013,** *34* (25), 2135-2145.

63. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J Comput Chem* **2010,** *31* (4), 671-690.

64. Yu, W. B.; He, X. B.; Vanommeslaeghe, K.; MacKerell, A. D., Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J Comput Chem* **2012,** *33* (31), 2451-2468.

65. Heinz, H.; Vaia, R. A.; Farmer, B. L.; Naik, R. R., Accurate Simulation of Surfaces and Interfaces of Face-Centered Cubic Metals Using 12-6 and 9-6 Lennard-Jones Potentials. *J. Phys. Chem. C* **2008,** *112* (44), 17281-17290.

66. Pall, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E., Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. *Lect Notes Comput Sc* **2015,** *8759*, 3-27.

67. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S., MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015,** *109* (8), 1528-1532.

68. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O., Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J Comput Chem* **2011,** *32* (10), 2319-2327.

69. An, Y. L.; Sherman, W.; Dixon, S. L., Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. *J Chem Inf Model* **2013,** *53* (9), 2312-2321.

TABLE OF CONTENTS IMAGE