JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

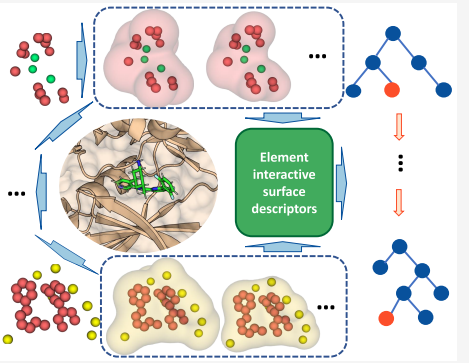
EISA-Score: Element Interactive Surface Area Score for Protein−Ligand Binding Affinity Prediction

Md Masud Rana and Duc Duy Nguyen*

Cite This: J. Chem. Inf. Model. 2022, 62, 4329−4341 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular surface representations have been advertised as a great tool to study protein structure and functions, including protein−ligand binding affinity modeling. However, the conventional surface-area-based methods fail to deliver a competitive performance on the energy scoring tasks. The main reason is the lack of crucial physical and chemical interactions encoded in the molecular surface generations. We present novel molecular surface representations embedded in different scales of the element interactive manifolds featuring the dramatically dimensional reduction and accurately physical and biological properties encoders. Those low-dimensional surface-based descriptors are ready to be paired with any advanced machine learning algorithms to explore the essential structure−activity relationships that give rise to the element interactive surface area-based scoring functions (EISA-score). The newly developed EISA-score has outperformed many state-of-the-art models, including various well-established surface-related representations, in standard PDBbind benchmarks.



1. INTRODUCTION

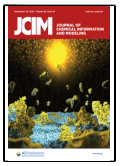
Geometric modeling of biomolecules concerns geometrical components at various scales and dimensions, including molecular surface generation, molecular visualization, curvature analysis, surface annotation, etc.[21,22,35,36,39,41,45,61,63,70,73] Among these ingredients, the molecular surface plays a significant role in visualizing and analyzing molecular structures and properties. Specifically, one can project the electrostatic potentials, flexibility indexes, and curvature magnitudes on the protein surface[60] to reveal protein structure and function, such as protein−ligand binding sites, protein−protein binding hot spots, and protein−DNA interactions.

There are various methods proposed to compute the biomolecular surfaces. One can classify these methods into three categories: analytical representation, partial differential equations (PDEs)-based generation, and explicit formulation. For the analytical calculation, the simplest model can be referred to a van der Waals surface (vdWS), formed by a union of the atomic sphere of the van der Waals radius. In addition, one can use the trajectory of the probe's center moving around the van der Waals surface to give rise to the solvent accessible surface (SAS).[41] Unfortunately, those vdWS and SAS approaches suffer the nonsmooth regions causing computation obstacles. For that reason, Connolly proposed a solvent-excluded surface (SES) to avoid these nonsmooth issues.[21] MSMS software was later developed to improve the speed and reliability of the SES calculation via the reduced surface.[63] There are other efficient algorithms for generating SES.[10,15,24,28,29,31,37,38,40,47−49,62,82] Among them, TMSmesh[49] used the boundary element method and finite element method to handle arbitrary sizes of molecules. By adapting a multistep region-growing EDT approach, Daberdaku and Ferrari[24] developed fast molecular surface representations for large molecules. Hermosilla et al.[38] utilized interactive GPU power to accelerate SES rendering at a fractional cost. Wei and his coworkers introduced ESES which accurately generates SES on the Cartesian mesh.[48]

To define the solute−solvent region, one can allow the overlap of the solvent and solute domains via the fuzzy characteristic or hypersurface functions. This approach was initially introduced in 2005 to generate the class of desirable biomolecular surfaces by curvature-driven geometric PDEs.[77] The other type used the mean curvature flow or Laplace-Beltrami equation to form the molecular surface by minimizing the surface energy.[3−5] Later, these optimal geometric flow schemes were extended to model the nonpolar energy of the biomolecular systems.[16−18,76,78]

Despite the fact that the analytical approaches are able to generate the accurate molecular surface, and PDE-based methods can embed the molecular energy information, they are not flexible when expressing the surface of local atoms. In addition, the analytical surfaces often consist of geometric

Received: June 1, 2022
Published: September 15, 2022



https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

singularities which obstruct the estimation of other geometry information such as curvature.[21,63] The prominent representative of the explicit surface is the Gaussian surfaces, in which the Gaussian functions are used as the density potential at each atom.[34,53,55,83] Those surfaces avoid the geometric singularities but are sensitive to level set values used to extract a specific surface candidate.[53]

Molecular surface representations have shown their important role in predictions of solvation-free energies and ion channel transport. However, they have to be incorporated in the realm of the physical models such as the Poisson−Boltzmann equation[19,71] and Poisson−Nernst−Planck model.[12−14,76,78] These dependencies limit the direct link of the molecular surface properties on the molecular properties. In addition, the dependence on the parametrized factors such as atomic charges and grid sizes of the predefined domain has restrained the capability of the molecular surface details on the diverse and complex biomolecular structures.[54]

Due to the essential physical and chemical properties captured on the molecular surface, its information has been widely used in quantitative and qualitative tasks in exploring molecule properties and activities. For the qualitative purpose, the biomolecular surface can be used to visualize protein folding,[66] protein−protein interactions,[23] DNA binding and bending,[27] molecular docking,[65] binding site classification,[25] and molecular dynamics.[33] In the quantitative effort, the molecular surface can be integrated with the implicit solvent model to predict solvation free energy,[1,11,55] incorporated with the Poisson−Nernst−Planck setting to compute the electrostatic and concentration profiles and current−voltage curves,[84,85] and used as a variable in the partial least-squares model to predict the solubility and permeability of the druglike molecules.[6] However, those approaches are limited in representing complex biomolecular structures from large and diverse data sets due to the lack of details of physical and chemical interactions.

Recently, we have unlocked the representation power of the curvatures of the molecular surface for massive and distinct molecular and biomolecular structures to predict drug toxicity, molecular solvation energy, and protein−ligand binding affinity.[57] However, the role of the surface area in capturing the crucial physical and chemical interactions in the biomolecular structures is not fully explored. Despite the recent efforts to integrate the surface area information into predictive models such as Cyscore[9] and GLXE[26] for protein−ligand binding affinity prediction, those surface area-based models are far from the competitive level with their counterparts.

To decipher the full potential of the surface area-based descriptors, we propose to construct the molecular surface at the pairwise element levels. The element-wise surfaces will effectively capture some specific types of noncovalent interactions, such as van der Waals interactions, hydrophobicity, and hydrogen bonds. Furthermore, the element-level surface area features highlight the scalability in the sense that the proposed representation will be independent of molecular sizes, that is, number of atoms, thus enabling the equal footing configuration for molecular structures from the highly diversified data sets. Given the information on atomic coordinates, there are several ways to construct the corresponding molecular surface. In this work, we extend our proposed molecular surface generation of small molecules in the implicit solvation modeling[55] to characterize the surfaces between protein and ligand at the element level. In general, the Riemannian manifolds are constructed on the subsets of the group of element types to allow a convenient formation of the structures of differential geometry. One can extract the manifold representations for the selected atoms via a discrete-to-continuum mapping that enables the embedding of the high dimensional data space of the biomolecular atoms into the low-dimensional model.[79−81]

Protein−ligand interactions trigger lots of biological processes, including signal transduction, gene regulation, and immunoreaction. Thus, understanding protein−ligand interactions will decipher the mechanism of biological regulation, providing a solid framework and theoretical support for drug design. As a result, various scoring functions (SFs) have been developed to capture and represent the protein−ligand binding process in recent years. The most popular type of SF is the empirical SFs that uses the physical terms' associated coefficients fitted to the existing data.[30,69,86] The completely data-dependent SFs are the machine learning-based SFs, where their performances are strongly affected by the quality of the training data. On a positive note, such data-driven models can effectively handle large and diverse data sets. The descriptors of these SFs come in various forms, from the simple element-pair contact counts[2] to high-level abstract mathematical representations, namely graph theory,[56] differential geometry,[57] persistent homology,[7,50] hypergraphs,[51] spectral graphs,[52] and persistent curvatures.[74,75]

The objective of the present work is to introduce the element-interactive surface area (EISA) descriptors for the first time in the literature to accurately and effectively describe the molecular representations in the low-dimensional space. The interactive molecular surface is presented by the standard correlation functions, namely exponential and Lorentz kernel functions which give us the Gaussian-like surfaces. Moreover, those surfaces are infinitely differentiable and free of geometric singularities. In this work, we are interested in constructing a class of surfaces at the multiscale levels by varying the suitable kernel parameters and level set values via the multiscale discrete-to-continuum mapping. By pairing with the advanced machine learning architectures, the molecular surface-based model, named EISA-score, reveals its quantitative power in predicted drug-related molecular properties, such as protein−ligand binding affinity (BA). The accurate and robust method to calculate the BA values of the small molecules is the crucial component in speeding up the process of drug discovery to help design novel drugs. In this work, we test the scoring power of our proposed model against three common benchmarks in the drug design area, namely CASF-2007,[20] CASF-2013,[46] and CASF-2016.[68] Several experiments confirm that our EISA-score achieves state-of-the-art results and outperforms the other molecular surface-based models by a wide margin.

2. MODEL DEVELOPMENT

2.1. Element Interactive Manifolds. This section presents a background of the discrete-to-continuum mapping via the atomic density function formulated in the common choice of correlation kernel functions. Under the element-wise setting, that mapping extracts the low-dimensional manifolds targeting the specific element types to represent the high dimensional interactions for the group of atoms of interest.

2.1.1. Atomic Density. Given a molecule with N atoms, we denote $\mathcal{X} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\}$ as the set of N atomic coordinates.

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

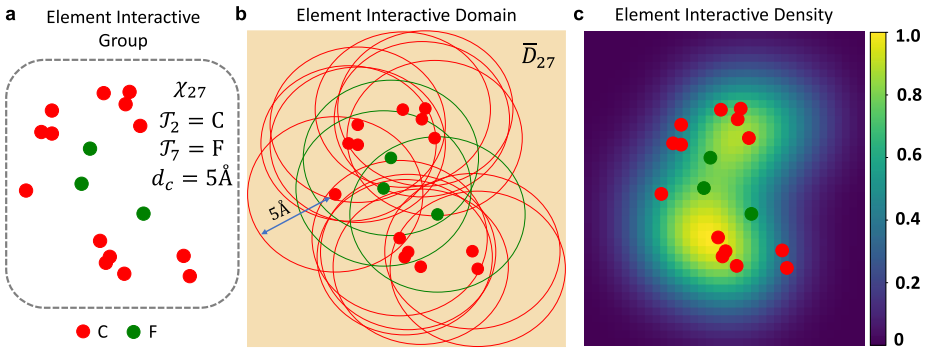


Figure 1. A 2D illustration of an element interactive information. (a) Element interactive collection $\mathcal{X}_{27}$, where $\mathcal{T}_2 = C$ and $\mathcal{T}_7 = F$ with a cutoff distance $d_c = 5$ Å. (b) The element interactive region D_{27} is the union of the balls with their centers belonging to $\mathcal{X}_{27}$ and radii being 5 Å. The element interactive domain $\bar{D}_{27}$ is the colored region. (c) The element interactive density ρ_{27} with an exponential kernel.

Let $\mathbf{r}_j \in \mathbb{R}^3$ be the position of jth atom in the molecule and $\|\mathbf{r} - \mathbf{r}_j\|$ be the Euclidean distance between the atom $\mathbf{r}_j$ and a point $\mathbf{r} \in \mathbb{R}^3$. The molecular density is given by a discrete-to-continuum mapping

$$\rho(\mathbf{r}) = \sum_{j=1}^{N} \omega_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) \tag{1}$$

where ω_j are the weights, η_j are characteristic distances, and Φ is a C^2 correlation kernel or statistical density estimator that satisfies the following admissibility conditions

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) = 1, \quad \text{as} \|\mathbf{r} - \mathbf{r}_j\| \to 0 \tag{2}$$

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) = 0, \quad \text{as} \|\mathbf{r} - \mathbf{r}_j\| \to \infty \tag{3}$$

As in our previous work,[56−58] the generalized exponential and generalized Lorentz functions have shown their robustness and efficiency in capturing the dynamic interactions between various types of atoms at different ranges. Their formulations are given as the following

generalized exponential kernel

$$\Phi_E(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) = e^{-(\|\mathbf{r}-\mathbf{r}_j\|/\eta_j)^\kappa}, \quad \kappa > 0 \tag{4}$$

generalized Lorentz kernel

$$\Phi_L(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) = \frac{1}{1 + (\|\mathbf{r} - \mathbf{r}_j\|/\eta_j)^\kappa}, \quad \kappa > 0 \tag{5}$$

In the present work, the atomic weights w_j are chosen to be 1 for simplicity. In other applications, one might consider the atomic charges to represent the atomic weights.[7,57] The kernel parameters η_j and κ need to be carefully selected to capture the crucial interactions between different atom types and consequently produce a meaningful molecular surface. The multiscale atom density that can be obtained by choosing different ranges for the kernel parameter sets η_j, and κ has shown its potential in covering different wide range intramolecular interactions from the diverse families of proteins.[58,59]

2.1.2. Element Interactive Densities. To account for details of physical interactions in a protein−ligand complex such as hydrophobic, hydrophilic, etc., we are interested in constructing the atomic densities in an element interactive manner. To this end, we consider the four most appearances element types in protein, namely C, N, O, and S, while there are 10 commonly occurring element types in ligand, namely, H, C, N, O, S, P, F, Cl, Br, and I. As a result, we have 40 element interactive possibilities between protein and ligand atoms: CH, CC, CN, CO, ..., NH, ..., and SI. It is worth noting that we do not consider protein's H in our current approach due to the missing hydrogen atoms in most of the protein's crystal structures. One might be concerned that the lack of hydrogen consideration in the protein will lead to the asymmetric representation of hydrogen bonds. However, the hydrogen bonds between protein and ligand could be implicitly illustrated in our element interactive scheme. Although our discussion of the element specifics is designed for the protein−ligand system, this approach, with minimal effort, can be applied to a single biomolecular setting and other interactive models in chemistry and biology.

For convenience, let $\mathcal{T} = \{H, C, N, O, S, P, F, Cl, ...\}$ be the set of all interested element types in a given biomolecular data set. To reduce the notation complexity, we denote the element type at the ith position in the set $\mathcal{T}$ as $\mathcal{T}_i$. For example, $\mathcal{T}_2$ indicates the element type carbon. Assuming that a biomolecule has N atoms of interest. Then, we assign

$$\mathcal{X} = \{(\mathbf{r}_i, \alpha_i) | \mathbf{r}_i \in \mathbb{R}^3; \alpha_i \in \mathcal{T}; i = 1, 2, ..., N\}$$

as the collection of these N atoms annotated by their coordinates $\mathbf{r}_i$ and element types α_i. For a molecular complex, the collection of all atoms of type $\mathcal{T}_k$ and $\mathcal{T}_{k'}$ within a binding site defined by a cutoff distance d_c is denoted as

$$\mathcal{X}_{kk'} = \{\{\mathbf{r}_i, \mathbf{r}_j\}: \|\mathbf{r}_i - \mathbf{r}_j\| \le d_c; \quad \alpha_i \in \mathcal{T}_k \quad \text{and} \quad \alpha_j \in \mathcal{T}_{k'}\} \tag{6}$$

Before constructing the element interactive densities, we define the element interactive region $D_{kk'}$ for element type $\mathcal{T}_k$ and $\mathcal{T}_{k'}$ as the following

$$D_{kk'} = \{\mathbf{r} \in \bigcup_i B(\mathbf{r}_i, d_c) | \mathbf{r}_i \in \mathcal{X}_{kk'}\} \tag{7}$$

where $B(\mathbf{r}_i, d_c)$ is a ball with a center $\mathbf{r}_i$ and a radius d_c. The element interactive domain $\bar{D}_{kk'}$ is defined to be the smallest cube that enclosed $D_{kk'}$.

We now can design the element interactive density $\rho_{kk'}$, an atomic density defined in eq 1 but with a restraint on the element interactive region $\bar{D}_{kk'}$:

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

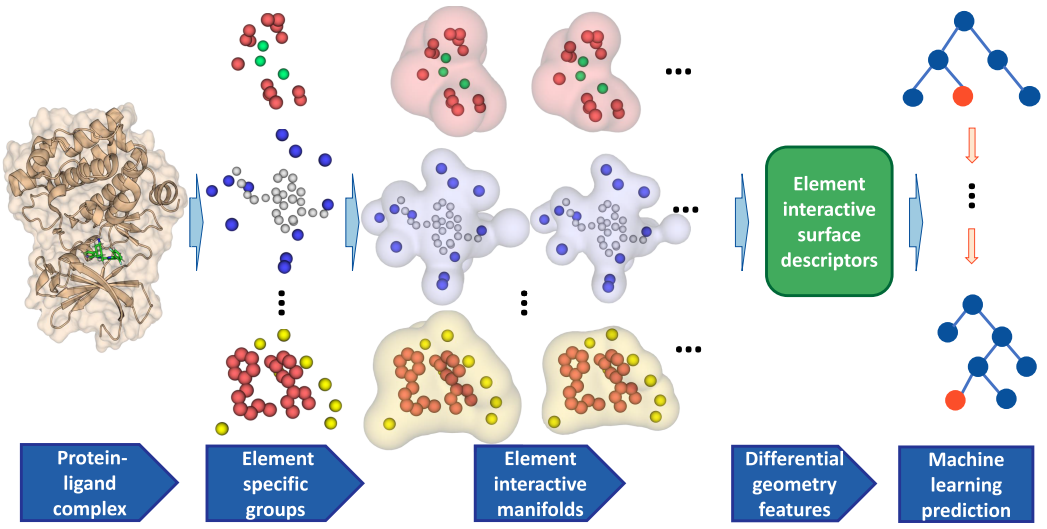


Figure 2. Illustration of EISA learning strategy using a complex with PDBID: 5dwr (first column). The second column represents the element-specific groups for carbon−fluoride, nitrogen−hydrogen, and oxygen−carbon from top to bottom, respectively. The corresponding element interactive manifolds are shown in the third column. Different manifolds are generated by varying the isovalue $0 < c < 1$. The fourth column presents the surface area-based descriptors obtained from various manifolds. In the final column, the advanced machine learning models such as the gradient boosting trees integrate these differential geometry features for training and prediction.

$$\rho_{kk'}(\mathbf{r}, \Phi) = \sum_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{kk'}), \quad \mathbf{r} \in \bar{D}_{kk'}, \mathbf{r}_j \in \mathcal{X}_{kk'} \tag{8}$$

and the normalized density function can be defined as

$$\hat{\rho}_{kk'}(\mathbf{r}, \Phi) = \frac{\rho_{kk'}(\mathbf{r}, \Phi)}{\max\{\rho_{kk'}(\mathbf{r}, \Phi)\}} \tag{9}$$

Figure 1 illustrates the atom collection $\mathcal{X}_{27}$, element interactive domain $\bar{D}_{27}$, and element interactive density ρ_{27} for element type $\mathcal{T}_2$ = C and $\mathcal{T}_7$ = F for the protein−ligand complex with PDBID: 5dwr.

In this work, we call the density function (8) the "global density" for the element types $\mathcal{T}_k$ and $\mathcal{T}_{k'}$. In addition, we desire to explore the "local density" formed by a single atom r_{i_o} with an element type $\mathcal{T}_k$ and all element type $\mathcal{T}_{k'}$ atoms:

$$\rho^{i_o}_{kk'}(\mathbf{r}, \Phi) = \Phi(\|\mathbf{r} - \mathbf{r}_{i_o}\|; \eta_{kk'}) + \sum_{j \neq i_o} \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{kk'}),$$
$$\mathbf{r} \in \bar{D}^{i_o}_{kk'}, \alpha_j = \mathcal{T}_{k'} \tag{10}$$

where $\bar{D}^{i_o}_{kk'}$ is a local element interactive cubic domain that enclosed $D^{i_o}_{kk'}$ which is defined as

$$D^{i_o}_{kk'} = \{\mathbf{r} \in \bigcup_{j \neq i_o} B(\mathbf{r}_j, d_c) \cup B(\mathbf{r}_{i_o}, d_c) | \alpha_j = \mathcal{T}_{k'},$$
$$j = 1, 2, ..., N\} \tag{11}$$

It is straightforward to verify that

$$D_{kk'} = \bigcup_{i_o} D^{i_o}_{kk'}, \quad i_o = 1, 2, ..., N \text{ and } \alpha_{i_o} = \mathcal{T}_k \tag{12}$$

The assembly of the local element interactive density $\rho^{i_o}_{kk'}$ enables our proposed model to examine the local interactions between a single atom of the element type $\mathcal{T}_k$ against a group of atoms of the element type $\mathcal{T}_{k'}$, capturing essential physical and chemical information across different biomolecular families that the global density might omit.

2.2. Element Interactive Surface Area. With ρ being a level set function defined on every grid point in an interested domain Ω, the isosurface Γ induced by ρ is given by $\Gamma = \{(x, y, z) \in \Omega: \rho(x, y, z) = c\}$, where c is the recommended isovalue. Assume $f(x, y, z)$ is the surface density function defined in Γ, the surface integral of f in Cartesian grids with a uniform mesh can be evaluated by[33,64]

$$\int_\Gamma f(x, y, z)\, dS = \sum_{(i,j,k) \in I_o} \left(f(x_o, y_j, z_k) \frac{|n_{o,x}|}{h} + f(x_i, y_o, z_k) \frac{|n_{o,y}|}{h} + f(x_i, y_j, z_o) \frac{|n_{o,z}|}{h} \right) h^3 \tag{13}$$

where h is the mesh size, (x_o, y_j, z_k) is the intersection point between the interface Γ and the x mesh line going through (i, j, k), and $n_{o,x}$ is the x component of the unit normal vector at (x_o, y_j, z_k). Similar definitions are used for the y and z directions. In addition, I_o is the set of irregular grid points. In our numerical scheme, a grid point is classified as irregular if its numerical difference's stencil involves neighbor point(s) from the other side of the interface Γ. One can find the surface area of Γ by considering the density function $f = 1$ in eq 13.

The intersection point (x_o, y_j, z_k) can be determined as described in ref 53 by

$$(x_o, y_j, z_k) = \left(\frac{\rho(x_o, y_j, z_k) - \rho(x_i, y_j, z_k)}{\rho(x_{i+1}, y_j, z_k) - \rho(x_i, y_j, z_k)} (x_{i+1} - x_i), y_j, z_k \right) \tag{14}$$

where $\rho(x_o, y_j, z_k) = c$, and the corresponding normal vector at (x_o, y_j, z_k) is interpolated by

$$\mathbf{N}_{o,j,k} = \frac{\rho(x_o, y_j, z_k) - \rho(x_i, y_j, z_k)}{\rho(x_{i+1}, y_j, z_k) - \rho(x_i, y_j, z_k)} (\mathbf{N}_{i+1,j,k} - \mathbf{N}_{i,j,k}) + \mathbf{N}_{i,j,k} \tag{15}$$

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

where $\mathbf{N}_{i,j,k}$ is the normal vector at the grid point (i, j, k) and is approximated by

$$\mathbf{N}_{i,j,k} = \left(\frac{\rho(x_{i+1}, y_j, z_k) - \rho(x_{i-1}, y_j, z_k)}{x_{i+1} - x_{i-1}}, \frac{\rho(x_i, y_{j+1}, z_k) - \rho(x_i, y_{j-1}, z_k)}{y_{j+1} - y_{j-1}}, \ldots, \frac{\rho(x_i, y_j, z_{k+1}) - \rho(x_i, y_j, z_{k-1})}{z_{k+1} - z_{k-1}} \right)$$

The volume integral of f is derived in a similar manner:

$$\int_{\Omega_m} f(x, y, z)\, \mathrm{d}S = \frac{1}{2}\left(\sum_{(i,j,k)\in I_1} f(x_i, y_j, z_k)h^3 + \sum_{(i,j,k)\in I_1 \cup I_o} f(x_i, y_j, z_k)h^3 \right) \quad (16)$$

Here I_1 contains all the grid points inside Ω_m, and I_o is the set of the irregular grid points defined at the surface area estimation eq 13. The desired volume of an enclosed molecular surface is attained by setting $f(x, y, z) = 1$.

2.3. Machine Learning Strategy with EISA. The descriptors of the element interactive surface area (EISA) for a molecule or molecular complex provide robustness and scalable features for machine learning or deep learning-based models to learn the diverse biomolecular data sets. The global and local element interactive densities, respectively defined in eqs 8 and 10, give rise to the corresponding global and local surface area descriptors. Furthermore, by varying the isovalue c for attaining the isosurface $\Gamma_{kk'} = \{(x, y, z) \in \Omega: \rho_{kk'} = c\}$ of the element interactive manifold, one can arrive at multiple surfaces for a given molecule at different resolutions. That enables us to capture molecular surfaces at various scales, which embed the physical and chemical interactions between protein and ligand atoms at different ranges. The learning strategy with EISA descriptors are summarized in Figure 2.

The EISA representations are ready to be integrated with wide variety of machine learning algorithms such as support vector machine,[8] random forest,[58] gradient boosting trees,[56] artificial neural networks,[32] and convolutional neural networks.[7] However, we only use gradient boosting trees (GBTs) in this work instead of optimizing machine learning algorithm selections. We use the GBT module in scikit-learn v0.24.1 package with the following parameters: `n_estimators` = 10000, `max_depth` = 7, `min_samples_split` = 3, `learning_rate` = 0.01, `loss` = ls, `subsample` = 0.3, and `max_features` = sqrt. These parameter values are selected from the extensive tests on PDBbind data sets and are uniformly used in all our validation tasks in this work.

3. RESULTS

In this section, we demonstrate the performance of the proposed element interactive surface area (EISA) strategy for protein−ligand binding affinity prediction from three standard benchmarks in drug design.

3.1. Model Parametrization. For convenience, we use the notation $\mathrm{EISA}^{\alpha}_{\kappa,\tau}$ to indicate the element interactive surface areas (EISAs) generated by using kernel type α and corresponding kernel parameters κ and τ. Here, $\alpha = E$ and $\alpha = L$ refer to the generalized exponential and generalized Lorentz kernels, respectively. And τ is used such that $\eta_{kk'} = \tau(\overline{r}_k + \overline{r}_{k'})$, where $\overline{r}_k$ and $\overline{r}_{k'}$ are the van der Waals radii of element type k and element type k', respectively. Kernel parameters κ and τ are selected based on the cross validation with a random split of the training data. We propose an EISA representation in which multiple kernels are parametrized at different scale (η) values. In this work, we consider at most two kernels. As a straightforward notation extension, two kernels can be parametrized by $\mathrm{EISA}^{\alpha_1,\alpha_2}_{\kappa_1,\tau_1;\kappa_2,\tau_2}$. Each of these kernels gives rise to one set of features. Since there are two ways of formulating the interactive surface areas, global surface (see eq 8) and local surface (see eq 10), we finalize our notation to demonstrate two different kinds of surface calculation: ${}^{\mathrm{glo}}\mathrm{EISA}^{\alpha_1,\alpha_2}_{\kappa_1,\tau_1;\kappa_2,\tau_2}$ and ${}^{\mathrm{loc}}\mathrm{EISA}^{\alpha_1,\alpha_2}_{\kappa_1,\tau_1;\kappa_2,\tau_2}$. While the first notation stands for the global interactive surface, the latter indicates the local surface area.

3.2. Data Sets. We are interested in using our EISA method to predict the binding affinities of protein−ligand complexes. A standard benchmark for such a prediction is the PDBbind database. Three popular PDBbind data sets, namely CASF-2007, CASF-2013, and CASF-2016, are employed to test the performance of our method. Each PDBbind data set has a hierarchical structure consisting of the following subsets: a general set, a refined set, and a core set. The latter set is a subset of the previous one. The PDBbind database provides 3D coordinates of ligands and their receptors obtained from experimental measurement via the Protein Data Bank. In each benchmark, it is standard to use the refined set, excluding the core set, as a training set to build a predictive model for the binding affinities of the complexes in the test set (i.e., the core set). More information about these data sets is offered on the PDBbind Web site http://pdbbind.org.cn/. A summary of the data set is provided in Table 1.

Table 1. Summary of PDBbind Datasets Used in the Present Work

Data set	Training set complexes	Test set complexes
CASF-2007 benchmark	1105	195
CASF-2013 benchmark	3516	195
CASF-2016 benchmark	3772	285

3.3. Model Performance and Discussion. *3.3.1. Hyperparameters and Model Setting.* To achieve the optimal EISA-score's performances on each benchmark, we carefully optimize its hyperparameters on each training set. These hyperparameters include kernel parameters $\tau \in [0.5, 6]$ and $\kappa \in [0.5, 10]$ with an increment of 0.5 and higher values in {15, 20}. Moreover, the cutoff distance d_c is between 5 Å and 14 Å with an increment of 1. (See Table 2 for the summary of the hyperparameters' domain.) The element interactive surface

Table 2. Ranges of EISA-Score's Hyperparameters

Parameter	Domain
τ	{0.5, 1.0, ..., 6}
κ	{0.5, 1, ..., 10} ∪ {15, 20}
d_c	{5, 6, ..., 14}
c	{0.05, 0.1, ..., 0.8}

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

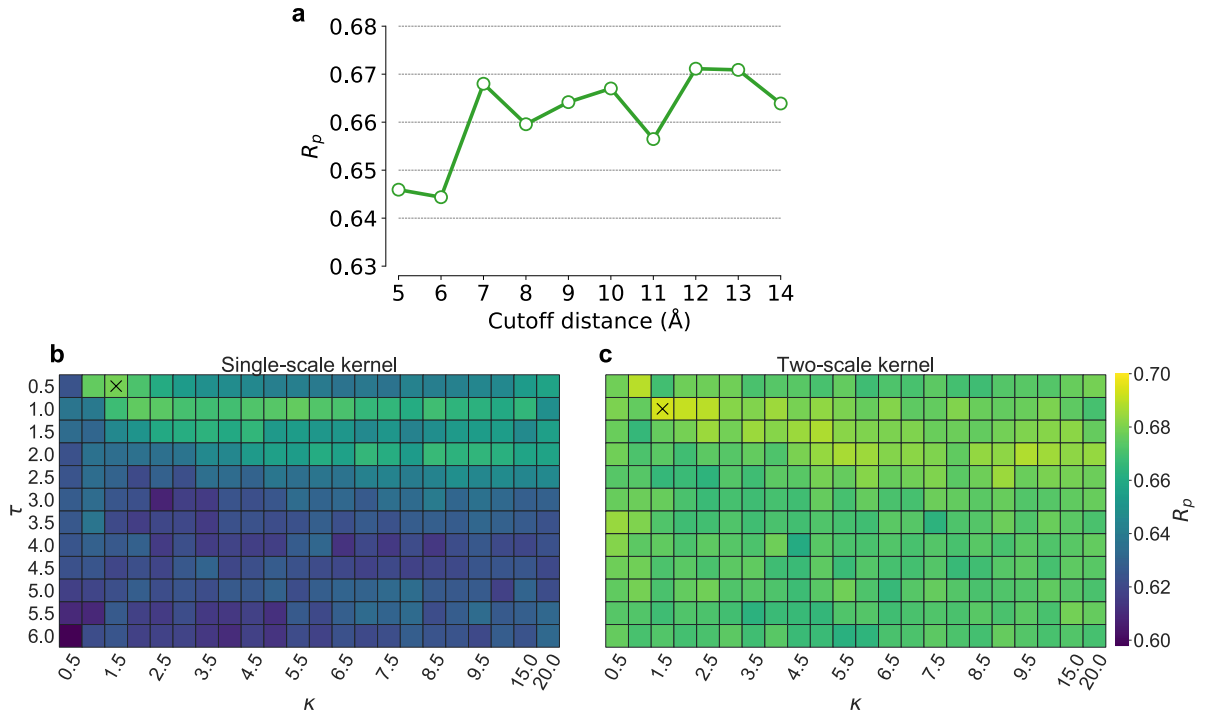


Figure 3. Optimized parameters for the global surface model on CASF-2007 data set. (a) Global surface 5-fold CV results for various cutoff distances on CASF-2007 training set. We fix the kernel parameters $(\kappa, \tau) = (2, 1)$. For a given cutoff distance value, we generate 16 surface areas based on 16 different isovalues in $[0.05, ..., 0.8]$. Global surface 5-fold CV results for CASF-2007 training set with (b) single-scale exponential kernel and (c) two-scale exponential kernel. The best parameter locations are marked by "x". Specifically, the best parameters for single-scale kernel model are $(\kappa, \tau) = (1.5, 0.5)$ with the corresponding Pearson's correlation coefficient $R_p = 0.678$. The second optimized kernel parameters are $(\kappa, \tau) = (1.5, 1)$ producing $R_p = 0.693$.

area is described by four commonly occurring atom types, C, N, O, S, in protein and 10 commonly atom types, H, C, N, O, F, P, S, Cl, Br, I, in ligands. Note that we only employ the generalized exponential kernel in the current work since the generalized Lorentz kernel yields similar accuracy.[57,58]

In global surface models, ^{glo}EISA, with a given set of kernel parameters (τ, κ), cutoff distance d_c, and a pair of element types $\mathcal{T}_k$ and $\mathcal{T}_{k'}$, we consider 16 isovalues c in the interval $[0.05, ..., 0.8]$ with an increment of 0.05. That results in 16 surface area values. We then achieve 6 descriptors by taking the sum, mean, median, maximum, minimum, and standard deviation of those area values. Furthermore, there are $4 \times 10 = 40$ combinations between protein and ligand element types. Finally, we encode the binding interaction in a protein−ligand complex into a vector of fixed length at $6 \times 40 = 240$ components.

In local surface models, ^{loc}EISA, besides using a similar hyperparameter setting of the global approach, we only select one single isovalue c in $[0.1, 0.75]$ for element types $\mathcal{T}_k$ and $\mathcal{T}_{k'}$. However, a different atomic position from the element type $\mathcal{T}_k$ will generate a different interactive manifold resulting in a different surface area. To get a scalable representation, we calculate the sum, mean, median, maximum, minimum, and standard deviation of those various values. With 40 possible combinations between protein and ligand atom types, one can attain a descriptor of size 240 for a given complex.

It is worth noting that our EISA features for both global and local surface models are simple and easy to calculate—the only required data input is the atomic names and coordinates of the complexes. For a given protein−ligand complex, it takes on average 1 min and 30 s to generate the features on a single processor with mesh size $h = 0.5$. Furthermore, our EISA models can be easily parallelized to accommodate the virtual high-throughput screening.

3.3.2. Results and Discussion. Several hyperparameters of our proposed models, ^{glo}EISA, need to be carefully optimized for each benchmark. For the sake of achieving fair performances, we only use the training data set to carry on the grid search on the designated domains mentioned in Table 2 via the results of the cross validation (CV) tests. We execute 20 CV runs for each hyperparameter set, and the criteria are based on the best median Pearson's correlation coefficient R_p.

CASF-2007. At first, we carry out the 5-fold CV of the global surface models ^{glo}EISA on the CASF-2007 training data. To explore the optimal cutoff distance values d_c, we fix the kernel parameters $(\kappa, \tau) = (2, 1)$, select 16 isovalues c in $[0.05, ..., 0.8]$, but vary the d_c between 5 Å and 14 Å with increments of 1. Figure 3a reveals that $d_c = 12$ Å yields the best median Pearson's correlation coefficient R_p.

We now explore the optimal exponential kernel parameters for a single-scale model $^{glo}\text{EISA}^{E,12}_{\kappa,\tau}$ where $\tau \in [0.5, 6]$ and $\kappa \in [0.5, 10]$ with increment of 0.5. We also consider high values of $\kappa \in \{15, 20\}$. Figure 3b plots all the CV results and shows that $(\kappa, \tau) = (1.5, 1)$ gives the best median $R_p = 0.678$ for the global surface model. The two-scale kernel model, $^{glo}\text{EISA}^{EE,12}_{1.5,0.5,\kappa_2,\tau_2}$, is built on top of the previously optimized single scale. The optimal second kernel parameters (κ_2, τ_2) are explored via CV experiments, and the result of each parameter combination is illustrated in Figure 3c. We found that $^{glo}\text{EISA}^{E,12}_{1.5,0.5,1.5,1}$ produces the best median $R_p = 0.693$ on the CASF-2007 training set. It is interesting to observe that the

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

single-kernel model ${}^{glo}\mathrm{EISA}^{EE,12}_{1.5,0.5,\kappa_2,\tau_2}$ performs well on the 195 complexes from the test set of the CASF-2007 benchmark with the reported R_p = 0.801 and the root-mean-square error (RMSE) = 2.01 kcal/mol. While the two-scale model ${}^{glo}\mathrm{EISA}^{EE,12}_{1.5,0.5,1.5,1}$ performs slightly better than its predecessor and achieves R_p = 0.807 and RMSE = 2.00 kcal/mol. Those results are reported in Table 3. Furthermore, we are interested to see the feature importance of our proposed model. Figure 6a reveals the top 10 most important features among the 240 features of the global surface model. It is interesting to find that the hydrophobic interactions (C−C) and polar−nonpolar interactions (C−N, C−O, C−S, N−C) are among the top 10 important features. The ranking of all 240 features of the global surface model is provided in Table S1 in the Supporting Information.

Table 3. Performance of Various EISA Models on the CASF-2007 Test Set

Model	R_p	RMSE (kcal/mol)
Results of Global Surface Model		
${}^{glo}\mathrm{EISA}^{E,12}_{1.5,0.5}$	0.801	2.01
${}^{glo}\mathrm{EISA}^{EE,12}_{1.5,0.5;1.5,1}$	0.807	2.00
Results with Local Surface		
${}^{loc}\mathrm{EISA}^{E;6.5;0.15}_{15,0.5}$	0.807	1.986
${}^{loc}\mathrm{EISA}^{EE;6.5;0.15}_{15,0.5;2,2}$	0.793	2.046
Results with Consensus Method		
Consensus$\{{}^{glo}\mathrm{EISA}^{E,12}_{1.5,0.5}, {}^{loc}\mathrm{EISA}^{E;6.5;0.15}_{15,0.5}\}$	**0.825**	**1.941**
Consensus$\{{}^{glo}\mathrm{EISA}^{EE,12}_{1.5,0.5;1.5,1}, {}^{loc}\mathrm{EISA}^{EE;6.5;0.15}_{15,0.5;2,2}\}$	0.817	1.984

The second kind of our surface-based model is the local surface based approach, ^{loc}EISA, that measures the various different surface areas between a single protein atom and all the ligand atoms. There is a slight difference in terms of the parameter choice between the global and local models. While the global surface areas utilize various isovalues between 0.05 and 0.8, the local surface approach will explore the isovalue to generate the best surface model. But at first, while we fix the isovalue c = 0.25, and cutoff distance d_c = 5 Å, we vary the kernel parameters κ and τ in their designated domains (see Table 2). Figure 4a visualizes that CV test and reports the best kernel parameters (κ = 15, τ = 0.5) with R_p = 0.688. In the next step, we investigate the best cutoff distance d_c for the local surface based model with previously optimized single-kernel parameters (κ = 15 and τ = 0.5) and an isovalue c = 0.25. In this experiment, we vary d_c between 4 Å and 7 Å, with increment of 0.5, then we find out the optimal cutoff distance is 6.5 Å, which produces the median R_p = 0.701 on the 5-fold CV of the CASF-2007 training set, see Figure 4c.

The isovalue c is the next parameter we would like to optimize for our local surface model, ${}^{loc}\mathrm{EISA}^{E,6.5,c}_{15,0.5}$. We search c in the discrete domain between 0.1 and 0.75 with increments of 0.5. Figure 4d reveals that using isovalue c = 0.15 will be the best choice for ${}^{loc}\mathrm{EISA}^{E,6.5,c}_{15,0.5}$ with the reported median R_p = 0.712. Similar to the global surface model, we are interested in extending the single-scale EISA-score ${}^{loc}\mathrm{EISA}^{E,6.5,0.15}_{15,0.5}$ to the two-scale one ${}^{loc}\mathrm{EISA}^{E,6.5,0.15}_{15,0.5,\kappa_2,\tau_2}$. Figure 4b summarizes the performances of the current model on the 5-fold experiments with respect to different values of κ_2 and τ_2, and we conclude

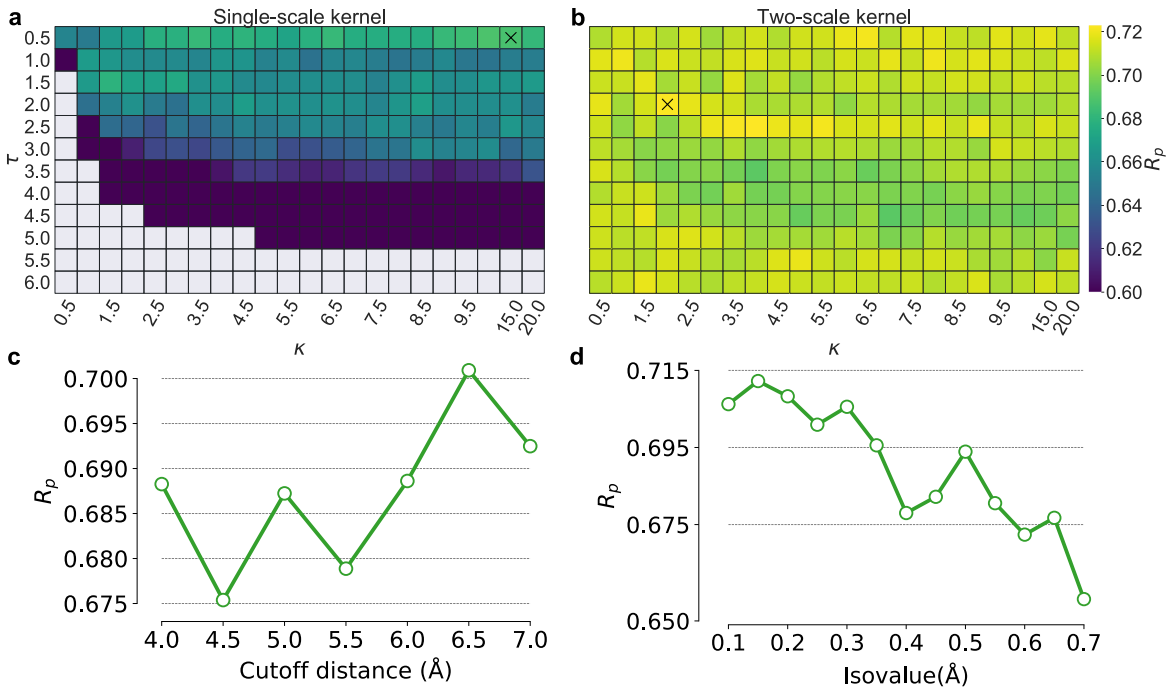


Figure 4. Optimized parameters for the local surface model on CASF-2007 data set. (a) Local surface 5-fold CV results for single-scale scenario, and it is found that the best kernel parameters (κ, τ) = (15, 0.5) and the corresponding median R_p = 0.688. Note that there are empty values in the panel a since R_p cannot be determined at the choice of κ and τ; (b) 5-fold CV results for single-scale approach and the best second kernel parameters are (κ, τ) = (2, 2) producing the best R_p = 0.726. The marker "x" indicates the position having the best R_p. (c) The 5-fold CV results of the local surface model with respect to the cutoff distance d_c. The best cutoff distance is d_c = 6.5 Å, and R_p = 0.701. (d) The 5-fold CV results of the local surface model when the isovalue c varies from 0.1 to 0.7. Optimal isovalue c is found to be 0.15 and the corresponding R_p is 0.712.

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

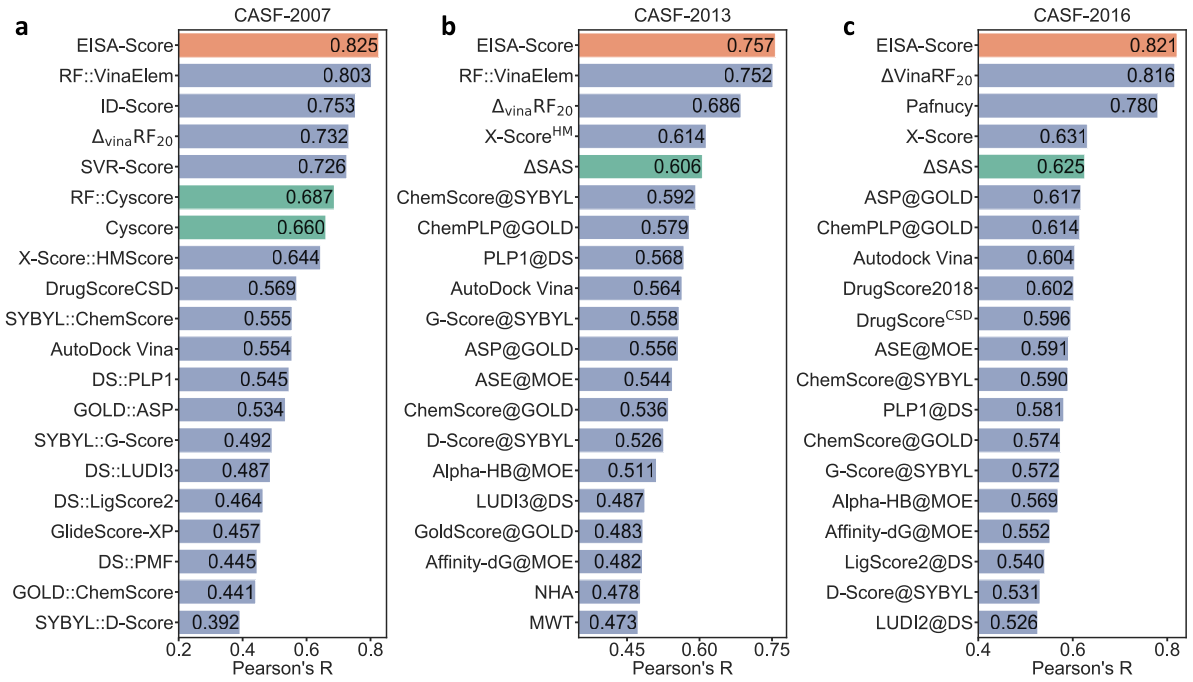


Figure 5. Performance comparison of different scoring functions on the CASF benchmarks. Our proposed model in this work, EISA-score, is highlighted in red, the other geometric based scoring functions are highlighted in green, and the rest is in purple. (a) CASF-2007: the performances of other methods taken from previous studies.[2,9,20,42−44,72] Our EISA-score achieves R_p = 0.825 and RMSE = 1.941 kcal/mol. (b) CASF-2013: the other results are extracted from previous studies.[44,46,72] Our EISA-score achieves R_p = 0.757 and RMSE = 2.113 kcal/mol. (c) CASF-2016: our EISA-score achieves R_p = 0.821 and RMSE = 1.835 kcal/mol, other scoring functions are discussed in the references.[67,68,72]

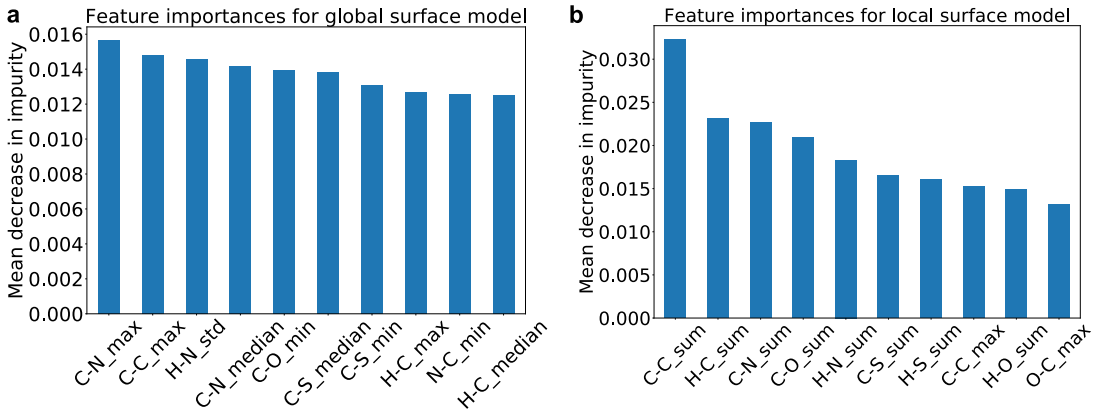


Figure 6. Feature importance for (a) global surface model and (b) local surface model using the mean decrease in impurity of the gradient boosting trees. The horizontal axis represents the feature names which are textualized as the statistical measures (sum, mean, median, etc.) of the ligand-protein element specific group.

that $(\kappa_2, \tau_2) = (2, 2)$ gives us that optimal two-scale learner achieving the best median R_p = 0.726.

Table 3 reports the efficiency of one-kernel and two-kernel local surface models on the CASF-2007 test set. Interestingly, with only one-single scale, the local surface model, ${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{E;6.5;0.15}}_{15,0.5}$ performs similarly to the two-scale approach using the global surface features. Its R_p value is 0.807 but its RMSE is as low as 1.986 kcal/mol and is lower than that of the global surface model. Unfortunately, the two-kernel version of the local EISA-score does not improve what the one-kernel has already achieved. In fact, the R_p of ${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE;6.5;0.15}}_{15,0.5;2,2}$ is just 0.793, and the corresponding RMSE is 2.046 kcal/mol. The consensus model which is the aggregation of the predicted values from unrelated models is acclaimed to often improve the overall performance.[57,57,58] For that reason, we include the consensus version in our proposed models. As seen from Table 3, the consensus approach, formed by a single scale between the global and local surface areas, Consensus$\{{}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{E,12}}_{1.5,0.5}, {}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{E;6.5;0.15}}_{15,0.5}\}$, gives rise to the best one with R_p = 0.825 and RMSE = 1.941 kcal/mol. While the consensus of the two-scales models produce the second best R_p at 0.817 and RMSE at 1.984 kcal/mol. Moreover, it is interesting to find that the top five important features among the 240 features of the local surface model are the summation of the surface area of the element type pairs (C−C, H−C, C−N, C−O, H−N). It is also exciting to see that the maximum of the local surface areas of the element type interactions C−C and O−O is in the top 10 positions. Figure 6b shows the 10 most important features for

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

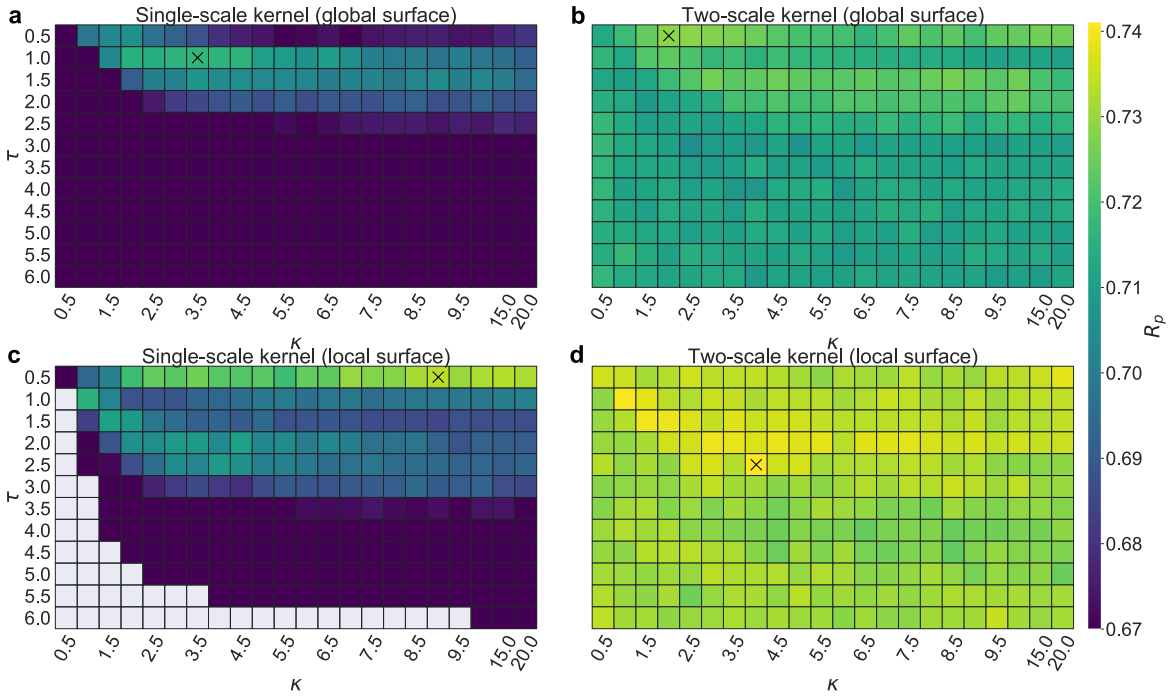


Figure 7. Optimized parameters of the global and local surface models for CASF-2013. Note that the marker "x" indicates the position having the best R_p and the empty values in the panels are due to the fact that R_p cannot be determined at the choice of κ and τ. (a) Single-scale global surface and its optimal kernel parameters $(\kappa, \tau) = (3.5, 1)$ and corresponding median $R_p = 0717$. (b) Two-scale kernel global surface and its optimal parameters for the second kernel $(\kappa, \tau) = (2, 0.5)$ and corresponding median $R_p = 0729$. (c) Single-scale local surface model and its optimal parameters for the second kernel $(\kappa, \tau) = (3, 1)$ and corresponding median $R_p = 0.715$. (d) Two-scale local surface model and its optimal parameters for the second kernel $(\kappa, \tau) = (3, 0.5)$ and corresponding median $R_p = 0.727$.

the local surface model. The ranking of all 240 features of the local surface model is provided in Table S2 in the Supporting Information.

In addition, we compare the scoring power of our proposed EISA-score against the state-of-the-art scoring functions in the literature.[2,9,20,42−44] Figure 5a plots the aforementioned comparison and clearly the dominance of our EISA model in the scoring power task. Note that the geometrical-based models, Cyscore[9] and RF::Cyscore,[43] are highlighted in the green color. Specifically, Cyscore used area and curvature dependent descriptors. However, its performance ($R_p = 0.660$) is not as good as our proposed EISA-score ($R_p = 0.825$) due to the lack of the examination of pairwise element types inducing interactive manifolds. Furthermore, one can cite another reason is the missing machine learning power in the Cyscore model. However, Li and his colleague[43] solved that concern by replacing the Cyscore's original scoring function by the random forest, and the result is not promising with the reported R_p as low as 0.687. These results confirm the efficiency and robustness of the proposed element specific surface area based descriptors for protein−ligand complexes.

CASF-2013. In this second benchmark among the CASF family, we carry out the similar hyperparameters optimization strategy to the CASF-2007 approach. For simplicity, we use the optimized cutoff distance $d_c = 12$ Å, found from the CASF-2007 data set for the global surface model. To explore the most optimal parameters for the first kernel, we again perform 5-fold CV on CASF-2013 training data and find out that ($\kappa_1 = 3.5$, $\tau_1 = 1$) gives the best $R_p = 0.717$ (see Figure 7a). To construct the second kernel, we simply fix the first kernel parameters, and vary the second kernel parameter in the interested domain (see Table 2). We found the best parameter for the second kernel $\kappa_2 = 2$ and $\tau_2 = 0.5$ with best median $R_p = 0.729$ (see Figure 7b). Finally, we achieve the optimal one-kernel global surface model ${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{E},12}_{3.5,1}$ and the optimal two-kernels global surface model ${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE},12}_{3.5,1;2,0.5}$. These models are utilized to predict the unseen complexes from the test set of CASF-2013. As seen from Table 4, the performances of single kernel and two kernels, respectively, achieve ($R_p = 0.684$, RMSE = 2.286 kcal/mol) and ($R_p = 0.724$, RMSE = 2.180 kcal/mol). There is a considerable improvement from the single kernel to two kernels model in compassion to the what we have observed in CASF-2007. The size of the training set (1105 for CASF-2007 and 3516 for CASF-2013) can play a huge factor role in our multiscale strategy.

Table 4. Performance of Various EISA Models on the CASF-2013 Test Set

Model	R_p	RMSE (kcal/mol)
Results with Global Surface		
${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{E},12}_{3.5,1}$	0.684	2.286
${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE},12}_{3.5,1;2,0.5}$	0.724	2.180
Results with Local Surface		
${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{E};6.5;0.15}_{9,0.5}$	0.749	**2.102**
${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE};6.5;0.15}_{9,0.5;4,2.5}$	0.741	2.129
Results with Consensus Method		
Consensus$\{{}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{E},12}_{3.5,1}, {}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{E};6.5;0.15}_{9,0.5}\}$	0.741	2.155
Consensus$\{{}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE},12}_{3.5,1;2,0.5}, {}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE};6.5;0.15}_{9,0.5;4,2.5}\}$	**0.756**	2.113

To reduce the search time cost of hyperparameters for the local surface approach, we use the optimized cutoff distance $d_c = 6.5$ Å, and the isovalue $c = 0.15$ which are explored from the

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

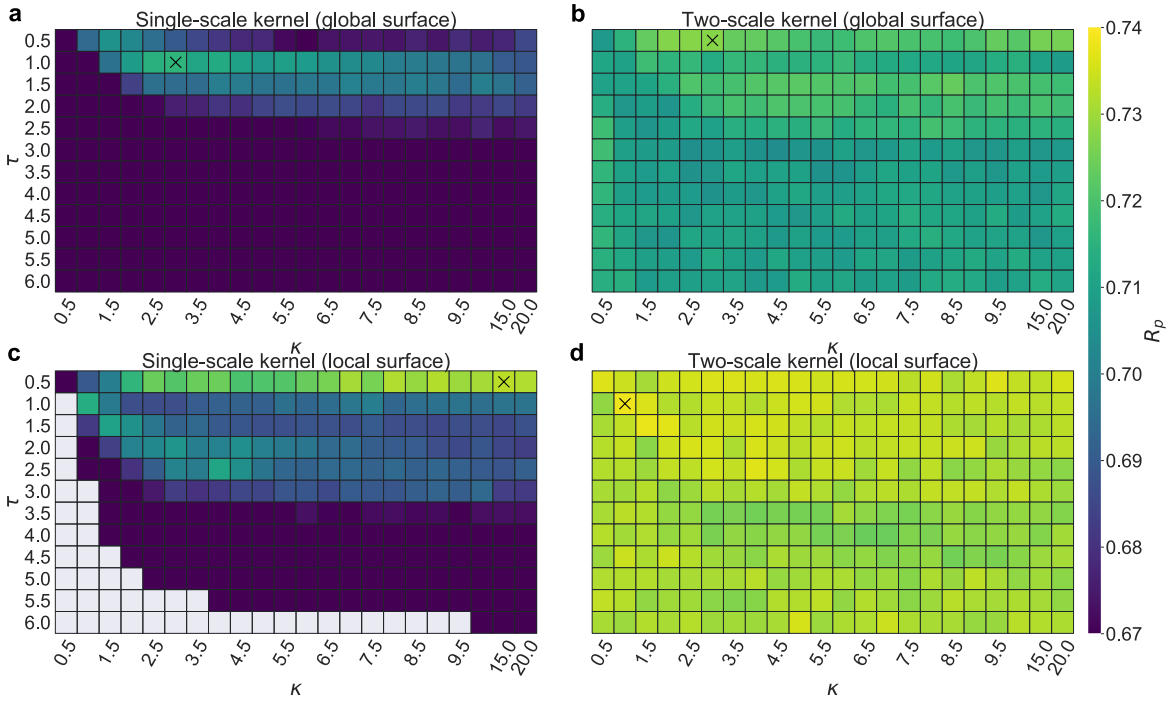


Figure 8. Optimized parameters of the global and local surface models for CASF-2016. Note that the marker "x" indicates the position having the best R_p, and the empty values in the panels are due to the fact that R_p cannot be determined at the choice of κ and τ. (a) Single-scale global surface and its optimal kernel parameters $(\kappa, \tau) = (3, 1)$ and corresponding median $R_p = 0.715$. (b) Two-scale kernel global surface and its optimal parameters for the second kernel $(\kappa, \tau) = (3, 0.5)$ and corresponding median $R_p = 0.727$. (c) Single-scale local surface model and its optimal parameters for the second kernel $(\kappa, \tau) = (15, 0.5)$ and corresponding median $R_p = 0.733$. (d) Two-scale local surface model and its optimal parameters for the second kernel $(\kappa, \tau) = (1, 1)$ and corresponding median $R_p = 0.738$.

CASF-2007 experiment. These parameters are pretty consistent among different protein−ligand complexes. Therefore, we speculate there is little room for improvement if we reoptimize those parameters. Similar to the global surface scheme, we first search for the optimal one-kernel model. Figure 7c plots the 5-fold CV results of $\mathrm{EISA}^{\mathrm{E;6.5;0.15}}_{\kappa_1,\tau_1}$ on the training set of CASF-2013, and we conclude that $k_1 = 0$ and $\tau_1 = 0.5$ will yield the best $R_p = 0.734$. Again, for the two-kernel model $\mathrm{EISA}^{\mathrm{EE;6.5;0.15}}_{9,0.5;\kappa_2,\tau_2}$, we use the optimized value from the single-scale model for the first kernel, and explore the optimal ones for the second kernel. As see in Figure 7d, $\kappa_2 = 4$ and $\tau_2 = 2.5$ produces the best $R_p = 0.741$. Finally, we evaluate the scoring power of two selected local surface models, ${}^{\mathrm{local}}\mathrm{EISA}^{E;6.5;0.15}_{9,0.5}$ and ${}^{\mathrm{loc}}\mathrm{EISA}^{EE;6.5;0.15}_{9,0.5;4,2.5}$, on the CASF-2013 test set. It is comparable to what we observed in CASF-2007, the one-scale local surface model ($R_p = 0.749$) performs a bit better than its counterpart ($R_p = 0.741$), albeit a bigger training data. Our optimal strategy still relies on the consensus design where the consensus between two two-scale models, $\mathrm{Consensus}\{{}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE,12}}_{3.5,1;2,0.5}, {}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE;6.5;0.15}}_{9,0.5;4,2.5}\}$, delivers the best R_p as high as 0.756 and the corresponding RMSE = 2.113 kcal/mol. (See Table 4 for the completion of results). Our EISA-score again tops other published models on CASF-2013 as indicated in Figure 5b. It is worth mentioning that we also include the other surface area-based model, ΔSAS,[46] which used the solvent-accessible surface area of the buried ligand molecule when forming the complex. However, ΔSAS's performance is not promising with an R_p as low as 0.606 due to the lack of the greater details of the buried surface for specific element types.

CASF-2016. For this final benchmark, we perform the hyperparameters search similar to what we proposed for CASF-2013. The global surface design will use the ideal distance cutoff $d_c = 12$ Å obtained from the CASF-2007 experiments. The optimal single-scale model for CASF-2016 is found to be ${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{E,12}}_{3,1}$, where its R_p from the 5-fold CV on 3772 complexes of CASF-2016 training set is equal to 0.715. On top of this single-scale model, the two-scale continues to improve the CV performances with its best model being ${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE,12}}_{3,1;3,0.5}$, and its $R_p = 0.727$. Figure 8 panels a and b summarize the CV results for various kernel parameter combinations.

The local surface approach uses the optimal isovalue $c = 0.15$ and distance cutoff $d_c = 6.5$ Å realized from CASF-2007 5-fold results. Figures 8c,d reports the CV results on CASF-2016 with respect to the single-scale and two-scale parameter choices. Specifically, ${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{E;6.5;0.15}}_{15,0.5}$ is the best single-scale representative with the median 5-fold CV $R_p = 0.727$. Furthermore, the best two-scale candidate is found to be ${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE;6.5;0.15}}_{15,0.5;1,1}$ with the corresponding median 5-fold CV $R_p = 0.738$.

Lastly, the aforementioned desirable EISA models are trained on the training data of CASF-2016 and are utilized to predict the binding energies of 285 complexes in CASF-2016 test set. Table 5 lists the results of these models including the consensus strategies. The familiar trend has been observed here. Two-scale models, ${}^{\mathrm{glo}}\mathrm{EISA}^{\mathrm{EE,12}}_{3,1;3,0.5}$ and ${}^{\mathrm{loc}}\mathrm{EISA}^{\mathrm{EE;6.5;0.15}}_{15,0.5;1,1}$, bring about the most outstanding performance among the nonconsensus ones. In addition, the consensus models improve the existing methods. Specifically, Consensus{

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

Table 5. Performance of Various EISA Models on the CASF-2016 Test Set

Model	R_p	RMSE (kcal/mol)
Results with Global Surface		
${}^{glo}EISA^{E,12}_{3,1}$	0.769	1.989
${}^{glo}EISA^{EE,12}_{3,1;3,0.5}$	0.798	1.888
Results with Local Surface		
${}^{loc}EISA^{E;6.5;0.15}_{15,0.5}$	0.791	1.883
$locEISA^{EE;6.5;0.15}_{15,0.5;1,1}$	0.795	1.881
Results with Consensus Method		
Consensus$\{{}^{glo}EISA^{E,12}_{3,1}, {}^{loc}EISA^{E;6.5;0.15}_{15,0.5}\}$	0.813	1.873
Consensus$\{{}^{glo}EISA^{EE,12}_{3,1;3,0.5}, locEISA^{EE;6.5;0.15}_{15,0.5;1,1}\}$	**0.821**	**1.835**

${}^{glo}EISA^{EE,12}_{3,1;3,0.5}$, ${}^{loc}EISA^{EE;6.5;0.15}_{15,0.5;1,1}\}$ reaches the $R_p = 0.821$ on the test set while its stand-alone models ${}^{glo}EISA^{EE,12}_{3,1;3,0.5}$ and ${}^{loc}EISA^{EE;6.5;0.15}_{15,0.5;1,1}$ scores $R_p = 0.798$ and $R_p = 0.795$, respectively. CASF-2016 is a prevalent benchmark which attracts numerous scoring functions relying on it to test their scoring power.[67,68,72] As seen in Figure 5c, it is encouraging to see our EISA-score outperforming other state-of-the-art methods. It is noted that, among other 20 scoring functions listed in Figure 5c, only ΔSAS[68] solely leans on the surface descriptors. However, its performance on CASF-2016 is unfavorable with $R_p = 0.625$ as opposed to 0.821 of our proposed EISA-score. This result again confirms the rigorous and robust capacity of our novel surface area-based descriptors for drug design.

4. CONCLUSION

Molecular surface representations are well-known for biological structure modeling to reveal biomolecular properties and activities. However, their relationship to the biological functions is often encoded in the realm of the physical models such as Poisson−Boltzmann equation and Poisson−Nernst−Planck model. Unfortunately, the problematic parameter choices of these physical models have overshadowed the valuable information extracted from the molecular surface. There are some recent efforts to directly incorporate the surface area descriptors to capture the protein−ligand potency.[9,26] However, conventional surface area models do not portray crucial physical and chemical interactions such as noncovalent bonds, hydrogen bonds, van der Waals interactions, etc., which lead to discouraging results and limited capacity to handle diverse biomolecular data sets. These issues call for robustness and scalable surface area representations for biomolecular structures.

This work proposes a novel element interactive surface area score (EISA-score) for protein−ligand binding prediction and can be extended to handle drug-related problems. Our proposed models construct scalable element interactive manifolds instead of a single surface representation for a whole complex often used in the standard approaches. The innovative surface areas help encode the physical and biological information mentioned above, which have been missed in conventional methods. Our EISA-score offers two types of surface area models, namely global and local surface. Specifically, while the global surface area strategy provides the overall molecular representation between protein and ligand atoms, the local approach focuses on describing the local manifold formed by a specific protein atom and ligand molecule. Our molecular surfaces are induced by the discrete-to-continuum mapping powered by the correlation function such as exponential and Lorentz kernels.

Due to the high sensitivity of the hyperparameters, including isovalue, kernel power, and kernel scalar factor in our surface generation, we carefully perform the cross validation on the training data to select the optimal surface descriptors for the protein−ligand complexes. As a result, our proposed EISA-score achieves superior performances over state-of-the-art methods on three mainstream benchmarks, namely CASF-2007,[20] CASF-2013,[46] and CASF-2016.[68] These encouraging results confirm our surface-area-based models' robustness, reliability, and accuracy in the binding affinity prediction for small molecules, which is an essential task in drug design.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code is available at Github: https://github.com/NguyenLabUKY/EISA-Score.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00697.

List of the feature names and their corresponding mean decrease in impurity (MDI) for both the Global Surface Model and the Local Surface Model (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Duc Duy Nguyen − *Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506, United States;* orcid.org/0000-0003-2215-0328; Email: ducnguyen@uky.edu

Author

Md Masud Rana − *Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00697

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by NSF Grants DMS-2053284, DMS-2151802, and University of Kentucky Startup Fund.

■ REFERENCES

(1) Baker, N. A. Improving Implicit Solvent Simulations: a Poisson-Centric View. *Curr. Opin. Struct. Biol.* **2005**, *15* (2), 137−143.

(2) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26* (9), 1169−1175.

(3) Bates, P. W.; Chen, Z.; Sun, Y.; Wei, G.-W.; Zhao, S. Geometric and Potential Driving Formation and Evolution of Biomolecular Surfaces. *J. Math. Biol.* **2009**, *59* (2), 193−231.

(4) Bates, P. W.; Wei, G. W.; Zhao, S. The Minimal Molecular Surface. *arXiv* **2006**, No. q-bio/0610038.

(5) Bates, P. W.; Wei, G.-W.; Zhao, S. Minimal Molecular Surfaces and Their Applications. *J. Comput. Chem.* **2008**, *29* (3), 380−391.

(6) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46* (4), 558−570.

(7) Cang, Z.; Mu, L.; Wei, G.-W. Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. *PLoS Comput. Biol.* **2018**, *14* (1), No. e1005929.
(8) Cang, Z.; Mu, L.; Wu, K.; Opron, K.; Xia, K.; Wei, G.-W. A Topological Approach for Protein Classification. *Comput. Math. Biophys.* **2015**, *3* (1), 140.
(9) Cao, Y.; Li, L. Improved Protein-Ligand Binding Affinity Prediction by Using a Curvature-Dependent Surface-Area Model. *Bioinformatics* **2014**, *30* (12), 1674−1680.
(10) Chan, S. L.; Purisima, E. O. Molecular Surface Generation Using Marching Tetrahedra. *J. Comput. Chem.* **1998**, *19* (11), 1268−1277.
(11) Chen, D.; Chen, Z.; Chen, C.; Geng, W.; Wei, G.-W. MIBPB: a Software Package for Electrostatic Analysis. *J. Comput. Chem.* **2011a**, *32* (4), 756−770.
(12) Chen, D.; Chen, Z.; Wei, G.-W. Quantum Dynamics in Continuum for Proton Transport II: Variational Solvent-Solute Interface. *Int. j. numer. method. biomed. eng.* **2012**, *28* (1), 25−51.
(13) Chen, D.; Wei, G.-W. Quantum Dynamics in Continuum for Proton Transport-Generalized Correlation. *J. Chem. Phys.* **2012**, *136* (13), 04B606.
(14) Chen, D.; Wei, G.-W. Quantum Dynamics in Continuum for Proton Transport I: Basic Formulation. *Commun. Comput. Phys.* **2013**, *13* (1), 285−324.
(15) Chen, W.; Zheng, J.; Cai, Y. Kernel Modeling for Molecular Surfaces Using a Uniform Solution. *Comput. Des.* **2010**, *42* (4), 267−278.
(16) Chen, Z.; Baker, N. A.; Wei, G.-W. Differential Geometry Based Solvation Model I: Eulerian Formulation. *J. Comput. Phys.* **2010**, *229* (22), 8231−8258.
(17) Chen, Z.; Baker, N. A.; Wei, G.-W. Differential Geometry Based Solvation Model II: Lagrangian Formulation. *J. Math. Biol.* **2011**, *63* (6), 1139−1200.
(18) Chen, Z.; Wei, G.-W. Differential Geometry Based Solvation Model. III. Quantum Formulation. *J. Chem. Phys.* **2011**, *135* (19), 194108.
(19) Chen, Z.; Zhao, S.; Chun, J.; Thomas, D. G.; Baker, N. A.; Bates, P. W.; Wei, G. W. Variational Approach for Nonpolar Solvation Analysis. *J. Chem. Phys.* **2012**, *137* (8), 84101.
(20) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49* (4), 1079−1093.
(21) Connolly, M. L. Depth-Buffer Algorithms for Molecular Modelling. *J. Mol. Graph.* **1985**, *3* (1), 19−24.
(22) Corey, R. B.; Pauling, L. Molecular Models of Amino Acids, Peptides, and Proteins. *Rev. Sci. Instrum.* **1953**, *24* (8), 621−627.
(23) Crowley, P. B.; Golovin, A. Cation-!!insert-Eqn352/!! Interactions in Protein-Protein Interfaces. *Proteins Struct. Funct. Bioinforma.* **2005**, *59* (2), 231−239.
(24) Daberdaku, S.; Ferrari, C. Computing Voxelised Representations of Macromolecular Surfaces: A Parallel Approach. *Int. J. High Perform. Comput. Appl.* **2018**, *32* (3), 407−432.
(25) Das, S.; Kokardekar, A.; Breneman, C. M. Rapid Comparison of Protein Binding Site Surfaces with Property Encoded Shape Distributions. *J. Chem. Inf. Model.* **2009**, *49* (12), 2863−2872.
(26) Dong, L.; Qu, X.; Zhao, Y.; Wang, B. Prediction of Binding Free Energy of Protein-Ligand Complexes with a Hybrid Molecular Mechanics/Generalized Born Surface Area and Machine Learning Method. *ACS omega* **2021**, *6* (48), 32938−32947.
(27) Dragan, A. I.; Read, C. M.; Makeyeva, E. N.; Milgotina, E. I.; Churchill, M. E. A.; Crane-Robinson, C.; Privalov, P. L. DNA Binding and Bending by HMG Boxes: Energetic Determinants of Specificity. *J. Mol. Biol.* **2004**, *343* (2), 371−393.
(28) Edelsbrunner, H.; Mücke, E. P. Three-Dimensional Alpha Shapes. *ACM Trans. Graph.* **1994**, *13* (1), 43−72.
(29) Egan, R.; Gibou, F. Fast and Scalable Algorithms for Constructing Solvent-Excluded Surfaces of Large Biomolecules. *J. Comput. Phys.* **2018**, *374*, 91−120.
(30) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11* (5), 425−445.
(31) Fraczkiewicz, R.; Braun, W. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J. Comput. Chem.* **1998**, *19* (3), 319−333.
(32) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D Fingerprints Still Valuable for Drug Discovery? *Phys. Chem. Chem. Phys.* **2020**, *22* (16), 8373−8390.
(33) Geng, W.; Wei, G.-W. Multiscale Molecular Dynamics Using the Matched Interface and Boundary Method. *J. Comput. Phys.* **2011**, *230* (2), 435−457.
(34) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99* (11), 3503−3510.
(35) Grant, J. A.; Pickup, B. T.; Nicholls, A. A Smooth Permittivity Function for Poisson−Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22* (6), 608−640.
(36) Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. The GaussianGeneralizedBorn Model: Application to Small Molecules. *Phys. Chem. Chem. Phys.* **2007**, *9* (35), 4913−4922.
(37) Hayryan, S.; Hu, C.-K.; Skřivánek, J.; Hayryane, E.; Pokornỳ, I. A New Analytical Method for Computing Solvent-Accessible Surface Area of Macromolecules and Its Gradients. *J. Comput. Chem.* **2005**, *26* (4), 334−343.
(38) Hermosilla, P.; Krone, M.; Guallar, V.; Vázquez, P.-P.; Vinacua, À.; Ropinski, T. Interactive GPU-Based Generation of Solvent-Excluded Surfaces. *Vis. Comput.* **2017**, *33* (6), 869−881.
(39) Koltun, W. L. Precision Space-Filling Atomic Models. *Biopolym. Orig. Res. Biomol.* **1965**, *3* (6), 665−679.
(40) Lange, A. W.; Herbert, J. M.; Albrecht, B. J.; You, Z.-Q. Intrinsically Smooth Discretisation of Connolly's Solvent-Excluded Molecular Surface. *Mol. Phys.* **2020**, *118* (6), No. e1644384.
(41) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379−IN4.
(42) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. ID-Score: a New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53* (3), 592−600.
(43) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Substituting Random Forest for Multiple Linear Regression Improves Binding Affinity Prediction of Scoring Functions: Cyscore as a Case Study. *BMC Bioinformatics* **2014**, *15* (1), 1−12.
(44) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: the Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inform.* **2015**, *34* (2−3), 115−126.
(45) Li, L.; Li, C.; Zhang, Z.; Alexov, E. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *J. Chem. Theory Comput.* **2013**, *9* (4), 2126−2136.
(46) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54* (6), 1717−1736.
(47) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. Analytical Shape Computation of Macromolecules: I. Molecular Area and Volume Through Alpha Shape. *Proteins Struct. Funct. Bioinforma.* **1998**, *33* (1), 1−17.
(48) Liu, B.; Wang, B.; Zhao, R.; Tong, Y.; Wei, G.-W. ESES: Software for Eulerian Solvent Excluded Surface. *J. Comput. Chem.* **2017**, *38* (7), 446−466.
(49) Liu, T.; Chen, M.; Lu, B. Efficient and Qualified Mesh Generation for Gaussian Molecular Surface Using Adaptive Partition and Piecewise Polynomial Approximation. *SIAM J. Sci. Comput.* **2018**, *40* (2), B507−B527.

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341

(50) Liu, X.; Feng, H.; Wu, J.; Xia, K. Dowker Complex Based Machine Learning (DCML) Models for Protein-Ligand Binding Affinity Prediction. *PLoS Comput. Biol.* **2022**, *18* (4), No. e1009943.
(51) Liu, X.; Wang, X.; Wu, J.; Xia, K. Hypergraph-Based Persistent Cohomology (HPC) for Molecular Representations in Drug Design. *Brief. Bioinform.* **2021**, *22* (5), bbaa411.
(52) Meng, Z.; Xia, K. Persistent Spectral-Based Machine Learning (PerSpect ML) for Protein-Ligand Binding Affinity Prediction. *Sci. Adv.* **2021**, *7* (19), No. eabc5329.
(53) Mu, L.; Xia, K.; Wei, G. Geometric and Electrostatic Modeling Using Molecular Rigidity Functions. *J. Comput. Appl. Math.* **2017**, *313*, 18−37.
(54) Nguyen, D. D.; Wang, B.; Wei, G.-W. Accurate, Robust, and Reliable Calculations of Poisson-Boltzmann Binding Energies. *J. Comput. Chem.* **2017**, *38* (13), 941−948.
(55) Nguyen, D. D.; Wei, G.-W. The Impact of Surface Area, Volume, Curvature, and Lennard-Jones Potential to Solvation Modeling. *J. Comput. Chem.* **2017**, *38* (1), 24−36.
(56) Nguyen, D. D.; Wei, G.-W. AGL-Score: Algebraic Graph Learning Score for Protein-Ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.* **2019**, *59* (7), 3291−3304.
(57) Nguyen, D. D.; Wei, G.-W. DG-GL: Differential Geometry-Based Geometric Learning of Molecular Datasets. *Int. j. numer. method. biomed. eng.* **2019**, *35* (3), No. e3179.
(58) Nguyen, D. D.; Xiao, T.; Wang, M.; Wei, G.-W. Rigidity Strengthening: A Mechanism for Protein–Ligand Binding. *J. Chem. Inf. Model.* **2017**, *57* (7), 1715−1721.
(59) Opron, K.; Xia, K.; Wei, G.-W. Communication: Capturing Protein Multiscale Thermal Fluctuations. *J. Chem. Phys.* **2015**, *142* (21), 211101.
(60) Petrey, D.; Honig, B. GRASP2: Visualization, Surface Properties, and Electrostatics of Macromolecular Structures and Sequences. *Methods Enzymol.* **2003**, *374*, 492−509.
(61) Richards, F. M. Areas, Volumes, Packing, and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6* (1), 151−176.
(62) Rychkov, G.; Petukhov, M. Joint Neighbors Approximation of Macromolecular Solvent Accessible Surface Area. *J. Comput. Chem.* **2007**, *28* (12), 1974−1989.
(63) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. Reduced Surface: an Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38* (3), 305−320.
(64) Smereka, P. The Numerical Approximation of a Delta Function with Application to Level Set Methods. *J. Comput. Phys.* **2006**, *211* (1), 77−90.
(65) Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M. Molecular Docking Using Surface Complementarity. *Proteins Struct. Funct. Bioinforma.* **1996**, *25* (1), 120−129.
(66) Spolar, R. S.; Record, M. T., Jr Coupling of Local Folding to Site-Specific Binding of Proteins to DNA. *Science (80-.).* **1994**, *263* (5148), 777−784.
(67) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning Model for Protein-Ligand Binding Affinity Prediction. *Bioinformatics* **2018**, *1*, 9.
(68) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: the CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895−913.
(69) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical Free Energy Calculations of Ligand-Protein Crystallographic Complexes. I. Knowledge-Based Ligand-Protein Interaction Potentials Applied to the Prediction of Human Immunodeficiency Virus 1 Protease Binding Affinity. *Protein Eng. Des. Sel.* **1995**, *8* (7), 677−691.
(70) Vorobjev, Y. N.; Hermans, J. SIMS: Computation of a Smooth Invariant Molecular Surface. *Biophys. J.* **1997**, *73* (2), 722−732.
(71) Wang, B.; Wei, G. W. Parameter Optimization in Differential Geometry Based Solvation Models. *J. Chem. Phys.* **2015**, *143* (13), 134119.
(72) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38* (3), 169−177.
(73) Wang, L.; Li, L.; Alexov, E. pKa Predictions for Proteins, RNAs, and DNAs with the Gaussian Dielectric Function Using DelPhipKa. *Proteins Struct. Funct. Bioinforma.* **2015**, *83* (12), 2186−2197.
(74) Wee, J.; Xia, K. Forman Persistent Ricci Curvature (FPRC)-Based Machine Learning Models for Protein-Ligand Binding Affinity Prediction. *Brief. Bioinform.* **2021**, *22* (6), bbab136.
(75) Wee, J.; Xia, K. Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein-Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2021**, *61* (4), 1617−1626.
(76) Wei, G.-W. Multiscale, Multiphysics and Multidomain Models I: Basic Theory. *J. Theor. Comput. Chem.* **2013**, *12* (08), 1341006.
(77) Wei, G. W.; Sun, Y.; Zhou, Y. C.; Feig, M. Molecular Multiresolution Surfaces. *arXiv* **2005**, No. ph/0511001.
(78) Wei, G.-W.; Zheng, Q.; Chen, Z.; Xia, K. Variational Multiscale Models for Charge Transport. *siam Rev.* **2012**, *54* (4), 699−754.
(79) Xia, K.; Opron, K.; Wei, G.-W. Multiscale Multiphysics and Multidomain Models-Flexibility and Rigidity. *J. Chem. Phys.* **2013**, *139* (19), 194109.
(80) Xia, K.; Wei, G.-W. A Review of Geometric, Topological and Graph Theory Apparatuses for the Modeling and Analysis of Biomolecular Data. *arXiv* **2016**, No. 01735.
(81) Xia, K.; Zhao, Z.; Wei, G.-W. Multiresolution Persistent Homology for Excessively Large Biomolecular Datasets. *J. Chem. Phys.* **2015**, *143* (13), 134103.
(82) Xu, D.; Zhang, Y. Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform. *PLoS One* **2009**, *4* (12), No. e8140.
(83) Yu, Z.; Holst, M. J.; Cheng, Y.; McCammon, J. A. Feature-Preserving Adaptive Mesh Generation for Molecular Shape Modeling and Simulation. *J. Mol. Graph. Model.* **2008**, *26* (8), 1370−1380.
(84) Zheng, Q.; Chen, D.; Wei, G.-W. Second-Order Poisson-Nernst-Planck Solver for Ion Transport. *J. Comput. Phys.* **2011**, *230* (13), 5239−5262.
(85) Zheng, Q.; Wei, G.-W. Poisson-Boltzmann-Nernst-Planck Model. *J. Chem. Phys.* **2011**, *134* (19), 194101.
(86) Zheng, Z.; Merz, K. M., Jr Ligand Identification Scoring Algorithm (LISA). *J. Chem. Inf. Model.* **2011**, *51* (6), 1296−1306.

https://doi.org/10.1021/acs.jcim.2c00697
J. Chem. Inf. Model. 2022, 62, 4329−4341