Journal of Computer-Aided Molecular Design (2020) 34:131–147
https://doi.org/10.1007/s10822-019-00237-5

MathDL: mathematical deep learning for D3R Grand Challenge 4

Duc Duy Nguyen[1] · Kaifu Gao[1] · Menglun Wang[1] · Guo-Wei Wei[1,2,3]

Received: 10 June 2019 / Accepted: 14 October 2019 / Published online: 16 November 2019

Abstract

We present the performances of our mathematical deep learning (MathDL) models for D3R Grand Challenge 4 (GC4). This challenge involves pose prediction, affinity ranking, and free energy estimation for beta secretase 1 (BACE) as well as affinity ranking and free energy estimation for Cathepsin S (CatS). We have developed advanced mathematics, namely differential geometry, algebraic graph, and/or algebraic topology, to accurately and efficiently encode high dimensional physical/chemical interactions into scalable low-dimensional rotational and translational invariant representations. These representations are integrated with deep learning models, such as generative adversarial networks (GAN) and convolutional neural networks (CNN) for pose prediction and energy evaluation, respectively. Overall, our MathDL models achieved the top place in pose prediction for BACE ligands in Stage 1a. Moreover, our submissions obtained the highest Spearman correlation coefficient on the affinity ranking of 460 CatS compounds, and the smallest centered root mean square error on the free energy set of 39 CatS molecules. It is worthy to mention that our method on docking pose predictions has significantly improved from our previous ones.

Keywords D3R—drug design data resource · Algebraic topology · Graph theory · Differential geometry · Binding affinity · Pose prediction · Docking · Deep learning · Generative adversarial network

Introduction

The drug design data resource (D3R) offers blind community-wide challenges of ligand pose and binding affinity ranking predictions [1–3]. Benchmarks in D3R contests contain high quality structures and reliable binding free energies supplied by experimental groups before the publication. These challenges provide computer-aided drug design (CADD) community a great opportunity to validate, calibrate, and develop drug virtual screening (VS) models. The latest D3R Grand Challenge 4 (GC4), took place from September 4th 2018 to December 4th, 2018. GC4 presented two different protein targets, Cathepsin S (CatS) and beta secretase 1 (BACE), which were generously supplied by Janssen Pharmaceuticals and Novartis, respectively. There were two stages in GC4. The first one has two subchallenges, namely Stage 1a and Stage 1b. In Stage 1a, participants were asked to predict the pose, rank the affinity, and estimate the free energy of BACE ligands. Following Stage 1a, Stage 1b revealed the receptor structures and participants were asked again to predict the crystallographic poses of 20 BACE ligands. There was no affinity calculation in this stage 1b. The second part of GC4 was called Stage 2 which contained the affinity rankings and free energy challenges for both BACE and CatS compounds. In this last stage, participants were able to take advantage of experimental structures of BACE complexes released right after stage 1b.

A successful VS model requires a reliable ligand conformation generation and highly accurate scoring function to predict binding affinities. There are several state-of-the-art software packages to take care of the first component of VS, for example, Autodock Vina [4], GOLD [5], GLIDE [6], ICM [7], etc. Unfortunately, one may fail dramatically to achieve decent poses if blindly using these software

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-019-00237-5) contains supplementary material, which is available to authorized users.

✉ Guo-Wei Wei
weig@msu.edu

[1] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

[2] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

[3] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA

programs. The pose prediction results in Grand Challenge 3 (GC3) clearly demonstrated this issue [3]. The second component of VS relates to the development of scoring function (SF) for binding affinity predictions. Basically, one can classify SF methods into four different types, namely force-field-based SF, knowledge-based SF, empirical-based SF, and machine learning-based SF [8]. The force-field-based SFs commonly emphasize van der Waals (vdW) interactions, electrostatic energy, hydrogen bonding descriptions, solvation effects, and so on. The well-known SFs for this category are COMBINE [9], MedusaScore [10], to name only a few. Typical examples of knowledge-based SFs are [11], DrugScore [12], KECSA [13], and IT-Score [14], which utilize protein–ligand pairwise statistical potentials in an additive manner to predict binding affinities. One can regard the empirical-based SFs as simple machine learning-based SFs since these SFs employ linear regression schemes to construct predictive models using various physical features, for instance vdW interactions, Lennard–Jones potentials, hydrogen bonds, electrostatics, solvation, and torsion information, etc. PLP [15], ChemScore [16], and X-Score [17] are the well-known representatives in this category. The last type of binding affinity SFs is machine learning-based approaches which have recently arise as the most advanced technique in CADD. One of the pioneer work on this SF category is RF-Score [18] based on the Random Forest (RF) algorithm [19] and their features as the numbers of atom pairwise contacts. Thanks to the nonlinear representation of the sophisticated machine learning frameworks, machine learning-based SFs can characterize the non-additive contributions from functional group interactions in the binding affinity calculations [20–26].

The availability of massive biological datasets, along with the accessibility to high-performance computing cluster (HPCC), has made machine learning-based models an emerging technology in biomolecular data analysis and prediction. However, the accuracy of machine learning-based SFs highly depends on whether their features are able to capture the physical and chemical information in protein–ligand interactions. Moreover, the direct use of three dimensional (3D) biomolecular structures in the deep learning network is immensely expensive. This hindrance mainly causes by the hefty number of degrees of freedom in the 3D macromolecular representations and the number of atoms varying among different structures. Therefore, there is a pressing need to develop innovative representations of protein–ligand complexes for machine learning methods.

Mathematical deep learning (MathDL) encompasses a family of scalable low-dimensional rotational and translational invariant mathematical representations integrated with advanced machine learning, including deep learning algorithms [27]. Its hypothesis is that the intrinsic physics of macromolecular interactions lie in low-dimensional manifolds. Based on such hypothesis, we have developed a number of mathematical tools originated from geometry, topology, graph theory, combinatorics, and analysis to simplify macromolecular complexity and reduce their dimensionality. For example, differential geometry provides a high-level abstraction of macromolecular complexes [28]. In molecular biophysics, differential geometry-based framework has shown its efficiency in modeling solvation-free energies [29, 30] and ion channel transport [31–35]. However, in those applications, differential geometry information is largely restricted to the separation of solvent and solute domains in facilitating the Poisson–Boltzmann model or the Poisson–Nernst–Planck model. In geometric modeling, differential geometry has been utilized for the qualitative analysis of biomolecule properties [36, 37]. Also, potential protein–ligand binding sites can be recognized via concave and convex regions of molecular surfaces indicated by minimum and/or maximum curvatures [37, 38]. Most recently, the roles of different kinds of curvature in solvation free energy models have been investigated [39]. However, the efficiency of the aforementioned differential geometry models is limited due to neglecting of atomic level information. Element interactive manifolds (EIM) were proposed to address this problem in differential geometry-based geometric learning (DG-GL) [25]. These EIMs successfully encode the pivotal physical, chemical, and biological information stored in high-dimensional data into low-dimensional manifolds, rendering a powerful approach for predicting solvation free energy, drug toxicity, and protein–ligand binding affinity [25].

Another low-dimensional mathematical approach is the topological representation of biomolecular structures. In topological data analysis, one can capture the connectivity of macromolecules or molecular components. Topological invariants, such as independent components, rings, cavities, and higher dimension faces in terms of Betti numbers help to characterize the conformation change upon the protein–ligand binding process, the folding and unfolding of proteins, and the opening or closing of ion channels [40]. The traditional topological descriptors, unfortunately, cannot discriminate the geometric difference among various macromolecular structures. Persistent homology (PH), a new branch of algebraic topology, utilizes a filtration parameter to generate a family of topological spaces and associated invariants, which contain richer geometric information [41, 42]. PH has been applied to computational biology [43–45]. However, these applications were mostly limited to qualitative analysis. Recently, we have devised PH for the quantitative analysis of protein folding energy, protein flexibility [46], ill-posed inverse problems of cryo-EM structures [47], predictive models of curvature energies of fullerene isomers [48, 49], and protein pocket detection [50]. In 2015, we introduced one of the first combinations

of PH descriptors and machine learning algorithms [51]. Since then, the integration of PH and machine learning has become a very popular approach in topological data analysis. Nonetheless, this approach is not good enough for biomolecular systems. It turns out that PH neglects chemical and biological information in its topological simplification of geometric complexity. Element-specific PH was introduced to retain chemical and biological information [22]. The integration of element-specific PH and machine learning algorithms has found great success in the predictions of protein folding free energy changes upon mutation [52], binding affinity [22–24], drug toxicity [53], partition coefficient, and aqueous solubility [54]. It has been employed for the classification of active ligands and decoys [24]. All of these new topological models outperformed other state-of-the-art methods on various common benchmarks.

Similarly to topology, graph theory also accentuates the connectivity between vertices to define graph edges. There are two major types of graphs: geometric graphs and algebraic graphs. Geometric graphs concern the pairwise connectivity between graph nodes and represent it in terms of "topological index" [55, 56], graph centrality [57–59], and contact map [60, 61]. The algebraic graph theory expresses the connectivity via eigenvalues, particularly, the second-smallest eigenvalue of the Laplacian matrix, known as Fiedler value, which is often used to analyze the stability of dynamical systems [62]. Graph theory has been widely used in many interdisciplinary studies. In biophysics, it is employed to model protein flexibility and long-time dynamics in normal mode analysis (NMA) [63–66] and elastic network model (ENM)[60, 67–72]. Since graph theory offers a nature representation of molecular structure, it is a common approach for analyzing chemical datasets [56, 73–77] and biomolecular datasets [60, 78–83]. Although there was much effort in constructing various graph representations in the past, graph based quantitative models are often less accurate than other competitive models in the analysis and prediction of biomolecular properties from massive and diverse datasets. Indeed, in the protein stability changes upon mutation analysis, the other models [23, 52, 84] are more accurate than the graph-based approach [85]. In addition, the graph theory based Gaussian network model (GNM) is not competitive in protein B-factor predcitions [86]. One of the main reasons is that there is no systematic representation of interactions among different chemical element types in a molecular structure. Additionally, many graph approaches do not describe non-covalent interactions. To overcome these limitations, we have proposed novel multiscale weighted colored subgraphs in both geometric graph and algebraic graph schemes to achieve the state-of-the-art performances in the predictions of protein B-factor [87], protein–ligand binding affinity [21, 26], docking [26], and virtual screening [26].

Our MathDL models using graph theory and algebraic topology were employed in the D3R Grand Challenges since GC2 and have obtained many encouraging results. Specifically, our prediction of the binding free energy set in Stage 2 was ranked the best in GC2 in our first participation of D3R competitions [27]. In our second participation, i.e. GC3, our submissions achieved the top places in 10 out of 26 official contests [27]. These achievements have confirmed the predictive power and efficiency of our MathDL models in drug design and discovery. However, there were still some shortcomings existing in our previous approaches mostly concerning the pose generation performance and ability to rank affinities of compounds with diverse chemical structures.

In the current D3R challenge, i.e. GC4, we have brought in two new technological aspects in our approach. First, we have further developed powerful differential geometry and algebraic graph-based MathDL models to assist our algebraic topology based methods. Additionally, we have extended our MathDL approach with more advanced deep learning architectures like generative adversarial networks (GAN) [88]. We have achieved very promising results with top places in pose prediction, affinity ranking and free energy estimation. The rest of this paper is devoted to more detailed discussions of our methodologies and their performances in D3R GC4.

Methods

We describe the mathematical methods underpinning our MathDL models in this sections.

Differential geometry representation

Multiscale discrete-to-continuum mapping

Given a molecule having N atoms. Denote $\mathbf{r}_i$ and q_j, $i = 1 \cdots N$, respectively, an atomic coordinate and a partial charge of the jth atom. A discrete-to-continuum mapping [89–91] represents the unnormalized molecular density at an arbitrary point $\mathbf{r} \in \mathbb{R}^3$ as follows

$$\rho(\mathbf{r}, \{\eta_k\}, \{w_k\}) = \sum_{j=1}^{N} w_j \Phi\big(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j\big), \tag{1}$$

where $\|\mathbf{r} - \mathbf{r}_j\|$ is the Euclidean distance of the point $\mathbf{r}$ and the jth atom in a given molecule. If all w_j are set to 1, $\rho(\mathbf{r}, \{\eta_k\}, \{w_k\})$ indicates a molecular density, whereas $\rho(\mathbf{r}, \{\eta_k\}, \{w_k\})$ serves as molecular charge density with $w_j = q_j$ for all j. In the present work, we utilize Autodock Tools (http://autodock.scripps.edu/resources/adt/index_html) to assign the Gasteiger charges for small molecules

and macromolecules. Additionally, η_j are characteristic distances and Φ is a monotonically decreasing kernel featuring the similarity between two 3D data points. To ensure the existence of the geometric representations such as curvatures, Φ is chosen to be monotonically decreasing C^2 function satisfying the following conditions

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

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

It is noted that radial basis functions meet admissibility conditions (2) and (3). Commonly used correlation kernels are generalized exponential functions

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

and generalized Lorentz functions

$$\Phi\big(\|\mathbf{r}-\mathbf{r}_j\|;\eta_j\big)=\frac{1}{1+\left(\|\mathbf{r}-\mathbf{r}_j\|/\eta_j\right)^{\nu}}, \quad \nu>0. \tag{5}$$

Moreover, one can use correlation kernels to model the electrostatic interaction between two charged articles as the following

$$\Phi(\|\mathbf{r}_i-\mathbf{r}_j\|,q_i,q_j;c)=\frac{1}{1+e^{-cq_iq_j/\|\mathbf{r}_i-\mathbf{r}_j\|}}, \tag{6}$$

where, q_i and q_j are the partial charges of two atoms, and c is a nonzero tunable parameter. It is noted that Φ described in Eq. (6) does not follow the admissible conditions (2) and (3). It is, therefore, only utilized to generate electrostatic persistent homology. All the Φs discussed in the current work were determined by one of Eqs. (4)–(5). Here, Φ takes 3D coordinates and kernel parameters as the input variables and maps them to a real number: $\mathbb{R}^3\to\mathbb{R}$. Therefore, Φ values totally depend on atom coordinates or grid point positions and are rotationally and translationally invariant.

It is expected that C^2 delta sequences of the positive type discussed in an earlier work [92] can function well for the correlation kernel purposes. To obtain multiscale discrete-to-continuum mapping, one can employ more than one set of scale parameters. In the current work, the aforementioned mapping was applied to protein–ligand complexes.

Element interactive densities

In order for differential geometry (DG) representations to effectively capture the crucial physical and biological information of large and diverse biomolecular datasets, we must employ DG to feature non-covalent intramolecular molecular interactions in a molecule and intermolecular interactions in molecular complexes, such as protein–protein and protein–ligand.

Additionally, the accuracy of the DG representations can be upgraded by element-level descriptions which result in scalable low-dimension manifold representations of high dimensional structures. For instance, to describe the pairwise interactions between protein and ligand, we consider frequently occurring element types in proteins and ligands. Particularly, the commonly occurring element types in proteins are C, N, O, S and commonly occurring element types in ligands are H, C, N, O, S, P, F, Cl, Br, I. That gives rise to 40 element pairwise groups. We do not include hydrogen in protein element types since H is usually absent from most datasets in the Protein Data Bank (PDB). Note that during our validation process, the pairwise interactions between different atom types did not enhance the overall performance of our models (this may be due to the limited data size.). Thus, we only carried out the element-specific interactions for the sake of simplicity.

Based on a statistical analysis, the frequently occurring element types in the biomolecular dataset are denoted as $\mathcal{C}=$ H, C, N, O, S, P, F, Cl, $\cdots$. For convenience, $\mathcal{C}_k$ represents the kth element in the set $\mathcal{C}$. For example, $\mathcal{C}_5=$ S. An ith atom in a given molecule is associated with its coordinate $\mathbf{r}_i$, element type α_i, and partial charge q_i. The non-covalent interactions between atoms of element type $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ are assumed to be described by the correlation kernel Φ

$$\begin{aligned}&\{\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta_{kk'})|\alpha_i=\mathcal{C}_k,\alpha_j=\mathcal{C}_{k'};\\&i,j=1,2,\ldots,N;||\mathbf{r}_i-\mathbf{r}_j||>r_i+r_j+\sigma\},\end{aligned} \tag{7}$$

where r_i and r_j are the atomic radii of ith and jth atoms, respectively and σ is the mean value of the standard deviations of r_i and r_j in the interested dataset. The covalent interactions are excluded due to the constraint $||\mathbf{r}_i-\mathbf{r}_j||>r_i+r_j+\sigma$. In addition, $\eta_{kk'}$ is a characteristic distance between the atoms, which depends only on their element types.

To construct the element interactive densities, we define *atomic-radius-parametrized* van der Waals domain of all atoms of kth element type as [25]

$$D_k:=\cup_{\mathbf{r}_i,\alpha_i=\mathcal{C}_k}B(\mathbf{r}_i,r_k), \tag{8}$$

in which $B(\mathbf{r}_i,r_i)$ is a ball with a center $\mathbf{r}_i$ and a radius r_i, and r_k is the atomic radius of the kth element type. Thus, D_k depends on atom coordinate $\mathbf{r_i}$ and its atomic radius. Note that, D_k does not define any vdW interactions but a domain to construct the surface density. The element interactive density between domain D_k and all atoms of k'th ($k\neq k'$) element type is given by

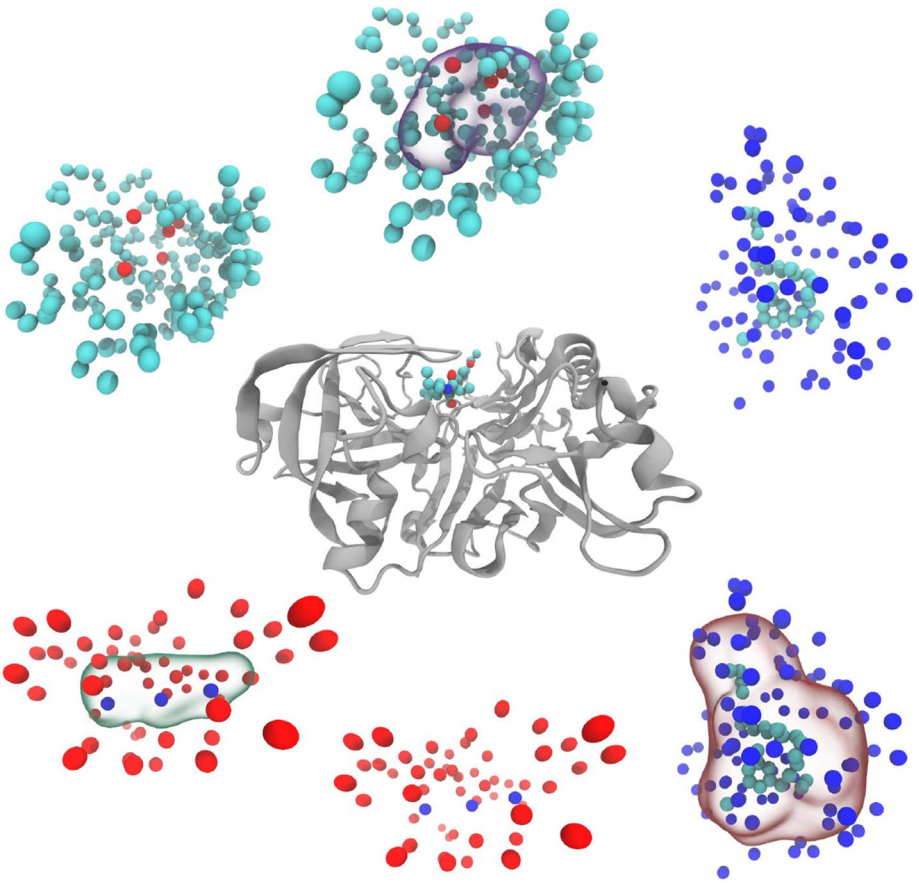

Fig. 1 Illustration of some element-specific selections and corresponding element interactive manifolds obtained at a given level set of the element interactive density. Each sphere illustrates the atomic positions. Cyan, red, and blue colors represent carbon, oxygen, and nitrogen, respectively. The transparent surfaces are the isosurface extracted from volume data represented in Eq. (8)

$$\rho_{kk'}(\mathbf{r},\eta_{kk'}) = \sum_{\substack{j \\ \alpha_j = \mathcal{C}_{k'} \\ ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma, \forall \alpha_i \in \mathcal{C}_k}} w_j \Phi(||\mathbf{r} - \mathbf{r}_j||;\eta_{kk'}), \quad \mathbf{r} \in D_k. \tag{9}$$

When $k' = k$, the element interactive density ρ_{kk} is now induced only by van der Waals domain D_k. In this case, we exclude the covalent interactions based on the position of the density input. Assuming $\mathbf{r} \in D_k^i$, with $D_k^i = B(\mathbf{r}_i, r_i)$, $\alpha_i = \mathcal{C}_k$, the element interactive density is then formulated by

$$\rho_{kk}(\mathbf{r},\eta_{kk}) = \sum_{\substack{j \\ \alpha_j = \mathcal{C}_k \\ ||\mathbf{r}_i - \mathbf{r}_j|| > 2r_j + \sigma}} w_j \Phi(||\mathbf{r} - \mathbf{r}_j||;\eta_{kk}). \tag{10}$$

For the sake of simplicity, we chose $w_j = 1$ for all cases. Since element interactive density is obtained by the addition of correlation kernels, it belongs to C^2 on the closed domain of D_k. We construct element interactive manifolds by restricting the set of points at a given level set of the density as shown in Fig. 1 .

Element interactive curvatures

Given an element interactive density $\rho(\mathbf{r})$, one can calculate the Gaussian curvature (K), the mean curvature (H), the minimum curvature ($\kappa_{\min}$), and the maximum curvature ($\kappa_{\max}$) for the resulting manifold as the following [37, 93]:

$$\begin{aligned} K = \frac{1}{g^2}\big[&2\rho_x\rho_y\rho_{xz}\rho_{yz} + 2\rho_x\rho_z\rho_{xy}\rho_{yz} + 2\rho_y\rho_z\rho_{xy}\rho_{xz} \\ &- 2\rho_x\rho_z\rho_{xz}\rho_{yy} - 2\rho_y\rho_z\rho_{xx}\rho_{yz} - 2\rho_x\rho_y\rho_{xy}\rho_{zz} \\ &+ \rho_z^2\rho_{xx}\rho_{yy} + \rho_x^2\rho_{yy}\rho_{zz} + \rho_y^2\rho_{xx}\rho_{zz} \\ &- \rho_x^2\rho_{yz}^2 - \rho_y^2\rho_{xz}^2 - \rho_z^2\rho_{xy}^2\big], \end{aligned} \tag{11}$$

$$H = \frac{1}{2g^{\frac{3}{2}}}[2\rho_x\rho_y\rho_{xy} + 2\rho_x\rho_z\rho_{xz} + 2\rho_y\rho_z\rho_{yz} - (\rho_y^2 + \rho_z^2)\rho_{xx} - (\rho_x^2 + \rho_z^2)\rho_{yy} - (\rho_x^2 + \rho_y^2)\rho_{zz}], \quad (12)$$

$$\kappa_{\min} = H - \sqrt{H^2 - K}, \quad (13)$$

$$\kappa_{\max} = H + \sqrt{H^2 - K}, \quad (14)$$

where $g = \rho_x^2 + \rho_y^2 + \rho_z^2$.

To construct unified curvature quantities for various biomolecular structures, we study the element interactive curvatures (EIC) at the atomic center and formulate them as [25]

$$K_{kk'}^{\mathrm{EI}}(\eta_{kk'}) = \sum_i K_{kk'}(\mathbf{r}_i, \eta_{kk'}), \quad \mathbf{r}_i \in D_k; k \neq k' \quad (15)$$

and

$$K_{kk}^{\mathrm{EI}}(\eta_{kk}) = \sum_i K_{kk}(\mathbf{r}_i, \eta_{kk}), \quad \mathbf{r}_i \in D_k^i, D_k^i \subset D_k. \quad (16)$$

Eqs. (15) and (16) are for the element interactive Gaussian curvature (EIGC), are applied to protein–ligand complexes in the current work. Thus, the atomic centers in Eqs. (15) and (16) can be either from ligand atoms or protein atoms. In a same manner, one can define $H_{kk'}^{\mathrm{EI}}(\eta_{kk'})$, $\kappa_{kk',\mathrm{min}}^{\mathrm{EI}}(\eta_{kk'})$ and $\kappa_{kk',max}^{\mathrm{EI}}(\eta_{kk'})$ for the element interactive mean curvature, element interactive minimum curvature, and element interactive maximum curvature, respectively.

It is worth noting that, the expressions of the curvatures defined in (11)–(14) are in the analytical forms. Thus, the EIC formulations are free from numerical error and totally preserve the reference geometric information of the molecules.

Multiscale weighted colored geometric subgraphs

For a given molecular datasets, we denote $\mathcal{C}$ a set consisting of the most frequently appearing element types. For a molecule of interest, we define a graph with the following vertices

$$\mathcal{V} = \{(\mathbf{r}_j, \alpha_j) | \mathbf{r}_j \in \mathbb{R}^3; \alpha_j \in \mathcal{C}; j = 1, 2, \ldots, N\}, \quad (17)$$

where N is the number of atoms, $\mathbf{r}_j$ and α_j are, respectively, coordinates and element type of the jth atom. Similarly to the discussion in the differential geometry representation section, we only consider non-covalent interactions represented by correlation kernels

$$\mathcal{E}_{kk'} = \{\Phi(||\mathbf{r}_i - \mathbf{r}_j||; \eta_{kk'}) | \alpha_i = \mathcal{C}_k, \alpha_j = \mathcal{C}_{k'}; i, j = 1, 2, \ldots, N; ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma\}, \quad (18)$$

all the notations in Eq. (18) are adopted from "Differential geometry representation" section. In which, Φ refers to the edge weight which represents the potential interaction between two nodes forming that edge. We now form weighted colored subgraphs $G(\mathcal{V}, \mathcal{E}_{kk'})$ to describe pairwise interactions in a given molecule. To unify the geometric graph-based descriptors for a diversity dataset, we construct multiscale weighted colored subgraph rigidity between kth element type $\mathcal{C}_k$ and k'th element type $\mathcal{C}_{k'}$ via a graph centrality type of scheme

$$\mathrm{RI}^G(\eta_{kk'}) = \sum_i \mu_i^G(\eta_{kk'}) = \sum_{\substack{i \\ \alpha_i = \mathcal{C}_k}} \sum_{\substack{j \\ \alpha_j = \mathcal{C}_{k'} \\ ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma}} \Phi(||\mathbf{r}_i - \mathbf{r}_j||; \eta_{kk'}). \quad (19)$$

The proposed subgraph rigidity index $\mathrm{RI}^G(\eta_{kk'})$ in Eq. (19) is the aggregation of the collective subgraph centrality $\mu_i^G(\eta_{kk'})$ which used in our previous B-factor prediction model [87]. That formulation represents a coarse-grained description at the element-level capturing important physical and biology information in a molecule or biomolecule such as van der Waals interactions, hydrogen bonds, electrostatics, etc. This description is scalable, i.e., independent of the size of an individual protein–ligand complex. In fact, when describing protein–ligand interactions, the labeled subgraph $G(\mathcal{V}, \mathcal{E}_{kk'})$ gives rise to a bipartite graph with its edges connecting protein atoms to ligand atoms. The positive and negative eigenvalues of the adjacency matrix of a bipartite graph are reflective, which enables us to select only positive or negative eigenvalues in machine learning. Moreover, Eq. (19) generalized our previous binding affinity prediction model [21] and was utilized for the D3R Grand Challenge 3 [27].

Multiscale weighted colored algebraic subgraphs

Still based on multiscale weighted colored subgraphs as defined in "Multiscale weighted colored geometric subgraphs" section, we have recently developed a novel algebraic graph approach or spectral graph formulation to describe molecules, biomolecules and their interactions at atomic levels [25]. We here utilize the Laplacian matrix and adjacency matrix to represent the interactions between nodes in a given subgraph.

Based on a weighted colored subgraph $G(\mathcal{V}, \mathcal{E}_{kk'})$, we define the weighted colored Laplacian matrix $L_{ij}(\eta_{kk'})$ as the following

$$L_{ij}(\eta_{kk'}) = \begin{cases} -\Phi(||\mathbf{r}_i - \mathbf{r}_j||; \eta_{kk'}) & \text{if} \quad i \neq j, \alpha_i = \mathcal{C}_k, \alpha_j = \mathcal{C}_{k'} \\ & \text{and} \quad ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma; \\ -\sum_j L_{ij} & \text{if} \quad i = j. \end{cases} \quad (20)$$

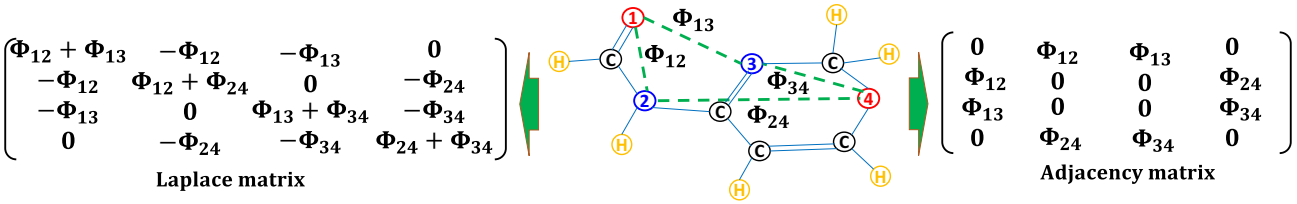

Fig. 2 Illustration of weight colored subgraphs G_{NO} including its Laplacian matrix (Left), and adjacency matrix (Right) deduced from molecule graph ($C_5H_6N_2O_2$) (Middle). Atoms 1 and 4 are oxygen, while atoms 2 and 3 are nitrogen. Graph edges, Φ_{ij}, are in the green-dashed lines representing the noncovalent bonds. In addition, one can get nine other nontrivial subgraph for this molecule, namely $G_{\mathrm{CC}}, G_{\mathrm{CN}}, G_{\mathrm{CO}}, G_{\mathrm{CH}}, G_{\mathrm{NN}}, G_{\mathrm{NH}}, G_{\mathrm{OO}}, G_{\mathrm{OH}}$, and G_{HH}

Due to the symmetric, diagonally dominant and positive-semidefinite, all eigenvalues of the Laplacian matrix $L_{ij}(\eta_{kk'})$ are nonnegative. Moreover, the smallest eigenvalues are zero. It is worth noting that the number of zero eigenvalues can equally referred to the zero-dimensional topological invariant which implies the number of the connected components in the graph. If a graph is connected, there exists one non-zero eigenvalue. Moreover, the smallest non-zero ones is called as Fiedler value representing algebraic connectivity. It is interesting to see that one can reconstruct the geometric graph rigidity via the following formulation

$$\mathrm{RI}^{G}(\eta_{kk'}) = \mathrm{Tr}L(\eta_{kk'}),$$

In addition, we can form the adjacency matrix A_{ij} for the aforementioned subgraph $G(\mathcal{V}, \mathcal{E}_{kk'})$ by

$$A_{ij}(\eta_{kk'}) = \begin{cases} \Phi(||\mathbf{r}_i - \mathbf{r}_j||;\eta_{kk'}) & \text{if } i \neq j, \alpha_i = \mathcal{C}_k, \alpha_j = \mathcal{C}_{k'} \\ & \text{and } ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma; \\ 0 & \text{if } i = j. \end{cases} \tag{21}$$

Clearly, adjacency matrix $A(\eta_{kk'})$ is a symmetric non-negative matrix. As a result, its spectrum is real. The Laplacian and adjacency matrices for subgraph including only oxygen and nitrogen atoms in molecule $C_5H_6N_2O_2$ are depicted in Fig. 2. Note that for different molecules, one can expect to have different graph structures. We only utilized one unique 3D representation for each ligand; thus there was only one single graph structure to represent one corresponding compound.

In general, the element-level information decoded from the Laplacian matrix and the adjacency matrix is quite similar despite of the different behaviors among their eigenvalues and eigenvectors. Specifically, the correlation between the adjacency matrix and the Laplacian matrix can be found in the Perron-Frobenius theorem via the following inequalities

$$\min_i \sum_j A_{ij} \leq \rho(A) \leq \max_i \sum_j A_{ij}. \tag{22}$$

In other words, one can state that the spectral radius $\rho(A)$ of the adjacency matrix A is bounded by diagonal element interval of the corresponding Laplacian matrix L.

In the algebraic approach, we are interested in describing the interactions between elements in the subgraph by the eigenvalues of its matrix. Thus, we design the weighted colored Laplacian matrix based descriptor at the element-level by

$$\mathrm{RI}^{L}(\eta_{kk'}) = \sum_i \mu_i^L(\eta_{kk'}), \tag{23}$$

and the weighted colored adjacency matrix based descriptor is proposed in a similar manner. Note that GNM [60] is a special case of the proposed Laplacian matrix $\mu_i^L(\eta_{kk'})$. Thus, one can utilize its spectrum $\mu_i^L(\eta_{kk'})$ for the protein B-factor prediction. To enrich the algebraic graph-based description information, we consider the statistics of the eigenvalues such as sum, mean, maximum, minimum and standard deviation.

Algebraic topology-based molecular signature

By employing powerful topological analysis, one can construct sophisticated topological spaces to capture the key interactions at the element level of an interested molecule or biomolecule. These physical and chemical information are encoded in different dimensional space under the topological invariant features, so-called Betti numbers. Upon the topological information, the rich and systematic descriptions are formulated and integrated with advanced machine learning framework.

Persistent homology

In the geometric point of view, the collection of points, edges, triangles, and higher-dimension representations

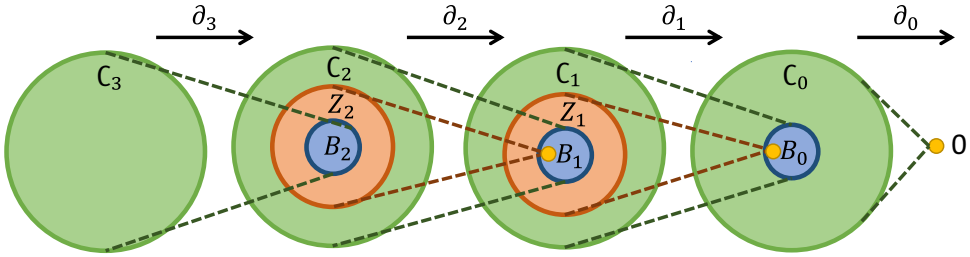

Fig. 3 Illustration of boundary operators, chain, cycle, and boundary groups in $\mathbb{R}^3$. Yellow circles are empty sets

form topological spaces. The general form of a triangle or a tetrahedron is called a simplex. Mathematically, a set of $(k+1)$ affinely independent points in $\mathbb{R}^n$ with $n \geq k$ gives rise to a simplex. To further characterize the topological spaces, face is introduced as a convex hull of a subset of points defining a simplex. In addition, a finite collection of simplices defines a simplicial complex X provided that two requirements are met. First, the faces of any simplex in X are also in X. Second, the intersection of two simplices σ_1 and σ_2 in X are either empty or a face of both σ_1 and σ_2. In a given simplicial complex X, a k-chain c is a formal sum of all the k-simplices in X which is defined as $c = \sum_i a_i \sigma_i$. Here, a_i is an integer coefficient chosen in a finite field $\mathbb{Z}_p$ with a prime p. With the additional operator on the coefficients of in the k-chain, one can form a group of k-chain denoted $\mathcal{C}_k(X)$. The boundary operator on simplices is defined as

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \cdots, \hat{v}_i, \cdots, v_k], \tag{24}$$

where $v_0, \cdots, v_k$ are vertices of the k-simplex σ and $[v_0, \cdots, \hat{v}_i, \cdots, v_k]$ means the codim-1 face of σ be omitting the vertex v_i. The boundary operator $\partial_k(\sigma)$ is homeomorphisms going from $\mathcal{C}_k(X)$ to $\mathcal{C}_{k-1}(X)$ with an important property $\partial_k \circ \partial_{k+1} = 0$. Therefore, one can form the following chain complex

$$\cdots \xrightarrow{\partial_{i+1}} \mathcal{C}_i(X) \xrightarrow{\partial_i} \mathcal{C}_{i-1}(X) \xrightarrow{\partial_{i-1}} \cdots \xrightarrow{\partial_2} \mathcal{C}_1(X) \xrightarrow{\partial_1} \mathcal{C}_0(X) \xrightarrow{\partial_0} 0. \tag{25}$$

In algebraic topology, homology is used to distinguish two shapes by detecting their holes. To define kth homology group, we consider the image of the boundary operator ∂_{k+1} denoted $\mathcal{B}_k(X) = \mathrm{Im}(\partial_{k+1})$ and the kernel of ∂_k denoted $\mathcal{Z}_k(X) = \mathrm{Ker}(\partial_k)$ which are all illustrated in Fig. 3. Then, the quotient group between the aforementioned kernel and image gives rise to the kth homology group

$$\mathcal{H}_k(X) = \mathcal{Z}_k(X)/\mathcal{B}_k(X). \tag{26}$$

The described above homology group is applied for a fixed topological space. To accommodate the objects related to multiscale, we can construct a sequence of subspaces of topological space. Such sequence is called a filtration $\emptyset = X_0 \subseteq X_1 \subseteq \cdots \subseteq X_{m-1} \subseteq X_m = X$ which naturally induces a series of homology groups of different dimensions connected by homomorphisms

$$I_k^{t,s} : \mathcal{H}_k(X_t) \to \mathcal{H}_k(X_s), \text{ with } 0 \leq t \leq s \leq m. \tag{27}$$

The images of these homomorphisms are called kth persistent homology groups, and ranks of these groups define kth persistent Betti numbers which are used to recognize topological spaces via number of k-dimensional holes. In the physical interpretation, Betti-0 counts the number of independent components, Betti-1 illustrates number of rings, and Betti-2 encodes the cavities.

Topological description of molecular systems

We carry out persistent homology on labels subgraph $G(\mathcal{V}, \mathcal{E}_{kk'})$ defined in the previous sections to describe molecular properties. The resulting topological formulation is called element specific persistent homology [22, 52].

There are two common types of filtration, namely Vietoris–Rips complex and alpha complex [94]. The Vietoris–Rips complex, a distance-based filtration, is used to directly address the protein–ligand interactions. For a set of atoms in subgraph $G(\mathcal{V}, \mathcal{E}_{kk'})$, the subcomplex associated to ϵ is defined as

$$X_{\mathrm{Rips}}(\epsilon) = \{\sigma \in X | \sigma = [v_0, \cdots, v_k],\ d(v_i, v_j) \leq 2\epsilon \text{ for } 0 \leq i, j \leq k\}, \tag{28}$$

where X is the collection of all possible simplices, d is the distance between two atoms. To capture a complex protein geometry, one can utilize alpha complex. The alpha filtration is built upon the non-empty intersection between a k-simplex and a $(k+1)$ Voronoi cells. In general, in the alpha filtration, the subcomplex associated to ϵ is defined as

$$X_{\mathrm{alpha}}(\epsilon) = \{\sigma \in X | \sigma = [v_0, \cdots, v_k],\ \cap_i \big(V(v_i) \cap B_\epsilon(v_i)\big) \neq \emptyset\}, \tag{29}$$

where $V(v_i)$ is the Voronoi cell of v_i and $B_\epsilon(v_i)$ is an ϵ ball centered at v_i. For the details of building an alpha filtration, we refer the interested readers to our published work [46].

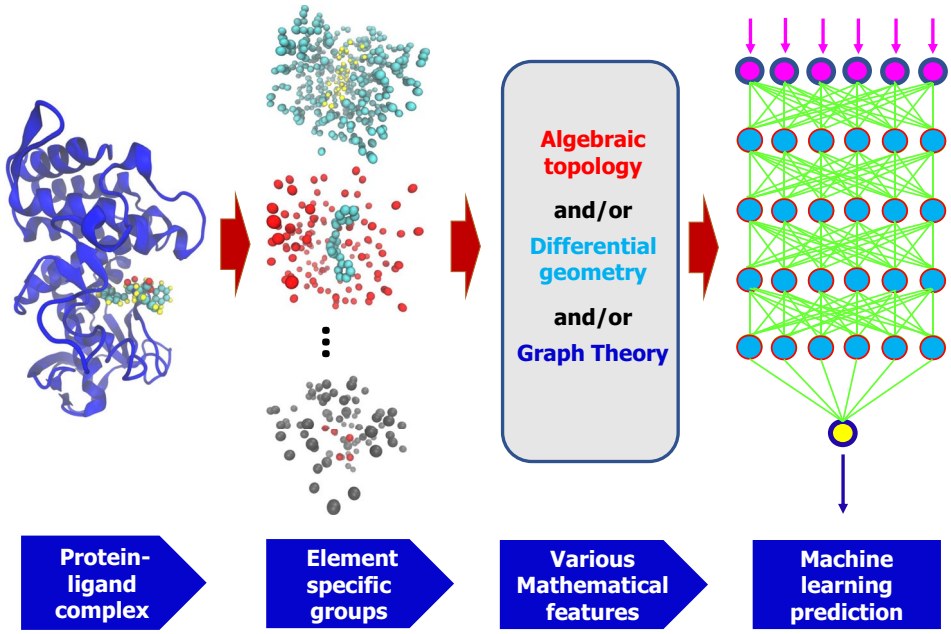

Fig. 4 A framework of MathDL energy prediction model which integrates advanced mathematical representations with sophisticated CNN architectures

Similarly to multiscale weight colored subgraphs in algebraic graph theory approaches, the element specific persistent homology has been shown to capture crucial physical interactions by tweaking the distance functions used in the filtration [22, 52]. Indeed, the hydrophobic effects can be described by considering the persistent homology computation on the collection of all carbon atoms. To describe the hydrophilic behavior of the molecular system, the element specific persistent homology is carried out only for nitrogen and oxygen atoms. In addition, an appropriate distance function selection can characterize the covalent bonds and non-covalent interactions in small molecules [24].

There are several ways to incorporate barcodes generated by persistent homology into machine learning models. One can use the Wasserstein metric to measure the similarities between two molecules' barcodes. As a result, the distance-based machine learning approaches such as nearest neighbors and kernel methods can be exploited [24]. To make use of advanced machine learning algorithms such as the ensemble of trees and deep neural networks, we vectorize persistent homology barcodes by discretizing them into bins and taking into account of the persistence, birth and death incidents in each bin. Furthermore, the statistics of element-specific persistent homology barcodes are included in fixed length features [24]. In the convolutional neural networks, such featurization of barcodes is represented in 1-dimensional and 2-dimensional like images [23, 24].

MathDL energy prediction models

We integrate the mathematical features with deep learning networks to form a powerful predictive model. The convolutional neural network (CNN) is a well-known algorithm with much success in image recognition and computer vision analysis. Essentially, CNN is a regularized version of the artificial neural network consisting of many convolutional layers, followed by several fully connected layers. To enhance the learning process, dropout techniques have been exploited in network layers [95]. The neural networks we use are classified as the feed-forward network where all the information in the current layer is linearly combined and then nonlinearized via an activation function before sending out to the next layer. The predictive power of the CNN models relies on the characterization of the local interactions in the spatial dimension under the discrete convolution operator. The choice of features inputs in the CNN networks gives rise to variants of binding energy predictive models. Figure 4 depicts MathDL energy prediction models and their network architectures are described in Fig. S1 in the Supporting Information. In the D3R GC4, we utilized two different models. In the first approach, the combination of algebraic topology and differential geometry features were employed in the network, we named this model BP1. In the second approach, algebraic topology, differential geometry, and algebraic graph representations were mixed to lead to another binding energy prediction model named BP2. The details of feature generation procedure of the algebraic topology, differential geometry, and algebraic graph models can be found in our earlier work [24–26].

MathDeep docking models

We here present an innovative pose generation scheme, denoted MGAN, using advanced mathematical

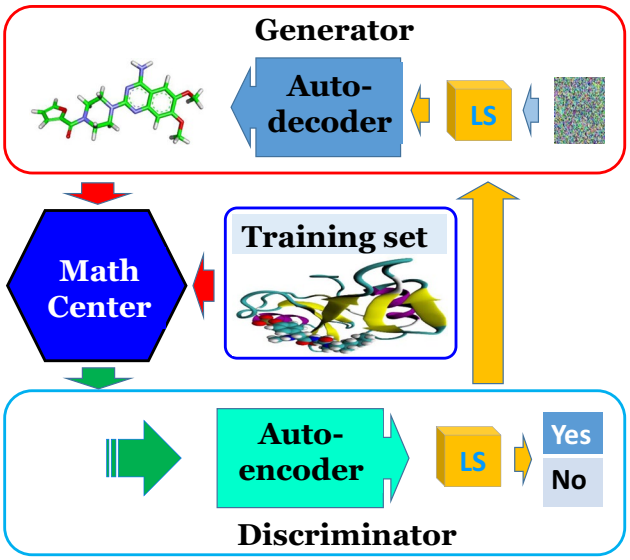

Fig. 5 Illustration of our docking approach using mathematical representations integrated with GAN architectures. The generator contains an autodecoder, a latent space (LS), and a noise source. The discriminator consists of an autoencoder and latent space. The Math center encodes 3D structures into low-dimensional mathematical representations using algebraic topology, differential geometry, and/or graph theory

representation pre-conditioned generative adversarial networks (GAN). GAN is a kind of deep learning model consisting of a generator G to learn the data distribution, and a discriminator D to discriminate training set structural information from that of the generator G [88]. The G model is iteratively improved from the D feedback until the D cannot tell the difference between training set structural information and D set one. To improve the GAN performance and avoid vanishing gradient and mode collapse, we employ Wasserstein GAN (WGAN) [96] in our model. To further enhance the quality of the generated structures, we take advantage of the conditional GAN technique [97]. The deep learning (DL) models G and D are partially adapted from our binding energy prediction networks which are fed with data encoded in intrinsically low-dimensional manifolds with differential geometry, algebraic topology and graph theory. Figure 5 depicts the MGAN's framework. Network architectures of autodecoder and autoencoder are illustrated in Figs. S2 and S3, respectively. By varying combinations of different mathematics, we end up with several docking models. Specifically, If DL networks G and D only exploit algebraic topology, we name this docking model DM1. Similarly, we attain DM2 and DM3 when GAN model includes only algebraic graph and differential geometry based representations, respectively. Finally, DM4 is constructed with the assistance of algebraic topology, algebraic graph, and differential geometry. We employed the PDBbind v2018 dataset to train MathDL and MGAN models. The optimal hyperparameters of the MathDL model were selected by experience and finalized by hyperopt python package (http://github.com/hyperopt/hyperopt). The MGAN model was trained based on the setting of Wasserstein GAN network discussed in this work [96]. Furthermore, to enhance the pose generation quality, we carry out the transfer learning to further optimize the MGAN model with the protein family-specific structures.

Results and discussion

In this section, we present MathDL results and discuss our performances in the latest Grand Challenge named GC4.

Pose prediction results and discussion

We have participated in the docking challenge task since D3R GC2. Before the current challenge, i.e., GC4, our docking results in term of RMSE were not competitive in comparison to those of other participants. Specifically, our mean RMSD values are 6.03 Å and 3.78 Å for GC2 and GC3, respectively. These results reflect an improvement in our docking approaches but their accuracy is still behind the top submissions in GC3. Instead of depending on the docking programs such as Autodock Vina [4] and GLIDE [6] as we did in the previous challenges, our GC4 docking schemes were driven by advanced mathematical representations and sophisticated deep learning architectures. Consequently, we achieved remarkable performances on the pose prediction tasks. The rest of this section is devoted to result discussions.

Despite having two protein receptors in GC4, all the pose predictions were only for BACE ligands and were organized in two stages, Stage 1a and Stage 1b. In Stage 1a, participants were provided SMILES strings of 20 ligands to be docked, the FASTA sequence of the BACE protein, and the reference protein structure (PDBID: 5ygx, chain A) for the superimposition process. Stage 1b took place right after the end of Stage 1a. Stage 1b provided the experimental protein structures in the complexes with 20 ligands requested for pose predictions, in which the structures of these ligands were removed. Participants were still asked to predict their poses. Therefore, Stage 1b is often referred to a self-docking challenge. There are two evaluation metrics for the pose prediction tasks, namely median and mean calculated over all RMSD values between the predicted poses and crystal structures.

In Stage 1a, we submitted two results. Figure 6 illustrates the performances of 70 submissions having median RMSD < 10 Å . Our best submission having receipt ID 5t302 with median RMSD = 0.53 Å and being highlighted in the red color. This docking model was DM1. In Stage 1b, we delivered four submissions; unfortunately, none of them was ranked the first place in either the median or mean metric.

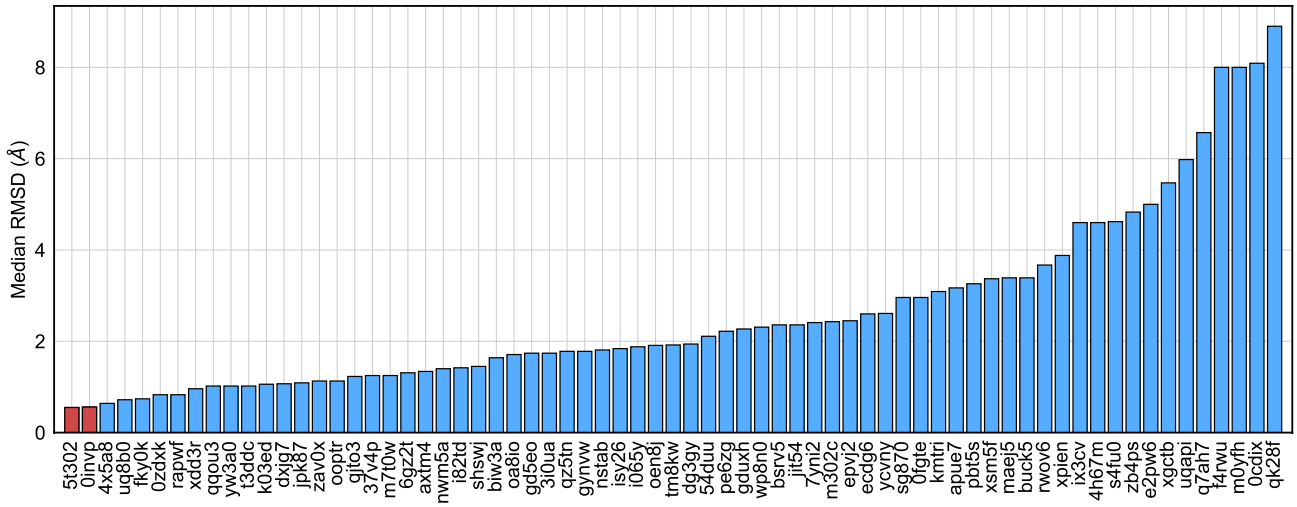

Fig. 6 Performance comparison of different submissions on pose prediction challenge of Stage 1a for the BACE dataset in term of median RMSD. Our submissions are highlighted in the red color, in which the best one is 5t302 with median RMSD = 0.55 Å

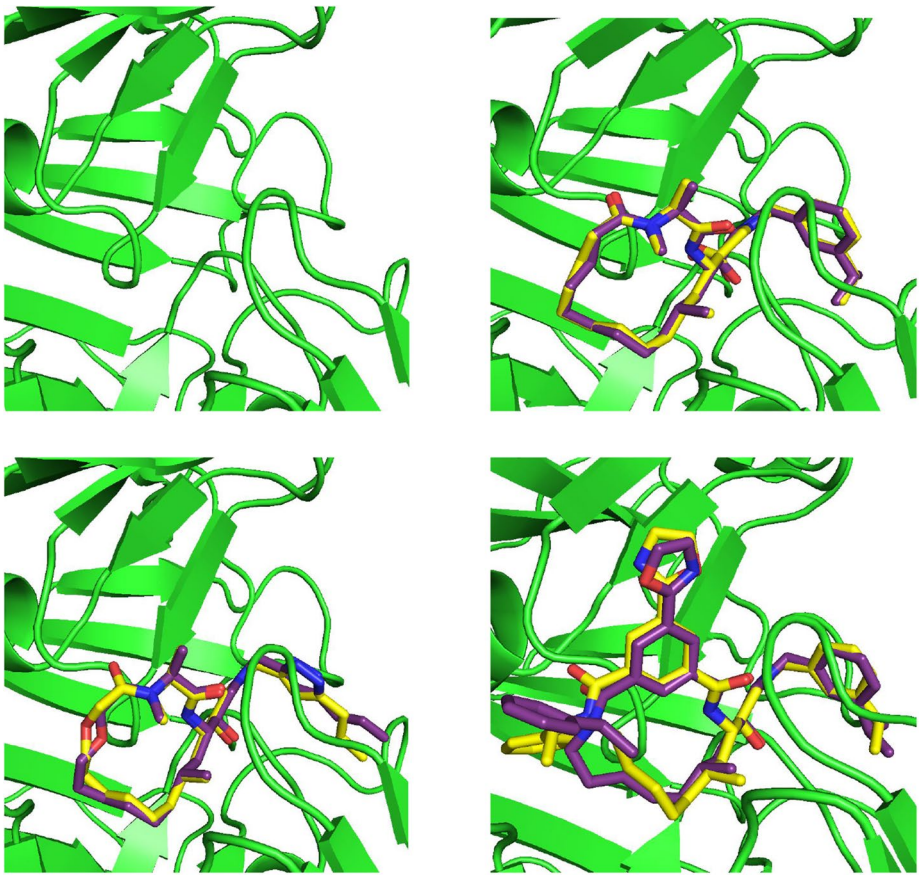

Fig. 7 Illustration of pose predictions by our MathGAN docking model with receipt ID 5t302. The top-left corner is original binding pocket of the BACE receptor. The top-right corner is our best pose prediction accuracy obtained when predicting BACE03's pose with RMSD = 0.23 Å . The bottom-left corner is our middle performance when predicting BACE05's pose with RMSD = 0.53 Å . The bottom-right is our worst performance when predicting BACE07's pose with RMSD = 2.63 Å . The experiment structures are in yellow while the predicted structures are in purple

However, our results were very promising. Particularly, our submission based on docking model DM3 with receipt ID itzv6 achieved mean RMSD of 0.73 Å which is at the second place and is a bit less accurate than the top submission with mean RMSD being 0.61 Å (receipt ID 5od5g). It may be noted that the best result in Stage 1b is not as good as that in Stage 1a. Figure 7 compares the poses predicted by

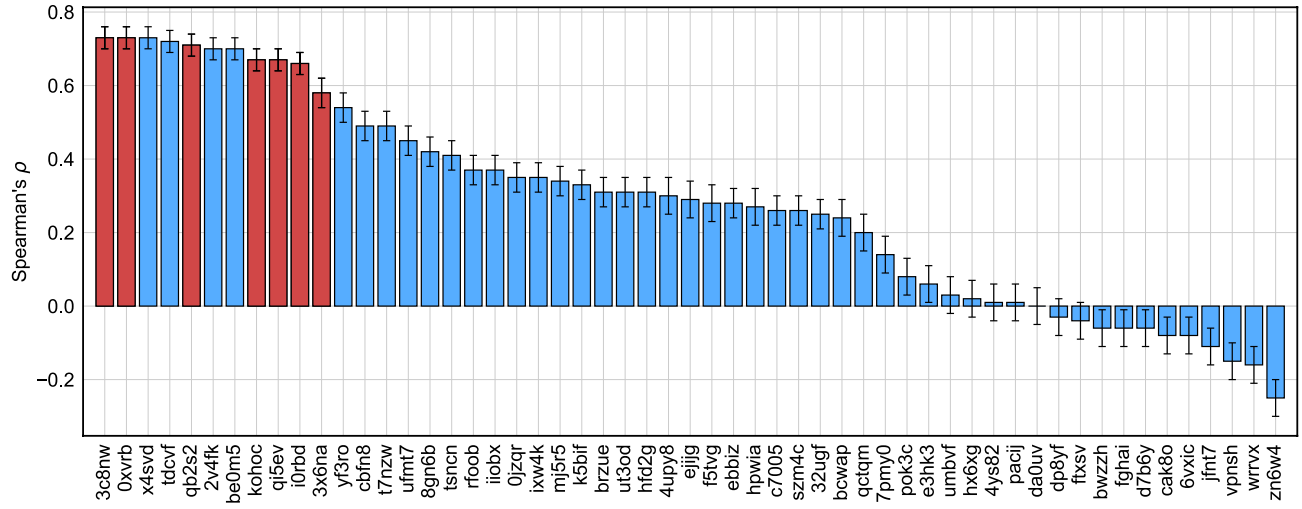

Fig. 8 Performance comparison of different submissions on the combined ligand and structure based scoring of CatS dataset in term of Spearman's ρ. Our submissions are highlighted in the red color, in which our top-ranked submissions are 3c8nw and 0xvrb with ρ=0.73

our submission ID 0invp to the corresponding experimental structures at different levels of accuracy.

It is interesting to find out that, the additional information of the co-crystal structures did not help our docking models. For example, our docking approach DM4 with submission ID Oinvp attained median RMSD of 0.53 Å and mean RMSD of 0.8 Å , respectively in Stage 1a. However, in Stage 1b, the same model labeled by receipt ID 2ieqo produced median RMSD and mean RMSD as high as 0.55 Å and 0.84 Å , respectively. These observations can confirm the robustness of our models and predictive value for the realistic situations in CADD when little or no co-crystal information is provided.

Affinity prediction results and discussion

There were two subchallenges for affinity prediction tasks. Subchallenge 1 regarded BACE ligands while Subchallenge 2 concerned CatS ligands. Both subchallenges were interested in affinity ranking of a diversity datasets and relative binding affinity predictions on the designated free energy set. There were two stages on BACE affinity prediction task, namely Stage 1 and Stage 2, whereas there was only one stage on CatS ligands. Unfortunately, we did not participate in Stage 1 of the BACE target since the announcement email made us overlook this contest.

Statistically, there were 154 compounds in the BACE dataset for affinity ranking contest, while there were 34 compounds for the calculation of relative or absolute binding affinities of the same receptor target. In CatS dataset, participants were asked to rank affinities of 459 ligands and predicted the binding energies of a smaller subset with 39 molecules. Moreover, Kendall's τ and Spearman's ρ were the evaluation metrics for affinity ranking challenges. In the binding free energy predictions, besides the aforementioned metrics, Pearson's r and centered root mean square error ($RMSE_c$) were utilized.

Overall, the official results from the D3R organizer have placed us among the top performers on these energy prediction contests. By considering specific evaluation metrics, we were ranked first place in *combined ligand and structure based scoring*[1], *structure based scoring*, and *free energy set* subcategories all belonging to the CatS dataset. For illustration, Figure 8 presents the Spearman's ρ performance of different submissions on the CatS affinity ranking contest combining ligand and structure based scoring models. Our best submission are highlighted in the red color with receipt IDs 3c8nw and 0xvrb. Both of them achieved the same Spearman's ρ as high as 0.73 and shared the first place with another group's submission having ID x4svd. In submission ID 3c8nw, we employed docking model DM4 for pose generation and model BP2 for the affinity prediction. While in submission ID 0xvrb, docking approach was DM3 and binding prediction protocol was BP2. In addition, our best result with ID ar5p6 achieved the lowest $RMSE_c$ for the free energy prediction of 39 designated CatS molecules. This successful submission utilized docking model DM4 and affinity prediction model BP2 for the calculations. Figure 9 presents $RMSE_c$ performance of various groups for the free energy prediction of CatS dataset. Table 1 summarizes the performances of our group at all categories in D3R GC4. We only counted the number of our submissions in the top three including ties. "No participation" at the results column

[1] This subcategory is the common list of ligand based and structure based scoring subcategories.

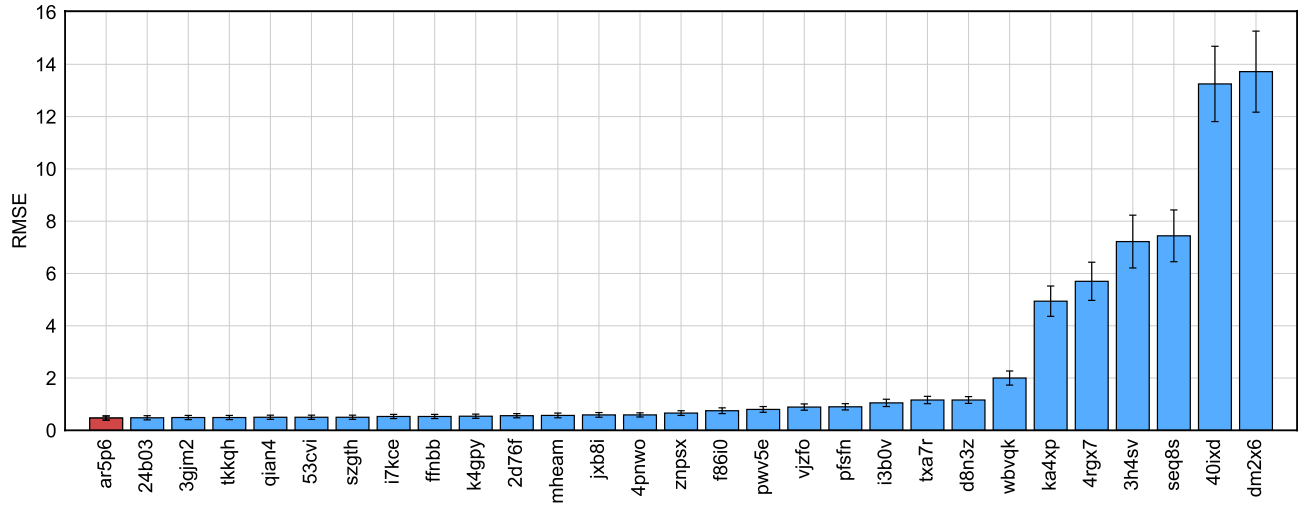

Fig. 9 Performance comparison of D3R GC4 participants on free energy set for CatS contest in term of centered RMSE $RMSE_c$. Our submissions are highlighted in the red color, in which our top-ranked prediction is ar5p6 with $RMSE_c = 0.47$ kcal/mol

implies that we did not participate in the corresponding contest. The blank results indicate that our predictions were not ranked within the top three.

It is noted that in the BACE affinity prediction, our results were not in the top three. In fact, our team was behind only to two teams that collected all the top three places in BACE affinity ranking, which indicates the consistence of our MathDL models in GC4 competitions.

Overall, the model BP2 was our best model for binding affinity prediction for both CatS and BACE datasets (see Table S1). The great performance of BP2 was expected since it combines algebraic topology, differential geometry, and graph theory features which help to enrich feature space and cover the most important aspects of physical and biological properties. However, there was a mixed conclusion when finding the best solution for pose prediction. Indeed, models DM3 and DM4 worked well for the CatS dataset, while DM1 was an only good solution for producing high quality poses for the BACE dataset (see Table S1). They helped the predictor BP2 achieved the best rankings among our submitted models. One can argue that DM1 achieved the best pose prediction for BACE ligands in Stage 1A; therefore it was foretasted to help BACE energy prediction tasks. The same behavior was observed for CatS dataset. According to our pre-validation results, DM4 which was our best model for the CatS pose prediction, achieved mean RMSD of 1.8 Å for the CatS pose prediction Stage 1B challenge in GC3. Note that the best submission in that subchallenge accomplished mean RMSD as low as 2.13 Å. It seems that the pose quality of our pose generation models correlates well to the accuracy of our binding affinity predictors.

Conclusion

The performances of our mathematical deep learning (MathDL) models on D3R GC4 are presented and discussed in this paper. We participated in a variety of D3R GC4 contests including pose predictions, affinity ranking, and absolute free energy predictions. Overall, our submissions were ranked the first in pose prediction in Stage 1a, affinity ranking and free energy predictions for Cathepsin ligands. Unfortunately, we did not get the first place on BACE datasets. Our best submission was only at the second place in free energy set for BACE in Stage 2 contest. In comparison to our previous D3R challenges, i.e., D3R GC2 and D3R GC3, we had two improvements in D3R GC4. The first improvement was the pose prediction. This was the first time we won this contest thanks to our newly developed docking model which integrates scalable low-dimensional rotational and translational invariant mathematical representations, such as differential geometry, algebraic graph, and algebraic topology, with well-designed generative adversarial networks. The second improvement was the affinity ranking for a dataset with diverse chemical properties. In previous challenges, our approaches performed well on free energy predictions but not on affinity ranking. In GC4, we successfully unified our newly established models, i.e., differential geometry and algebraic graph, and our well-known algebraic topology into powerful and robustness convolutional neural network models for binding affinity predictions.

In terms of efficiency, at this point, our MathDL models are quite automated. With sufficient computer resources, our MathDL models can finish all the GC4 competition tasks in a week or so.

Table 1 Overview of MathDL's performance in D3R GC4

Dataset	Contest	Results
Pose prediction		
BACE Stage 1A	Pose prediction	Ranked 1st (1/2)[i]; Ranked 2nd (3/3)[ii]
BACE Stage 1B	Pose prediction	Ranked 2nd (2/2)[iii]; Ranked 3rd (1/2)[iv]
Affinity predictions		
Cathepsin Stage 2	Combined ligand and structure based scoring	Ranked 1st (2/5)[v]; Ranked 2nd (2/3)[vi]; Ranked 3rd (2/4)[vii]
Cathepsin Stage 2	Ligand based scoring	No participation
Cathepsin Stage 2	Structure based scoring	Ranked 1st (2/4)[viii]; Ranked 2nd (3/3)[ix]; Ranked 3rd(3/3)[x]
Cathepsin Stage 2	Free energy set	Ranked 1st (1/7)[xi]; Ranked 2nd (1/7)[xii]; Ranked 3rd(3/5)[xiii]
BACE Stage 1	Combined ligand and structure	No participation
BACE Stage 1	Ligand based scoring	No participation
BACE Stage 1	Structure based scoring	No participation
BACE Stage 1	Free energy set	No participation
BACE Stage 2	Combined ligand and structure	
BACE Stage 2	Ligand based scoring	No participation
BACE Stage 2	Structure based scoring	
BACE Stage 2	Free energy set	Ranked 2nd (3/4)[xiv]; Ranked 3rd (1/4)[xv]

Superscript	Submission ID	Evaluation metric	Docking protocol	Scoring protocol
i	5t302	Median RMSD	DM1	
ii	5t302	Mean RMSD	DM1	
	0invp	Median RMSD	DM4	
	0invp	Mean RMSD	DM4	
iii	2ieqo	Median RMSD	DM4	
	itzv6	Mean RMSD	DM3	
iv	4myne	Mean RMSD	DM1	
v	0xvrb	Spearman's ρ	DM3	BP2
	3c8nw	Spearman's ρ	DM4	BP2
vi	0xvrb	Kendall's τ	DM3	BP2
	3c8nw	Kendall's τ	DM4	BP2
vii	qb2s2	Kendall's τ	DM1	BP2
	qb2s2	Spearman's ρ	DM1	BP2
viii	0xvrb	Spearman's ρ	DM3	BP2
	3c8nw	Spearman's ρ	DM4	BP2
ix	0xvrb	Kendall's τ	DM3	BP2
	3c8nw	Kendall's τ	DM4	BP2
	qb2s2	Spearman's ρ	DM1	BP2

Table 1 (continued)

Superscript	Submission ID	Evaluation metric	Docking protocol	Scoring protocol
x	qb2s2	Kendall's τ	DM1	BP2
	qi5ev	Spearman's ρ	DM3	BP1
	kohoc	Spearman's ρ	DM2	BP2
xi	ar5p6	$RMSE_c$	DM4	BP2
xii	24b03	$RMSE_c$	DM3	BP2
xiii	24b03	Kendall's τ	DM3	BP2
	24b03	Spearman's ρ	DM3	BP2
	24b03	Pearson's r	DM3	BP2
xiv	8frur	Kendall's τ	DM1	BP2
	8frur	Spearman's ρ	DM1	BP2
	8frur	$RMSE_c$	DM1	BP2
xv	8frur	Pearson's r	DM1	BP2

The numbers in "*(a / b)*" indicates that *a* number of our predictions had the ranking and there was a total of *b* submissions sharing the ranking

It is worth noting that our models for GC4 was the less competitive performance in BACE affinity ranking and free energy predictions. Additionally, it seems that our docking model did not upgrade when the co-crystal structures became available. These issues are under our investigation.

Acknowledgements This work was supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473 and NIH Grant GM126189. DDN and GWW are also funded by Bristol-Myers Squibb and Pfizer.

References

1. Gathiaka S, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN, Delproposto J, Kubish G, Dunbar JB, Carlson HA et al (2016) D3r grand challenge 2015: evaluation of protein-ligand pose and affinity predictions. J Comput-Aided Mol Des 30(9):651–668
2. Gaieb Z, Liu S, Gathiaka S, Chiu M, Yang H, Shao C, Feher VA, Walters WP, Kuhn B, Rudolph MG et al (2018) D3r grand challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. J Comput-Aided Mol Des 32(1):1–20
3. Gaieb Z, Parks CD, Chiu M, Yang H, Shao C, Walters WP, Lambert MH, Nevins N, Bembenek SD, Ameriks MK et al (2019) D3r grand challenge 3: blind prediction of protein-ligand poses and affinity rankings. J Comput-Aided Mol Des 33(1):1–18
4. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461
5. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3):727–748
6. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, JK JKP, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. J Med Chem 47:1739
7. Abagyan R, Totrov M, Kuznetsov D (1994) Icm-a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15(5):488–506
8. Liu J, Wang R (2015) Classification of current scoring functions. J Chem Inf Model 55(3):475–482
9. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem 38:2681–2691
10. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV (2008) Medusascore: an acurate force field-based scoring function for virtual drug screening. J Chem Inf Model 48:1656–1662
11. Muegge I, Martin Y (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42(5):791–804
12. Velec HFG, Gohlke H, Klebe G (2005) Knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem 48:6296–6303
13. Zheng Z, Wang T, Li P, Merz KM Jr (2015) KECSA-Movable type implicit solvation model (KMTISM). J Chem Theor Comput 11:667–682
14. Huang SY, Zou X (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: I. derivation of interaction potentials. J Comput Chem 27:1865–1875

15. Verkhivker G, Appelt K, Freer ST, Villafranca JE (1995) Empirical free energy calculations of ligand-protein crystallographic complexes. i. Knowledge based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus protease binding affinity. Protein Eng 8:677–691
16. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) 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 11:425–445
17. Wang R, Lai L, Wang S (2002) Further development and validation of empirical scoring functions for structural based binding affinity prediction. J. Comput-Aided Mol. Des 16:11–26
18. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein -ligand binding affinity with applications to molecular docking. Bioinformatics 26(9):1169–1175
19. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
20. Li H, Leung K-S, Wong M-H, Ballester PJ (2014) Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. BMC Bioinform 15(1):1
21. Nguyen DD, Xiao T, Wang ML, Wei GW (2017) Rigidity strengthening: a mechanism for protein-ligand binding. J Chem Inf Model 57:1715–1721
22. Cang ZX, Wei GW (2018) Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. Int J Numer Method Biomed Eng. https://doi.org/10.1002/cnm.2914
23. Cang ZX, Wei GW (2017) TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLOS Comput Biol 13(7):e1005690. https://doi.org/10.1371/journal.pcbi.1005690
24. Cang ZX, Mu L, Wei GW (2018) Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLOS Comput Biol 14(1):e1005929. https://doi.org/10.1371/journal.pcbi.1005929
25. Nguyen DD, Wei G-W (2019) Dg-gl: differential geometry-based geometric learning of molecular datasets. Int J Numer Method Biomed Eng 35(3):e3179
26. Nguyen D, Wei G-W (2019) Agl-score: algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. J Chem Inf Model 59(7):3291–3304
27. Nguyen DD, Cang Z, Wu K, Wang M, Cao Y, Wei G-W (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in d3r grand challenges. J Comput-Aided Mol Des 33(1):71–82
28. Wei GW (2010) Differential geometry based multiscale models. Bull Math Biol 72:1562–1622
29. Chen Z, Zhao S, Chun J, Thomas DG, Baker NA, Bates PB, Wei GW (2012) Variational approach for nonpolar solvation analysis. J Chem Phys 137:084101
30. Wang B, Wei G-W (2015) Parameter optimization in differential geometry based solvation models. J Chem Phys 143:134119
31. Chen D, Wei GW (2012) Quantum dynamics in continuum for proton transport III: generalized correlation. J Chem Phys 136:134109
32. Chen D, Wei GW (2012) Quantum dynamics in continuum for proton transport—generalized correlation. J Chem Phys 136:134109
33. Wei G-W, Zheng Q, Chen Z, Xia K (2012) Variational multiscale models for charge transport. SIAM Rev 54(4):699–754
34. Wei GW (2013) Multiscale, multiphysics and multidomain models I: basic theory. J Theor Comput Chem 12(8):1341006
35. Chen D, Wei GW (2013) Quantum dynamics in continuum for proton transport I: basic formulation. Commun Comput Phys 13:285–324
36. Feng X, Xia K, Tong Y, Wei G-W (2012) Geometric modeling of subcellular structures, organelles and large multiprotein complexes. Int J Numer Method Biomed Eng 28:1198–1223
37. Xia KL, Feng X, Tong YY, Wei GW (2014) Multiscale geometric modeling of macromolecules i: Cartesian representation. J Comput Phys 275:912–936
38. Mu L, Xia K, Wei G (2017) Geometric and electrostatic modeling using molecular rigidity functions. J Comput Appl Math 313:18–37
39. Nguyen DD, Wei GW (2017) The impact of surface area, volume, curvature and Lennard-Jones potential to solvation modeling. J Comput Chem 38:24–36
40. Kaczynski T, Mischaikow K, Mrozek M (2004) Computational homology. Springer-Verlag, Berlin
41. Edelsbrunner H, Letscher D, Zomorodian A (2001) Topological persistence and simplification. Discret Comput Geom 28:511–533
42. Zomorodian A, Carlsson G (2005) Computing persistent homology. Discret Comput Geom 33:249–274
43. Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS (2007) Persistent voids a new structural metric for membrane fusion. Bioinformatics 23:1753–1759
44. Dabaghian Y, Mémoli F, Frank L, Carlsson G (2012) A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput Biol 8(8):e1002581
45. Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V (2014) Topological measurement of protein compressibility via persistence diagrams. Jpn J Ind Appl Math 32:1–17
46. Xia KL, Wei GW (2014) Persistent homology analysis of protein structure, flexibility and folding. Int J Numer Method Biomed Eng 30:814–844
47. Xia KL, Wei GW (2015) Persistent topology for cryo-EM data analysis. Int J Numer Method Biomed Eng 31:e02719
48. Xia KL, Feng X, Tong YY, Wei GW (2015) Persistent homology for the quantitative prediction of fullerene stability. J Comput Chem 36:408–422
49. Wang B, Wei GW (2016) Object-oriented persistent homology. J Comput Phys 305:276–299
50. Liu B, Wang B, Zhao R, Tong Y, Wei G-W (2017) Eses: software for e ulerian solvent excluded surface. J Comput Chem 38(7):446–466
51. Cang ZX, Mu L, Wu K, Opron K, Xia K, Wei G-W (2015) A topological approach to protein classification. Mol Based Math Biol 3:140–162
52. Cang ZX, Wei GW (2017) Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. Bioinformatics 33:3549–3557
53. Wu K, Wei GW (2018) Quantitative toxicity prediction using topology based multitask deep neural networks. J Chem Inf Model 58:520–531
54. Wu K, Zhao Z, Wang R, Wei GW (2018) TopP-S: persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. J Comput Chem 39:1444–1454
55. Hosoya H (1971) Topological index. a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Soc Jpn 44(9):2332–2339
56. Hansen PJ, Jurs PC (1988) Chemical applications of graph theory. Part i. Fundamentals and topological indices. J Chem Educ 65(7):574
57. Newman M (2010) Networks: an introduction. Oxford University Press, Oxford
58. Bavelas A (1950) Communication patterns in task-oriented groups. J Acoust Soc Am 22(6):725–730
59. Dekker A (2005) Conceptual distance in social network analysis. J Soc Struct 6:31

60. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2:173–181
61. Yang LW, Chng CP (2008) Coarse-grained models reveal functional dynamics-I. Elastic network models-theories, comparisons and perspectives. Bioinf Biol Insights 2:25–45
62. Wei GW, Zhan M, Lai CH (2002) Tailoring wavelets for chaos control. Phys Rev Lett 89:284103
63. Go N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80:3696–3700
64. Tasumi M, Takenchi H, Ataka S, Dwidedi AM, Krimm S (1982) Normal vibrations of proteins: glucagon. Biopolymers 21:711–714
65. Brooks BR, Bruccoleri RE, Olafson BD, States D, Swaminathan S, Karplus M (1983) Charmm: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
66. Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J Mol Biol 181(3):423–447
67. Flory PJ (1976) Statistical thermodynamics of random networks. Proc R. Soc. Lond. A 351:351–378
68. Bahar I, Atilgan AR, Demirel MC, Erman B (1998) Vibrational dynamics of proteins: significance of slow and fast modes in relation to function and stability. Phys Rev Lett 80:2733–2736
69. Atilgan AR, Durrell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505–515
70. Hinsen K (1998) Analysis of domain motions by approximate normal mode calculations. Proteins 33:417–429
71. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng 14:1–6
72. Cui Q, Bahar I (2010) Normal mode analysis: theory and applications to biological and chemical systems. Chapman and Hall, London
73. Balaban AT (1976) Chemical applications of graph theory. Academic Press, Cambridge
74. Trinajstic N (1983) Chemical graph theory. CRC Press, Boca Raton
75. Schultz HP (1989) Topological organic chemistry. 1. Graph theory and topological indices of alkanes. J Chem Inf Comput Sci 29(3):227–228
76. Foulds LR (2012) Graph theory applicatons. Springer, Berlin
77. Ozkanlar A, Clark AE (2014) Chemnetworks: a complex network analysis tool for chemical systems. J Comput Chem 35(6):495–505
78. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
79. Canutescu AA, Shelenkov AA, Dunbrack RL (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12(9):2001–2014
80. Ryslik GA, Cheng Y, Cheung K-H, Modis Y, Zhao H (2014) A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. BMC Bioinform 15(1):86
81. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. Proteins-Struct Funct Genet 44:150–165
82. Vishveshwara S, Brinda K, Kannan N (2002) Protein structure: insights from graph theory. J Theor Comput Chem 1(01):187–211
83. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2017) Moleculenet: A benchmark for molecular machine learning. arXiv preprint arXiv:1703.00564
84. Quan L, Lv Q, Zhang Y (2016) Strum: structure-based prediction of protein stability changes upon single-point mutation. Struct Bioinform (In press)
85. Pires DEV, Ascher DB, Blundell TL (2014) mcsm: predicting the effects of mutations in proteins using graph-based signatures. Struct Bioinform 30:335–342
86. Park JK, Jernigan R, Wu Z (2013) Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations. Bull Math Biol 75:124–160
87. Bramer D, Wei GW (2018) Weighted multiscale colored graphs for protein flexibility and rigidity analysis. J Chem Phys 148:054103
88. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 2672–2680
89. Xia KL, Opron K, Wei GW (2013) Multiscale multiphysics and multidomain models—flexibility and rigidity. J Chem Phys 139:194109
90. Opron K, Xia KL, Wei GW (2014) Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. J Chem Phys 140:234105
91. Nguyen DD, Xia KL, Wei GW (2016) Generalized flexibility-rigidity index. J Chem Phys 144:234106
92. Wei GW (2000) Wavelets generated by using discrete singular convolution kernels. J Phys A 33:8577–8596
93. Soldea O, Elber G, Rivlin E (2006) Global segmentation and curvature analysis of volumetric data sets using trivariate b-spline functions. IEEE Trans PAMI 28(2):265–278
94. Edelsbrunner H (1992) Weighted alpha shapes. Technical Report. University of Illinois, Champaign
95. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
96. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning, pp 214–223
97. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784