PNAS

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Knot data analysis using multiscale Gauss link integral

Li Shen[a,1], Hongsong Feng[a,1], Fengling Li[b], Fengchun Lei[b], Jie Wu[c], and Guo-Wei Wei[a,d,e,2]

Affiliations are included on p. 12.

Edited by Kenneth Ribet, University of California, Berkeley, CA; received April 27, 2024; accepted July 17, 2024

In the past decade, topological data analysis has emerged as a powerful algebraic topology approach in data science. Although knot theory and related subjects are a focus of study in mathematics, their success in practical applications is quite limited due to the lack of localization and quantization. We address these challenges by introducing knot data analysis (KDA), a paradigm that incorporates curve segmentation and multiscale analysis into the Gauss link integral. The resulting multiscale Gauss link integral (mGLI) recovers the global topological properties of knots and links at an appropriate scale and offers a multiscale geometric topology approach to capture the local structures and connectivities in data. By integration with machine learning or deep learning, the proposed mGLI significantly outperforms other state-of-the-art methods across various benchmark problems in 13 intricately complex biological datasets, including protein flexibility analysis, protein–ligand interactions, human Ether-à-go-go-Related Gene potassium channel blockade screening, and quantitative toxicity assessment. Our KDA opens a research area—knot deep learning—in data science.

knot data analysis | Gauss link integral | multiscale analysis

Significance

Knot theory is a pivotal mathematical branch and has garnered tremendous research interest for over 200 years. Despite its broad applications, it has been limited to qualitative analysis. We introduce a multiscale knot theory paradigm that extends its scope from qualitative to quantitative analysis, providing a cutting-edge computational biology tool. For instance, drug discovery, a highly challenging scientific task, draws contributions from various disciplines. Our paradigm redefines the state-of-the-art in protein–ligand binding affinity scoring. Its utility in molecular science is also validated by other applications, including quantitative protein flexibility analysis, drug toxicity evaluation, and drug side-effect screening. Finally, our paradigm opens an era of exploration in mathematical science that underpins data science and artificial intelligence.

Knots are ubiquitous in nature, from animal nests, interlocked tree branches, vines, tendrils, chromosome chains, to DNA double helices. Humans have been intrigued by knot tying due to their practical functions, aesthetic appeal, and spiritual symbolism since prehistoric times. Mathematical theory of knots dated back to 1,771 by Alexandre-Théophile Vandermonde. Knot theory is one of the most active areas of mathematical studies, concerning the embeddings of a closed circle S^1 into the three-dimensional (3D) Euclidean space, their classification, equivalence after continuous deformations, or ambient isotopy (1). Some of the most important knot invariants, which differentiate knots, include knot crossing number, knot group (2), knot polynomials (1), knot Floer homology (3), Khovanov homology (4), etc.

Knot theory has been applied to various fields such as physics (5), biochemistry (6), and biology (7–9), with limited success. Most real-world objects might not be a closed circle. In applications, ambient isotopy typically has major different properties, while keeping the global knot information unchanged. For instance, the realization of many object functions, such as the molecular recognition of DNA, depends on local structures. Therefore, it is imperative to develop knot theory-based tools that are robust and effective for applications.

Several attempts have been made to address the aforementioned challenge. Jamroz et al. proposed the protein topology database KnotProt to study knot and slipknot type of proteins (10). Dabrowski-Tumanski et al. extend the database to include links and spatial graphs, and also enable the calculation of topological polynomials invariant of those structures (11). Recently, Panagiotou and Kauffman have proposed new invariants for open curves in 3-space (12). In addition, Baldwin et al. (13) attempted to localize knot information by intercepting some specific intervals in the linear structure of an open curve. Nevertheless, these approaches are still global topological in nature.

Multiscale analysis can offer a viable localization scheme for knot data analysis, given its remarkable success in diverse areas such as wavelet theory and topological data analysis (TDA). Persistent homology, as a prominent technique in TDA, combines concepts from algebraic topology, geometry, and multiscale analysis to analyze complex datasets (14, 15). It uncovers the complex topological invariants and patterns of data at various scales, which are not easily discernible with traditional geometric and statistical techniques. Topological features facilitate valuable representation learning, and their efficacy is demonstrated through integration with deep learning models, specifically in the context of topological deep learning (TDL) coined by us in 2017 (16). Compelling applications which consistently demonstrate the relevant advantages of TDL over existing methods are the victories of TDL in the D3R Grand Challenges, a worldwide annual

Author contributions: G.-W.W. designed research; L.S., H.F., and G.-W.W. performed research; L.S., H.F., F. Li, F. Lei, J.W., and G.-W.W. analyzed data; and L.S., H.F., and G.-W.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] L.S. and H.F. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: weig@msu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2408431121/-/DCSupplemental.

Published October 11, 2024.

competition series in computer-aided drug, (17), the discovery of SARS-CoV-2 evolution mechanisms (18), and the successful forecasting of SARS-CoV-2 variants BA.2 (19), and BA.5 (20) about two months in advance.

Mathematically, linking number is a knot invariant that measures the extent of linkage between two closed curves in 3D space, representing the number of times that each curve winds around the other. The Gauss linking integral (21), also known as Gauss's integral, gives an explicit formulation for the linking number. It serves as a fundamental tool for studying knots, links, and other topological structures within 3D space. This tool holds significance in various fields, including knot theory, geometric topology, differential geometry, and quantum field theory. For example, for idealized Dirac-string center vortices, the Chern–Simons number can be given by the Gauss link integral (22). High-order link integrals were proposed (23). However, these approaches are typically global and qualitative.

The objective of this work is to introduce knot data analysis (KDA) as a paradigm for data science. To this end, we propose a framework called multiscale Gauss linking integral (mGLI) by integrating multiscale analysis with classical knot and knot-related theories. The proposed mGLI can capture both local and global information of knots, curves, and other curve-like objects by admitting a family of open balls around each segment on the objects. We define a metric to describe the degree of the local entanglement within each ball. By increasing the ball radius, the metric will incorporate additional local information in objects and finally reveal the global properties of the original structure such as knots and entangled links. The proposed mGLI effectively captures intrinsic structures, and patterns in complex data offering valuable low-dimensional embeddings of the data. To assess the performance of mGLI, we consider 13 benchmark datasets across various domains, including protein flexibility analysis, protein–ligand binding affinity prediction, human Ether-à-go-go-Related Gene (hERG) blockade classification, and quantitative toxicity predictions. The performance of mGLI is compared with that of other state-of-art approaches, including TDA, unlocking geometric topology's potential.

In contrast to the previous qualitative and descriptive knot theory approaches, the mGLI is a quantitative and predictive strategy. It offers a tool in knot theory analysis and opens an area in data analysis and knot learning.

Results

Overview of KDA. Fig. 1 outlines the proposed KDA platform. Like TDA, KDA utilizes a multiscale strategy to capture local structural information of data at various scales and represent the information in a knot invariant, the Gauss link integral or Gauss link number. While globally the Gauss link number quantifies the linking or entanglement between two curves or loops in 3D space, our mGLI further measures local entanglements at each pair of link or curve segments. As shown in Fig. 1*A*, such local information is systematically collected across scales and assembled over all segments, giving rise to a vectorization of the original structure.

A specific application of mGLI to a protein–ligand complex is given in Fig. 1*B*. An element-specific mGLI strategy is introduced to elucidate physical and chemical interactions (Fig. 1*C*) and to ensure the scalability across different complexes via statistics (Fig. 1*D*). In the case of protein–ligand complex characterization, chemical and biological information, such as hydrogen bonds, electrostatics, hydrophilicity, and hydrophobicity can be delineated by element-specific mGLI strategy. The intrinsic molecular properties in the 3D structures are properly decoded into low-dimensional topological representations, which are suitable for downstream molecular property analysis and prediction. Theoretical details are provided in *Methods*.

The proposed mGLI method captures stereochemical information that is crucial for molecular interactions. In complement, pretrained deep language models are able to access evolutionary and constitutional information of the problem under study. Specifically, we use a transformer-based pretrained model for protein embedding (24), while transformer and autoencoder-based pretrained models are utilized for small molecule embedding (25, 26) as indicated in Fig. 1*E*. These embeddings are paired with mGLIs for downstream prediction tasks as shown in Fig. 1*F*.

mGLI. It is intrinsic to describe real-world data by mathematical objects, such as knots, knotoids, lassos, links, linkoids, cysteine knots, etc. (Fig. 5*A*). The mGLI involves partitioning knots and other curved objects into segments and conducting a multiscale analysis at each segment. Upon curve segmentation, Gauss link integrals are defined at various scales to quantitatively capture structure, connectivity, and entanglement. The global topological invariant properties are ultimately recovered when a sufficiently large scale is reached. Below, we give some essential formulations of the proposed mGLI method.

Definition 1 [Gauss linking integral]: Given two disjoint open or closed curves l_1 and l_2, parameterized as $\gamma_1(s)$ and $\gamma_2(t)$, respectively, the following double integral gives the Gauss linking integral that characterizes the degree of interlinking between l_1 and l_2 (27):

$$L(l_1, l_2) = \frac{1}{4\pi} \int_{[0,1]} \int_{[0,1]} \frac{\det(\dot{\gamma_1}(s), \dot{\gamma_2}(t), \gamma_1(s) - \gamma_2(t))}{|\gamma_1(s) - \gamma_2(t)|^3} \, ds\, dt, \tag{1}$$

where $\dot{\gamma_1}(s)$ and $\dot{\gamma_2}(t)$ are derivative of $\gamma_1(s)$ and $\gamma_2(t)$, respectively.

***Definition 2 [Segmentation of Gauss linking integral]*:** Given finite curve segments P_n and Q_m for disjoint open or closed curves $l_1$and l_2, respectively, the segmentation of Gauss linking integral induced by the curve segments is defined as the following $n \times m$ segmentation matrix:

$$G = \begin{pmatrix} L(p_1, q_1) & L(p_1, q_2) & \cdots & L(p_1, q_m) \\ L(p_2, q_1) & L(p_2, q_2) & \cdots & L(p_2, q_m) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, q_1) & L(p_n, q_2) & \cdots & L(p_n, q_m) \end{pmatrix}, \tag{2}$$

where $p_i \in P_n$ and $q_j \in Q_m$ are curve segments of l_1 and l_2, respectively. Examples of segmentation of Gauss linking integral for Hopf link are offered in *SI Appendix*, section 1A.

Remark 1: The segmentation of the Gauss linking integral serves as the basis for our multiscale modeling. Since the objects in the segmentation of Gauss linking integral are curve segments, we define the distance of curve segments $d(p_i, q_j)$ with Euclidean distance.

Definition 3 [Scaled Gauss linking integral]: Given a finite set of real numbers $R = \{r_0, r_1, r_2, r_3, \cdots, r_k\}$, where $0 = r_0 < r_1 < r_2 < \cdots < r_k$, the Gauss linking integral at scale $[r_t, r_{t+1}]$ is defined as Eqs. **3** and **4**.

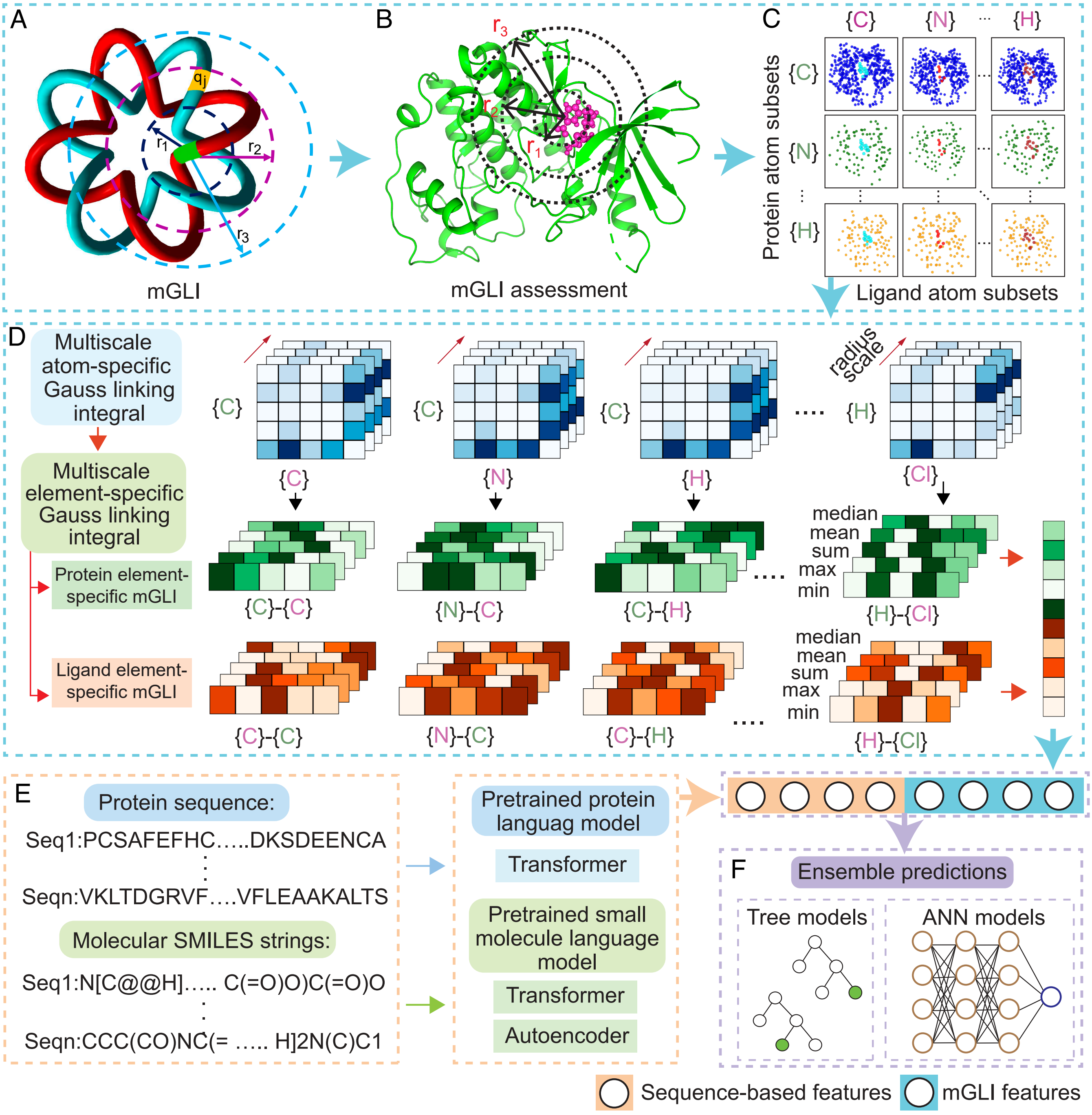
A
qj
r1
r2
r3
mGLI
B
r3
r2
r1
mGLI assessment
C
{C}
{N}
{H}
Protein atom subsets
{C}
{N}
{H}
Ligand atom subsets
D
Multiscale atom-specific Gauss linking integral
Multiscale element-specific Gauss linking integral
Protein element-specific mGLI
Ligand element-specific mGLI
{C}
{C}
{C}
{H}
radius scale
{C}
{N}
{H}
{Cl}
median
mean
sum
max
min
{C}-{C}
{N}-{C}
{C}-{H}
{H}-{Cl}
median
mean
sum
max
min
{C}-{C}
{N}-{C}
{C}-{H}
{H}-{Cl}
E
Protein sequence:
Seq1:PCSAFEFHC…..DKSDEENCA
Seqn:VKLTDGRVF….VFLEAAKALTS
Molecular SMILES strings:
Seq1:N[C@@H]….. C(=O)O)C(=O)O
Seqn:CCC(CO)NC(= ….. H]2N(C)C1
Pretrained protein languag model
Transformer
Pretrained small molecule language model
Transformer
Autoencoder
F
Ensemble predictions
Tree models
ANN models
Sequence-based features
mGLI features

Fig. 1. The conceptual diagram of the knot data analysis (KDA) platform for biological data learning. (*A*) An illustration of multiscale Gauss linking integral-based KDA on a (2, 8) torus. (*B*) mGLI is applied to the assessment of biomolecular 3D structures with multiple radius scales applied around each atom. (*C*) An element-specific mGLI strategy is introduced to embed physical and chemical interactions. (*D*) Atom-specific mGLI features are extracted to characterize atomic interactions in the protein–ligand complex. Statistics is used to ensure the scalability across different complexes. (*E*) Sequence-based features are generated for the amino acid sequence and the SMILES string, respectively, using pretrained natural language processing models. (*F*) The mGLI features and sequence-based features are paired for downstream predictions and analysis using gradient-boosting decision tree models or deep neural networks. Colors of frames and large arrows indicate the workflows in different modules: (*A*–*D*) denote a structure-based module (blue), (*E*) highlights a sequence-based module (orange), and (*F*) represents a prediction module (purple).

$$G^{r_t,r_{t+1}} = \begin{pmatrix} \chi_{[r_t,r_{t+1}]}(d(p_1,q_1))L(p_1,q_1) & \chi_{[r_t,r_{t+1}]}(d(p_1,q_2))L(p_1,q_2) & \cdots & \chi_{[r_t,r_{t+1}]}(d(p_1,q_m))L(p_1,q_m) \\ \chi_{[r_t,r_{t+1}]}(d(p_2,q_1))L(p_2,q_1) & \chi_{[r_t,r_{t+1}]}(d(p_2,q_2))L(p_2,q_2) & \cdots & \chi_{[r_t,r_{t+1}]}(d(p_2,q_m))L(p_2,q_m) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_{[r_t,r_{t+1}]}(d(p_n,q_1))L(p_n,q_1) & \chi_{[r_t,r_{t+1}]}(d(p_n,q_2))L(p_n,q_2) & \cdots & \chi_{[r_t,r_{t+1}]}(d(p_n,q_m))L(p_n,q_m) \end{pmatrix}, \qquad [3]$$

where

$$\chi_{[r_t,r_{t+1}]}(x)=\begin{cases}1, \text{if } x\in[r_t,r_{t+1}]\\ 0, \text{else}\end{cases}. \quad [4]$$

Remark 2: The scaled Gauss linking integral is used to extract appropriate linking integral within the scale. As shown in the curve segmentation for a (2, 8) torus of Fig. 1*A*, each torus has a collection of segments. We have $G_{ij}^{0,r_1}=0$, $G_{ij}^{r_1,r_2}=L(p_i,q_j)$, and $G_{ij}^{r_2,r_3}=0$. The scaled integral provides a way to capture local interactions between segments for a given scale. Cumulative integrals across expanding scales offer additional local structural insights, gradually unveiling broader global characteristics and relationships. Accordingly, multiscale Gauss linking integral features can be designed for various systems (*Methods*).

Definition 4 [Localized scaled Gauss linking integral]: For given scale $[r_t,r_{t+1}]$, we can define the localized scaled Gauss linking integral at p_i or q_j by the followings:

$$J^{r_t,r_{t+1}}(p_i)=\sum_{s=1}^{m}G_{is}^{r_t,r_{t+1}}, \quad [5]$$

$$J^{r_t,r_{t+1}}(q_j)=\sum_{s=1}^{n}G_{sj}^{r_t,r_{t+1}}. \quad [6]$$

Remark 3: By examining Gauss linking integrals at different scales, we obtain multiscale representation. The localized scaled Gauss linking integral gives rise to a measurement for each curve segment in the curve. By considering different scales, the localized scaled Gauss linking integral provides a featurization of each curve segment u:

$$Feature(u)=(J^{r_1,r_2}(u),J^{r_2,r_3}(u),\cdots,J^{r_{k-1},r_k}(u)). \quad [7]$$

In the case of biomolecular data characterization, curve segmentation is centered at atoms. Consequently, a scaled Gauss linking integral is tailored in an atom-specific or element-specific manner. Localized scaled Gauss linking integrals characterize atomic interactions across various scales, facilitating molecular multiscale analysis.

KDA of Biological Data. Biological systems are intricately complex and pose grand challenges. We evaluate the performance of mGLI with 13 benchmark datasets in four classes of biological systems, including protein flexibility analysis, protein–ligand binding affinity prediction, the classification of hEGR channel blockers, and quantitative toxicity prediction. To develop predictive machine learning models, we incorporate mGLI features with linear regression algorithm, gradient boosting decision trees (GBDT), deep neural networks (DNN), and multitask deep neural networks (MTDNN). Extensive comparison with the state-of-the-art is carried to demonstrate utility, reliability, and robustness of the proposed mGLI-based KDA platform.

Protein flexibility analysis. Proteins are inherently flexible and undergo various motions to maintain their functions. Protein flexibility is often experimentally measured with B-factors, also known as temperature factors or atomic displacement parameters. High B-factors indicate increased atomic mobility, suggesting the location of the protein that is flexible or involves conformational changes. Low B-factors, on the other hand, indicate rigid regions with limited atomic motion. We assess the effectiveness of the proposed mGLI-base features in predicting protein B-factors (*Methods*). The mGLI features are integrated with linear regression algorithm. It has been a tradition in B-factor predictions for all methods to utilize the same simple machine learning algorithm, thereby ensuring a fair comparison of various approaches.

Typically, the B-factor prediction focuses on C_α atoms in a protein as shown in Fig. 2*A* for protein (PDBID: 1J27). We segment the protein polymer chain structure into C_α atoms to facilitate Gauss linking integral calculations of atomic interactions among C_α atoms. The resulting atom-wise mGLI matrix is depicted in Fig. 2*B* with reference to the secondary structure. It is noteworthy that the Gauss linking integral depends on the orientations of segments or curves. Eliminating this orientation factor may lead to a more insightful analysis for specific tasks, regardless of curve orientation. To completely disregard orientation impact, we consider the absolute Gauss linking integral as

$$\bar{L}(l_1,l_2)=\frac{1}{4\pi}\int_{[0,1]}\int_{[0,1]}\left|\frac{\det(\dot{\gamma}_1(s),\dot{\gamma}_2(t),\gamma_1(s)-\gamma_2(t))}{|\gamma_1(s)-\gamma_2(t)|^3}\right|ds\,dt, \quad [8]$$

along with its corresponding integral segmentation matrix. The absolute Gauss linking integral of Fig. 2*B* is given in Fig. 2*C*. In the rest of this work, we use absolute Gauss linking integral in our computations.

Fig. 2 *C–H* show the absolute mGLIs at various scales from large to small. At the smallest scale (Fig. 2*H*), only the nearest neighbor interactions are recorded in Gauss linking integral. This multiscale analysis characterizes each C_α atom's local environment and interactions.

Numerous computational methods have been developed for B-factor predictions, such as Gauss network model (GNM) (28), anisotropic network model (29), normal mode analysis (NMA) (30). However, Park et al. (31) demonstrated that both GNM and NMA were ineffective in analyzing a wide range of protein structures. Their findings revealed that, on average, the correlation coefficients for GNM and NMA, across three protein sets categorized by size (small, medium, and large), were consistently below 0.6 and 0.5, respectively. Recently, advanced methods have emerged to address this challenge, including flexibility rigidity index-based approaches such as pfFRI (32) and opFRI (32), as well as topology-based methods like atom-specific persistent homology (ASPH) (33) and evolutionary homology (EH) (34).

To evaluate the performance of the proposed mGLI for protein flexibility analysis, we employed a dataset consisting of 364 protein structures, sourced from ref. 32. This dataset served as a benchmark for comparing mGLI against established methods, specifically opFRI (32), pfFRI (32), and GNM (31).

In *SI Appendix*, Table S11, we present the comparative results of mGLI with previous methods for each protein in the dataset. Remarkably, mGLI outperformed previous methods in 320 out of 364 proteins. On average, mGLI achieved the highest correlation coefficient of 0.725, surpassing the values of 0.673 for opFRI, 0.626 for pfFRI, and 0.565 for GNM, as illustrated in Fig. 2*I*. This represents a significant improvement of 7.7%, 15.8%, and 28.3%, respectively.

In addition, to validate the effectiveness of mGLI for predicting C_α atom B-factors in proteins of different sizes, we compared our method with previous approaches including EH (34), ASPH (33), opFRI (32), pfFRI (32), GNM (31), and NMA (31) on three protein sets, as shown in Fig. 2*J*. mGLI achieved average correlation coefficients of 0.899, 0.776, and 0.708 for the small, medium, and large protein sets, respectively. Our results on the

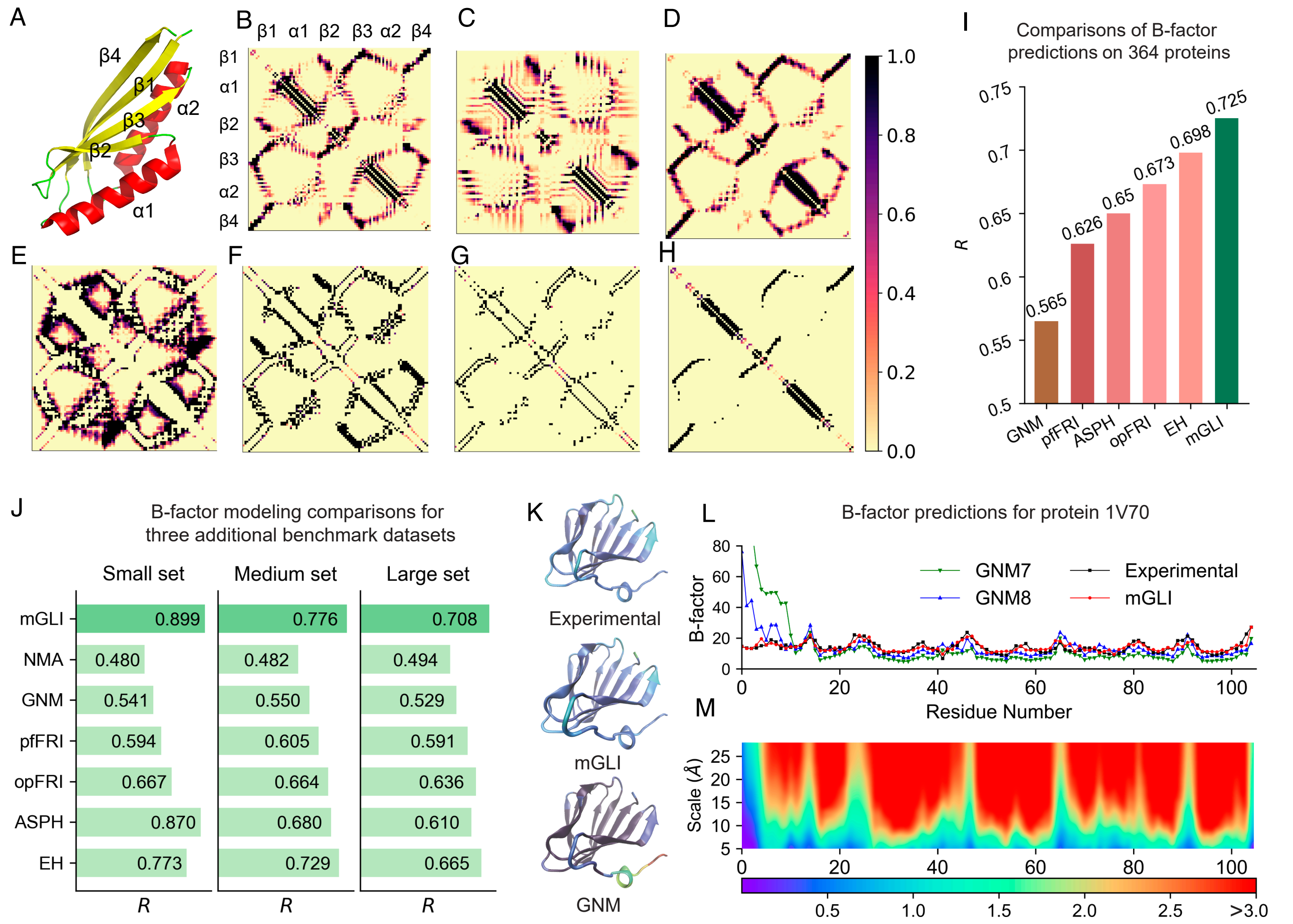

Fig. 2. An illustration of mBLI analysis for protein B-factor predictions. (*A*) The 3D structure of protein 1J27 consisting of two α-helices and four β-sheets. (*B*) The segmentation of the Gauss linking integral of protein 1J27. (*C*) The absolute value of Gauss linking integral matrix of protein 1J27. (*D*) The absolute Gauss linking integral matrix of protein 1J27. (*E*–*H*) Absolute Gauss linking integral matrices of protein 1J27 at different scales. (*I*) The comparison of B-factor predictions between our mGLI method and other literature approaches on a benchmark dataset of 364 proteins. (*J*) The comparison of B-factor predictions on three additional benchmark datasets between our mGLI method and other literature approaches (refer to *SI Appendix*, Table S2 for detailed information). (*K*) The visualization of protein 1J27 B-factors obtained from experiments, mGLI, and GNM (28). (*L*) Comparison of protein 1J27 B-factors obtained from experiments, mGLI, and GNM (28). Here, GNM7 and GNM8 indicate the cutoff value at 7 Å and 8 Å for the GNM. The *x*-axis represents the residue number, and the *y*-axis represents the B-factor value. (*M*) The visualization of mGLI features with the maximal cutoff at 30Å. The *x*-axis represents the residue number and the *y*-axis represents the scale range. Note that all values that exceed 3.0 are labeled as red.

three datasets significantly outperformed the previous methods, demonstrating improvements of 16.3%, 6.4%, and 6.5% on the small, medium, and large protein sets, respectively, compared to the previous state-of-the-art method EH (34).

To understand mGLI's performance, we present a case study with a potential antibiotic synthesis protein (PDBID: 1V70) 105 residues. Fig. 2*K* shows the protein colored with B-factor values. Apparently, mGLI-predicted B-factor values are very close to those of the experimental ones, whereas, GNM predicted values are unmatched. Fig. 2*L* presents detailed comparison. GNM methods have large errors around residues 1 to 10, which can also be seen in Fig. 2*K*. In contrast, mGLI gives accurate B-factor prediction for these residues. The mGLI features are presented in Fig. 2*M*. For each scale, we calculate the cumulative absolute Gauss linking integral, represented by a colored bar along with its accumulated value below. We designate the values exceeding a specific threshold (3.0 in this case) as red. Consequently, it becomes evident that the pattern of mGLI values in Fig. 2*M* matches the experimental B-factors in Fig. 2*L* directly. This observation holds true in a broader sense and is further validated in *SI Appendix*, Figs. S6 and S7.

Protein–ligand binding affinity predictions. Protein–ligand binding affinity describes the interaction strength between a potential drug molecule and its target protein or receptor, and its prediction plays a crucial role in drug design and discovery (35, 36). The development of machine learning models for protein–ligand binding affinity prediction represents a pivotal advancement in computational biology (37). We explore the utility of mGLI for machine learning predictive models. The PDBbind database (38) offers a comprehensive repository of protein–ligand complex structures along with their corresponding binding affinity data (36). In our study, we have included two of the most commonly utilized protein–ligand databases, namely, PDBbind-v2013 and PDBbind-v2016 (39). It is challenging to improve performance on these datasets as they have been studied by numerous researchers. The detailed information for the two datasets and related rigorous training-test splittings can be found in *SI Appendix*, Table S1.

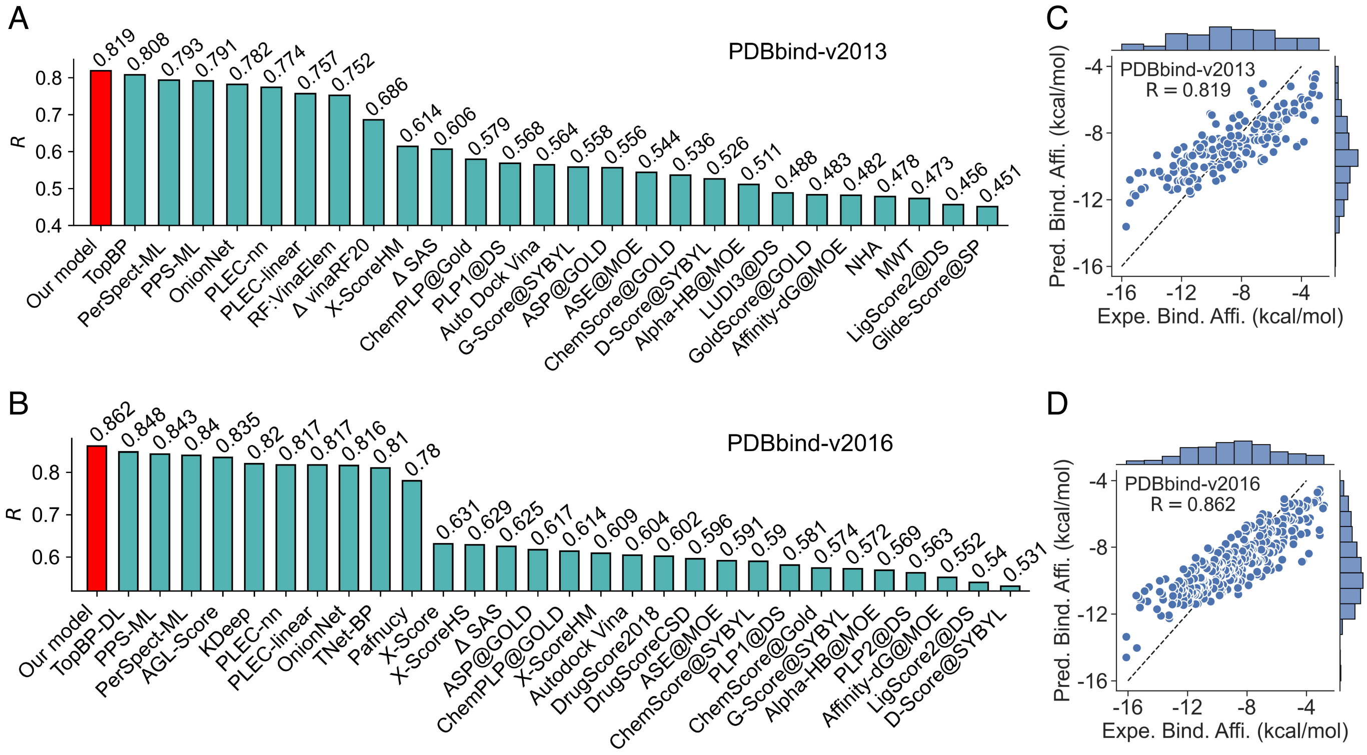

Fig. 3. The performance summary of our mGLI-assisted machine learning predictions for two PDBbind datasets. (*A* and *B*) The Pearson correlation coefficient (*R*) comparison for the binding affinity predictions of PDBbind-v2013 and PDBbind-v2016 core sets. Our models outperform other state-of-art methods (refer to *SI Appendix*, Table S4 for detailed information). (*C* and *D*) The comparison between the experimental binding affinity (BA) and the predicted BA from our best models across the two PDBbind datasets.

In *Methods*, we propose two mGLI featurization approaches on two distinct scale intervals $[r_t, r_{t+1}]$ or $[0, r_{t+1}]$, on which localized scaled Gauss linking integral is given. We use notations mGLI-bin and mGLI-all to indicate the protein–ligand complex features and mGLI-lig-bin and mGLI-lig-all to indicate two sets of ligand features. The mGLI-lig-all features can be used as additional features for protein–ligand interactions. We also utilize pretrained natural language processing (NLP) models, i.e., transformer features (TF), to complement mGLI features (see details in *Methods*). Gradient boosting decision algorithm is used for the predictions. Given a training dataset, models are built 20 times with different random seeds to address initialization-related errors. The median of Pearson correlation coefficient (*R*) values from the 20 experiments are reported below.

Fig. 3*A* illustrates the comparison of Pearson correlation coefficients (*R*) obtained from our model and the literature ones. Our mGLI-assisted model outperforms existing models for the two PDBbind datasets. The *R* values of 0.819 and 0.862, are achieved by our models in modeling PDBbind-2013 and PDBbind-2016, respectively, and are the highest values ever reported in the literature. This highlights our model's superiority and establishes it as a state-of-the-art protein–ligand binding affinity prediction model. Notably, our model demonstrates a significant improvement in *R* values in modeling the PDBbind-v2013 and PDBbind-v2016 datasets compared to others. The PDBbind-v2013 and PDBbind-v2016 datasets contain 2,764 and 3,767 complexes, respectively.

Persistent homology (40) and persistent spectral theories (41–43) give rise to competitive molecular representation and are widely utilized for molecular properties predictions. For example, TopBP (40), PerSpect-ML (42), and PPS-ML (43) rank among the top-performing models in binding affinity prediction, as demonstrated in Fig. 3*A*. The efficacy of these models can be further augmented when additional physical information is integrated. For instance, the average *R* value of PerSpect-ML (42) across the two datasets increased from 0.806 to 0.817, while that of PPS-ML (43) increased from 0.804 to 0.817. Our mGLI-assisted models, which are based on mGLI-all&mGLI-lig-all or mGLI-bin&mGLI-lig-all features, provide accurate predictions across the two PDBbind datasets, as shown in *SI Appendix*, Table S3. The symbol "&" denotes feature concatenation. The average *R* values of the two mGLI-based models across the two PDBbind datasets are 0.814 and 0.818. The best consensus models, formed by averaging predictions from mGLI-all&mGLI-lig-all or mGLI-bin&mGLI-lig-all feature-based models along with the transformer feature-based models further enhance the modeling performance, achieving an average *R* value of 0.838 and 0.841 across the two PDBbind datasets. This exceeds the average *R* of 0.835 obtained from persistent homology (40), as well as the averages of 0.817 from PerSpect-ML (42) and 0.817 from PPS-ML (43).

Fig. 3*B* offer visualization comparison between the experimental and predicted binding affinities generated by our best models for the two PDBbind datasets. The details of our models are provided in *SI Appendix*, Table S3.

hERG blockade classification predictions. Ligand-based virtual screening plays a significant role in drug discovery. Appropriate molecular descriptors are of vital importance for predictive accuracy. We investigate the performance of our mGLI molecular features in several ligand-based virtual screening prediction tasks. Predictions for hERG blockage are critically important in drug discovery due to the potential cardiac safety risks associated with drugs that inhibit the hERG potassium channel (44).

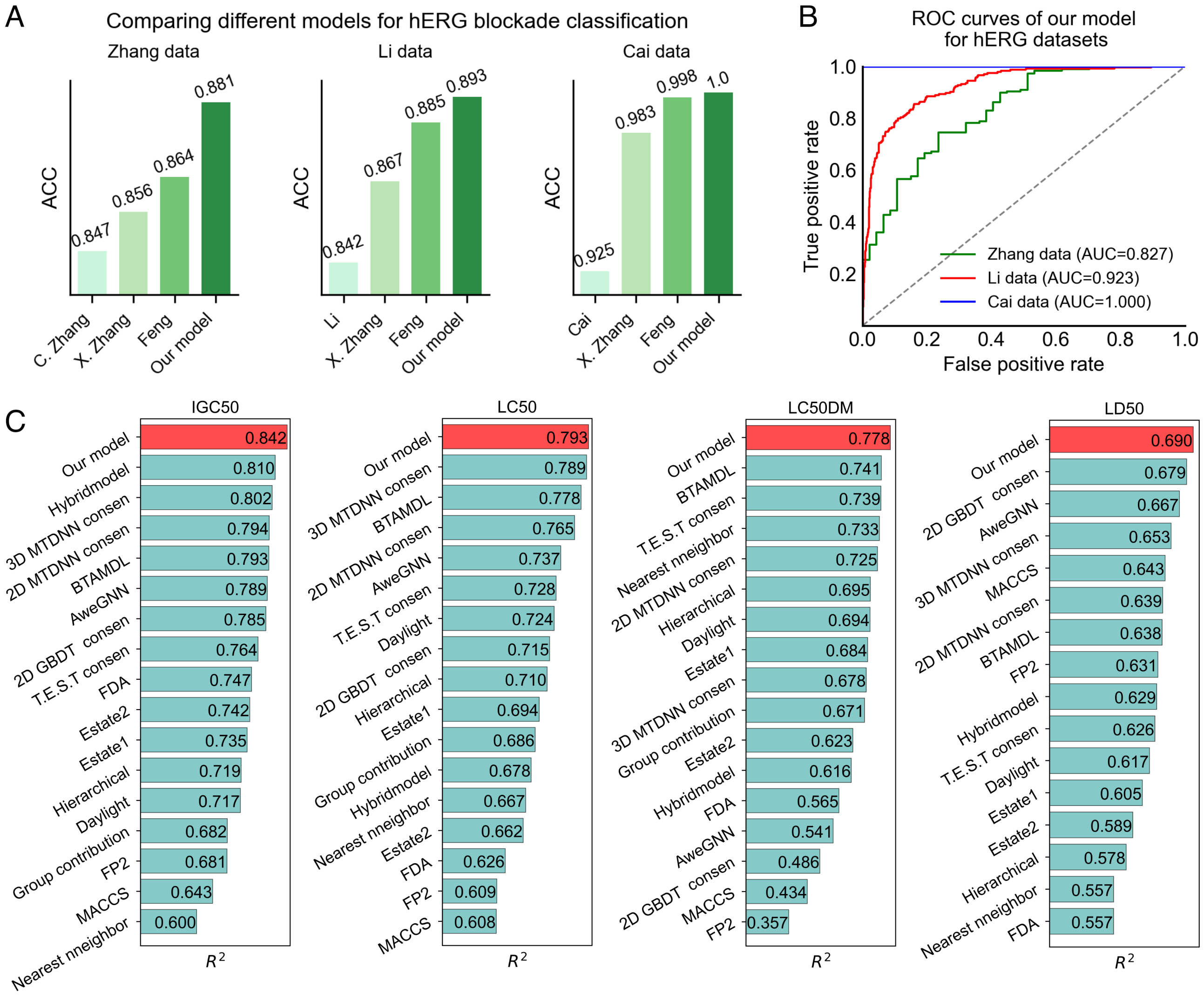

Fig. 4. The performance summary of our machine learning models for hERG blockade classification and drug toxicity predictions. (*A*) Accuracy (ACC) comparisons of our mGLI-assisted consensus model with literature models. These comparisons indicate that our model represents the state-of-the-art machine learning predictive tool. (*B*) ROC curves of our model for four hERG blockade classification tasks. (*C*) Prediction comparisons of our model with literature models for the four toxicity datasets in terms of the squared Pearson correlation coefficient (R^2) (Refer to *SI Appendix*, Table S7 for detailed comparative information).

Several machine learning predictive models are available in the literature (44–48), and we benchmark our mGLI-based models against them. Among these models, the persistent Laplacian theory (41, 44) was used in conjunction with several NLP molecular embeddings (26, 49) to build predictive models, yielding the best hERG blockade prediction model. The persistent Laplacian approach, rooted in spectral graph theory, can be regarded as an extension of persistent homology theory. It preserves the topological persistence as persistent homology, while revealing additional geometric insights from those nonharmonic portions of the spectrum. We provide a detailed discussion of these two theories in *SI Appendix*, section 7. Here, we employ mGLI theory alongside several other molecular descriptors, including the same two NLP embeddings as in ref. 44, and algebraic graph (AG)-based molecular features (50). The NLP embeddings are paired with artificial neural network algorithms, while mGLI and AG features are used with GBDT algorithms. Our final prediction model is obtained with the consensus prediction of these four models.

Three hERG blockade datasets with binary classification labels from the literature were used to investigate the performance of our models. Details of these datasets and five utilized evaluation metrics including area under the curve (ACC), Accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity, and specificity are included in *SI Appendix*, Table S1 and section 1. Among these metrics, ACC gives the percentage of the correctly predicted blockers and nonblockers. Given a training dataset, each individual model was built ten times with different random seeds. In the comparison with other literature models, the highest ACC scores, along with corresponding metrics evaluations from the ten prediction results, are reported in *SI Appendix*, Table S5. Our models yield state-of-the-art predictions. Fig. 4*A* displays the ACC score comparisons across the three datasets, while the comparison in terms of AUC and MCC is displayed in *SI Appendix*, Fig. S12. Fig. 4*B* exhibits the receiver operating characteristic (ROC) curves of our model in predicting the test sets of the three datasets.

Zhang et al. (45) investigated their model performance with a hERG dataset containing 1,163 compounds. Different training and test sets were partitioned from the 1,163 compounds. Various thresholds defined by IC_{50} values were used to discriminate hERG blockers from nonblockers. Their support vector machine model had the best ACC scores of 0.848 on the test set with threshold of 30 μM. Zhang et al.'s model (46) had a boosted prediction ACC score of 0.856. Feng et al.'s model (44) achieved much higher improvement in many metric. Our model has significantly higher predictive power than Feng et al.'s model (44) with ACC scores increased from 0.864 to 0.881, and MCC results boosted from 0.518 to 0.587, respectively, while it also achieved high sensitivity and specificity scores.

Li et al. (47) constructed two consensus models based on their dataset composed of 3,721 compounds with a threshold of IC_{50} equals to 1 μM classifying blockers and nonblockers. Their best consensus results on a test set of 1,092 compounds achieved an ACC score of 0.842. Feng et al.'s model (44) improved the results of Li et al. (47) and Zhang et al.(46). The AUC, ACC, and MCC scores of our mGLI-assisted model are 0.924, 0.893, and 0.661, which are even higher than the corresponding scores of 0.917, 0.885, and 0.629 in Feng et al.'s model (44).

Cai et al. (48) developed a multitask deep neural network-based model and had their best predictive power on a hERG dataset with blockade threshold value of 80 μM. The reported AUC and ACC scores achieved 0.967 and 0.925. Feng et al.'s (44) model had boosted performance. Our model accomplished perfect scores of 1.000 in all the five evaluation metrics. The detailed performance of our individual models is provided in *SI Appendix*, Table S6 or Fig. S13. The mGLI models outperform or achieve comparable results. This indicates the critical impact of mGLI modeling on the resulting consensus predictions. Our model consistently exhibits outstanding predictive performance, placing it among the top-tier machine learning models for hERG blocker/nonblocker classification.

Quantitative toxicity predictions. Toxicity in drug discovery refers to the potential harmful effects or adverse reactions that a drug or chemical compound may have on living organisms (51). Assessing drug toxicity is essential in drug discovery. We assess the performance of our mGLI-assisted predictive models on four toxicity datasets, including IGC50, LC50, LC50DM, and LD50. Information about the toxicity datasets is provided in *SI Appendix*, Table S1 and section 5B.

In addition to mGLI, we also employ transformer (TF) (49) and autoencoder (AE) models (26) to enhance the modeling performance. We pair GBDT with mGLI features to model the four datasets. Due to the similarity of the toxicity datasets, a MTDNN was employed to enhance modeling performance (25, 51, 52). We employed TF and AE features to build two MTDNN models, resulting in two additional sets of predictions. Our final predictive model is obtained by averaging these three sets of predictions. Given a training dataset, models are built 10 times with random seeds.

SI Appendix, Table S7 presents the detailed comparison in terms of squared Pearson correlation coefficients (R^2) and root mean squared error. The comparisons in terms of R^2 are depicted in Fig. 4*B*. Our model stands out in toxicity predictions, achieving the higher R^2 values of 0.842, 0.793, 0.778, and 0.690 for the IGC50, LC50, LC50DM, and LD50 datasets, respectively. *SI Appendix*, Fig. S16 presents a comparison between the experimental toxicity and our predicted toxicity values for the four datasets. The high consistency underscores the effectiveness of our machine learning models.

Two competitive models were proposed by Gao et al. (52), namely the 2D-GBDT and 2D-MTDNN consensus models, which utilize traditional 2D molecular fingerprints along with various machine learning algorithms. Their multitask learning consensus model achieved R^2 values of 0.794, 0.765, 0.725, and 0.639 for the IGC50, LC50, LC50DM, and LD50 datasets, respectively. They surpassed many other models in the literature, including those from Toxicity Estimation Software Tool (T.E.S.T) and related approaches, such as hierarchical, FDA, nearest neighbor, and T.E.S.T consensus (53). Wu et al. (51) introduced molecular fingerprints using persistent homology theory and developed a consensus multitask learning model. Additional molecular descriptors based on physical attributes, including energy, surface energy, and electric charge, were incorporated into their consensus model, significantly enhancing predictive performance. Their model achieved R^2 values of 0.802, 0.789, 0.678, and 0.653 for the aforementioned datasets. Our model outperforms these exceptional models. Several other models have recently been developed based on traditional molecular fingerprints such as estate1, estate2, daylight MACCS, or other advanced strategies. However, our model outperforms them by a significant margin, as observed in Fig. 4, and detailed comparisons are provided in *SI Appendix*, Table S7. This demonstrates that our mGLI-based knot theory provides an effective approach for molecular representation learning.

In addition, *SI Appendix*, Table S7 or Fig. S15 displays the detailed performance results of our GBDT and MTDNN models. We compared the mGLI-based GBDT model with GBDT models based on TF or AE features. The mGLI-GBDT model is competitive across the four prediction tasks, outperforming the TF-GBDT model in all tasks except for LC50DM. The inferior performance for the LC50DM task can be primarily attributed to overfitting issues. The large number of features in the mGLI model makes it less suitable for the LC50DM dataset, whose training set only has 283 molecules. The comparisons indicate that mGLI provides valuable 3D structure-based features for small molecule representations compared to NLP molecular features and is competitive in modeling individual tasks.

Discussion

Generalization to Other Topological Objects and Real-World Structures. It is intriguing to consider the range of data to which the present KDA can be applied. Mathematically, the multiscale Gauss link integral theory proposed in this work can naturally extend to a wide variety of other topological objects, such as knotoids (54), links, linkoids (55), lassos (56), and cysteine knots (57) in Fig. 5*A*, as well as curve segments in Fig. 5 *B* and *C*, tangles, and braids. These types of curved structures are ubiquitous in real-world objects, ranging from ropes, shoelaces, highways, and powerline networks to polymers, DNA, RNA, nucleosomes, chromosomes, and the trajectories of space vehicles and interceptor missiles. In a comparative analysis, our KDA deals with curved data, whereas TDA handles point cloud data defined on simplicial complexes, graphs, hypergraphs, etc. Additionally, our earlier persistent Hodge Laplacian is defined on manifolds and addresses volumetric data (58).

Curve Segment Size and Multiscale Granularity. In principle, our method allows for the arbitrary combination of curve segmentation with any multiscale schemes. However, in practical applications, the performance of mGLI is highly dependent on their selection. First and foremost, the values of the Gauss linking integral of a local curve segment depend not only on their spatial alignment but also on their relative lengths compared to the global curve. When the length of a curve segment approaches zero, the corresponding Gauss linking integral approximates to 0. Similarly, as curve segments expand to cover the global curve, the Gauss linking integral returns global information.

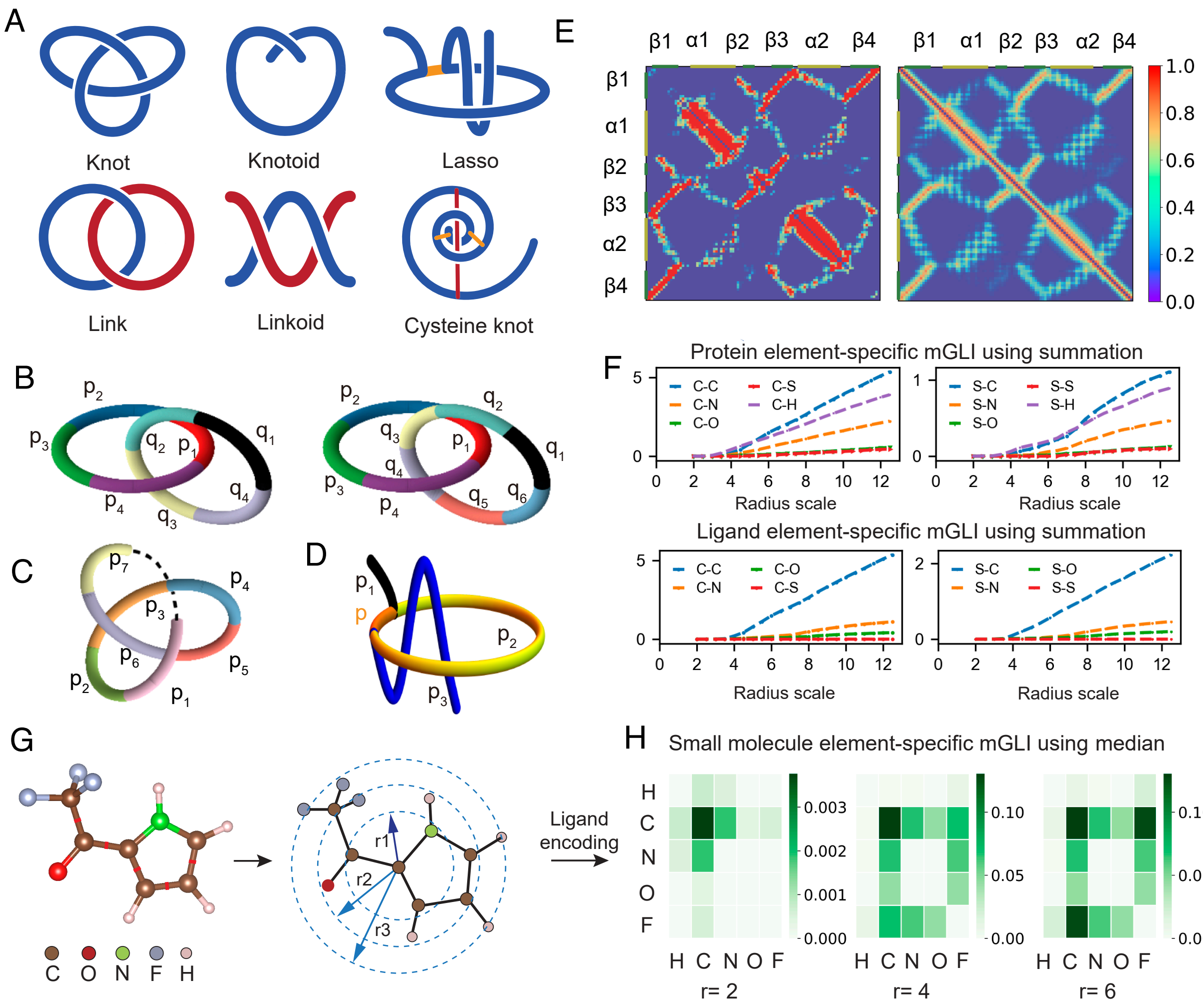

Fig. 5. (*A*) Examples of topological objects which can be studied by the multiscale Gauss linking integral. (*B*) Hopf link with two types of curve segmentations. (*C*) Slipknot with seven curve segments. (*D*) Lasso with four curve segments. (*E*) *Left* is the absolute Gauss linking integral matrix for protein 1J27. *Right* is the transient probability matrix (TPM) for protein 1J27. Points in *Top* row and *Left* column are colored green or yellow, denoting β-sheet or α-helix of 1J27. (*F*) The protein or ligand element-specific mGLI features based on summation statistics for protein 1PXO, as formulated in Eq. **16**. Additional features for other element-specific cases are offered in *SI Appendix*, Fig. S2, while features based on median statistics are provided in *SI Appendix*, Fig. S3. (*G*) The curve segmentation illustration of molecule 2-Trifluroacetyl along with radius scales centered at each atom. (*H*) The feature of element-specific mGLI under three scales for the molecule using median statistics, as formulated in Eq. **16**. The magnitude of feature values increases as the scales increase. Features with alternative statistics measures for element-specific mGLI features are presented in *SI Appendix*, Figs. S4 and S5.

In both cases, the Gauss linking integral fails to extract useful spatial information regarding local alignments. The choice of segmentation depends on the specific application. For example, in dealing with molecular properties, atomic segments are needed. In modeling a crowded highway, the segment of individual car size is a natural choice. Second, the selection of the multiscale range impacts the featurization of the Gauss linking integral. Ideally, different scales should capture distinct spatial structure information, and the choice of scales should reflect important interactions in the data. If the information between different scales is negligible, it can result in a large number of identical or trivial features. Conversely, if the scale is too coarse, it may lead to information loss.

The Superiority of mGLI for Biomolecular Data. Proteins, DNA, and RNA are polymers and are naturally modeled as curved structures at certain scales. The proposed multiscale Gauss linking integral proves to be a superior tool for biomolecular data analysis compared to previous methods. The analysis of biomolecular structures using mGLI can lead to insights. To demonstrate this, we conducted a structural analysis of protein 1J27 by segmenting the absolute multiscale Gauss linking integral and compared it with the previous transient probability matrix (TPM) (59). The structural information that was previously obscured becomes considerably more evident and clear when using mGLI, as depicted in Fig. 5*E*. For instance, in the TPM, interactions such as $\alpha 1$-$\alpha 1$ and $\alpha 2$-$\alpha 2$ are represented as slightly thicker yellow blocks along the diagonal. In contrast, mGLI portrays these interactions as larger, more expressive, and prominently red blocks. This enhanced visualization enables a more precise distinction between the self-interaction of the alpha chain and other structural elements, such as the self-interaction between the $\beta 2$ and $\beta 3$ regions. Furthermore, the contrast between different values within each block is more pronounced in mGLI compared to TPM. This distinction is particularly noticeable in blocks representing interactions like $\alpha 1$-$\alpha 2$, $\beta 1$-$\alpha 2$, $\beta 1$-$\beta 2$, and so forth.

Topological Data Analysis vs. Knot Data Analysis. Recent years have witnessed the rapid growth of TDA in data science, driving its success in various domains, particularly in computational biology (17–19). However, the major tool of TDA, persistent homology, has many drawbacks (20), including its qualitative and global nature, as well as the lack of localization. It is imperative to develop new mathematical/topological methods that overcome the drawbacks of TDA and potentially impact various domains of data science.

The proposed mGLI is a local method but recovers global topological properties at sufficiently large scales. Therefore, mGLI-based KDA models can outperform TDA models, as shown in this work. A direct comparison of TDA and KDA in protein B-factor prediction shows that KDA has a 17.2% improvement over TDA as sown in Fig. 2*I* (ASPH vs. mGLI). Besides, our mGLI models demonstrate superiority over TDA models (42, 43) for predicting protein–ligand binding affinity. Our model, based on mGLI features, achieves an average *R* value of 0.818 across the two PDBbind datasets. This surpasses the *R* values of 0.806 from PerSpect-ML (42) and 0.804 from PPS-ML (43) as well. The proposed KDA is computationally efficient, as it takes only a few minutes on a personal computer to generate mGLI features for a moderately sized dataset. Recently, a new KDA tool, persistent Khovanov homology, has also been reported (60). Given the tremendous success of TDA, we expect that KDA will become a powerful new topological learning tool for a wide variety of problems in data science.

Methods

Multiscale Gauss Linking Integral. We introduced several essential definitions related to the Gauss linking integral in *Results*. Additional important proposition or theorems are presented below.

Proposition 1. *The Gauss linking integral in Eq.* **1** *is identical to the average of half the algebraic sum of intercrossings in the projection of the two curves in any possible projection direction for both open and closed curves.*

Theorem 1. [Panagiotou et al. (12)]. *For closed curves, the Gauss linking integral is an integer and a topological invariant. For open curves, the Gauss linking integral is a real number and a continuous function of curve coordinates.*

Theorem 2. [The grand sum of the segmentation matrix]. *The grand sum of the segmentation matrix of two curves equals the Gauss linking integral of the original curves:*

$$L(l_1, l_2) = \sum_i \sum_j L(p_i, p_j). \qquad [9]$$

Remark 4 **[Generalization of Gauss linking integral]:** Vassiliev measure, a generalization of Gauss linking integral, can be applied to open and closed curves in 3-space (55). Similarly, the proposed mGLI obtained by combining the Gauss linking integral and multiscale process can naturally be applied to links, linkoids, open and closed curves, and other segmentable objects as shown in Fig. 5*B*. It can be noticed that any element in the segmentation of the Gauss linking integral is defined on local curve segments. This indicates that one can define a generalized form of the multiscale Gauss linking integral if the segmentation of the Gauss linking integral is well defined on local curve segments. In fact, for any topological or geometric structure that can be segmented into curve segments P_n, Q_m, we can define the following segmentation matrix:

$$\bar{G} = \begin{pmatrix} g(p_1, q_1) & g(p_1, q_2) & \cdots & g(p_1, q_m) \\ g(p_2, q_1) & g(p_2, q_2) & \cdots & g(p_2, q_m) \\ \vdots & \vdots & \ddots & \vdots \\ g(p_n, q_1) & g(p_n, q_2) & \cdots & g(p_n, q_m) \end{pmatrix}, \qquad [10]$$

where

$$g(p_i, q_j) = \begin{cases} L(p_i, q_j) & \text{if } p_i \cap q_j \text{ is a null-set,} \\ 0 & \text{else.} \end{cases} \qquad [11]$$

In the above definition, unlike in Eq. **2**, the curve segments in P_n and Q_m are allowed to intersect or even be equal. Thus, the mGLI can be applied in multiple topological/geometric structures as long as they can locally be represented as curve segments. Featurization can be similarly derived as in Eq. **7**.

mGLI Featurization for B-Factor Prediction. We consider a protein as an open curve, acknowledging that the polypeptide chain of a protein molecule can be seen as an open polygon l whose vertices are corresponding to the C_α atoms, while the edges represent the pseudobonds that connect a C_α atom to another one in an adjacent amino acid residue. We propose a curve segmentation induced by C_α atoms:

$$p_i = \left\{ x \in l_1 | f(x, c_i) = \inf_{c \in C} f(x, c) \right\}, 1 \le i \le n, \qquad [12]$$

where $f(a, b)$ is the distance of points a and b along l, c_i is the 3D coordinates of a C_α atom, and C is the set of C_α atoms. Then, the $d(p_i, q_j)$ assumed in Eq. **3** can be defined:

$$d(p_i, q_j) = d_E(c_i, c_j), \qquad [13]$$

where d_E is the Euclidean distance in the 3D space.

Then, according to the generalized multiscale Guass linking integral, the segmentation of Gauss linking integral that investigates the intercrossings between segments of the protein can be given:

$$G = \begin{pmatrix} L(p_1, p_1) & L(p_1, p_2) & \cdots & L(p_1, p_n) \\ L(p_2, p_1) & L(p_2, p_2) & \cdots & L(p_2, p_n) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, p_1) & L(p_n, p_2) & \cdots & L(p_n, p_n) \end{pmatrix} = \begin{pmatrix} 0 & L(p_1, p_2) & \cdots & L(p_1, p_n) \\ L(p_2, p_1) & 0 & \cdots & L(p_2, p_n) \\ \vdots & \vdots & \ddots & \vdots \\ L(p_n, p_1) & L(p_n, p_2) & \cdots & 0 \end{pmatrix}.$$

The localized scaled Gauss linking integral, detailed in *Remark* 3, is a natural way to characterize each C_α atom in B-factor predictions. We naturally choose a segment that precisely covers a single C_α atom along the polymer chain. Additionally, in our study, the multiscale scheme is selected to start from 5 Å and extend up to 17 Å, with each scale interval set at 1 Å. This choice is based on the fact that the average distance between C_α atoms is approximately 3.8 Å. Such a selection of the multiscale scheme results in a powerful featurization method that provides abundant representations of local protein structures.

Traditional B-factor analysis methods predominantly concentrate on individual atoms and their spatial positions in three-dimensional space, accounting for the thermal motion and disorder of atoms within a protein structure. However, the incorporation of bonding interactions between atoms, which indirectly impacts the observed B-factor values, is rarely employed in B-factor analysis. Through the incorporation of mGLI, our method introduces the notion of pseudobonds between protein atoms, effectively capturing the influence of bonding interactions. The integration of knot theory with the multiscale procedure enables the precise localization of measurements, capitalizing on the spatial positions and atomic environments. The synergy between multiscale analysis and knot theory culminates in a robust method for predicting protein B-factors, showcasing the potential of multiscale approaches in effectively pinpointing measurements derived from knot theory.

mGLI Featurization for Protein–Ligand Complex. Localized scaled Gauss linking integral is also utilized to characterize protein–ligand interactions. This approach defines distinct curve segments and computes integrals with other segments across various scales. For molecular structures, we adopt atom-specific curve segmentation. Each atom c_j in a protein or ligand molecule is linked by multiple covalent bonds to neighboring atoms, determining the curve segmentation specific to c_j. These segments originate from the central atom and extend to the midpoint of associated covalent bonds, resulting in atom-specific curve segmentation.

$$p_i = \left\{ x \in l | f(x, c_i) \leq \frac{1}{2} f(c, c_i), c \in C \right\}, \qquad [\mathbf{14}]$$

Here, C represents the set of adjacent atoms connected to atom c_i by covalent bonds, and l denotes the straight line along each covalent bond.

We focus on the binding core region where protein–ligand interactions primarily occur, extracting protein atoms within a 12 Å cutoff distance from the ligand. We can obtain atom-specific curve segmentations for both the protein and ligand. Using these segmentations (p_i in protein and q_j in ligand), we compute atom-by-atom Gauss linking integrals (a-GLI) $L(p_i, q_j)$. Multiple segment pairs between the two atoms may exist, resulting in numerous Gauss linking integral between a segment pair. We consider the absolute Gauss linking integrals to mitigate curve orientation effects. Due to the multiple integrals between pairs, we utilize statistical analysis, specifically median and SD, to define $L(p_i, q_j)$.

Element-specific approach is used in designing mGLI protein–ligand features. Specifically, we primarily focus on the protein atom groups of four elements (C, N, O, and S) within the protein, while considering atom groups of ten elements (C, N, O, H, S, P, F, Cl, Br, and I) within the ligand. We extract these atom groups in the core binding region, and then apply mGLI to characterize pairwise interactions between these atom groups from the protein and ligand.

Let P_n^C and Q_m^N represent collections of carbon (C) atom-specific curve segmentations in the protein and nitrogen (N) atom-specific curve segmentations in the ligand, respectively, given by $P_n^C = \{p_i^C | i = 1, 2, \cdots, n\}$ and $Q_m^N = \{q_j^N | j = 1, 2, \cdots, m\}$. We use the two groups to illustrate element-specific mGLI for protein–ligand featurization. The atomic coordinates in the two groups are labeled as $\{\mathbf{r}_i^C | i = 1, 2, \cdots, n\}$ and $\{\mathbf{r}_j^N | j = 1, 2, \cdots, m\}$. With the atom-by-atom Gauss linking integral $L(p_i^C, q_j^N)$ defined, we further determine the multiscale element-by-element Gauss linking integral. Assuming a scale $R = \{r_0, r_1, r_2, r_3, \cdots, r_k\}$, where $0 = r_0 < r_1 < r_2 < \cdots < r_k$, the distance between p_i^C and q_j^N is denoted as $d(p_i^C, q_j^N) = d_E(\mathbf{r}_i^C, \mathbf{r}_j^N)$ (in Å), where $d_E(\cdot, \cdot)$ indicates the Euclidean distance. The scaled Gauss linking integral $G^{r_t, r_{t+1}}$ in Eq. **3** for curve segments generalizes to atom-by-atom Gauss linking integral. Atom-specific localized scaled Gauss linking integrals between two atom groups can be similarly derived as in Eqs. **5** and **6**:

$$J^{r_t, r_{t+1}}(p_i^C, Q_m^N) = \sum_{s=1}^{m} G_{is}^{r_t, r_{t+1}},$$

$$J^{r_t, r_{t+1}}(q_j^N, P_n^C) = \sum_{s=1}^{n} G_{sj}^{r_t, r_{t+1}},$$

where the second variable in $J^{r_t, r_{t+1}}$ indicates linking atom sets with the specified atom in the first variable. These expressions quantify the intercrossing between a C atom-specific segmentation p_i^C in the protein and a set of C atom-specific segmentations in the ligand within a given scale from r_t to r_{t+1}, or between a N atom-specific segmentation q_i^N in the ligand and a set of C atom-specific segmentations in the protein within a given scale.

To provide a scalable description of atomic interactions between two atom groups, we compute all atom-specific localized scaled Gauss linking integrals $J^{r_t, r_{t+1}}(p_i^C, Q_m^N)$ for $i = 1, 2, \cdots, n$, and $J^{r_t, r_{t+1}}(q_j^N, P_n^C)$ for $j = 1, 2, \cdots, m$. Statistical measures are then used to determine the multiscale element-specific Gauss linking integral (e-GLI) through the following formulations:

$$\begin{aligned} &J^{r_t, r_{t+1}}(P_n^C, Q_m^N) = \text{statistics of} \\ &\left\{ J^{r_t, r_{t+1}}(p_1^C, Q_m^N), J^{r_t, r_{t+1}}(p_2^C, Q_m^N), \cdots, J^{r_t, r_{t+1}}(p_n^C, Q_m^N) \right\}, \\ &J^{r_t, r_{t+1}}(Q_m^N, P_n^C) = \text{statistics of} \\ &\left\{ J^{r_t, r_{t+1}}(q_1^N, P_n^C), J^{r_t, r_{t+1}}(q_2^N, P_n^C), \cdots, J^{r_t, r_{t+1}}(q_m^N, P_n^C) \right\} \end{aligned} \qquad [\mathbf{15}]$$

We employ various statistical measures such as sum, minimum, maximum, mean, and median in Eq. **15**, which depict the atomic interactions between C atom-specific segmentations in the protein and N atom-specific segmentations in the ligand within the scale $[r_t, r_{t+1}]$. We consider the two formulations in Eq. **15** as protein and ligand element-specific Gauss linking integral, respectively.

We can extend the starting point of the scale interval to 0, giving rise to following formulation:

$$\begin{aligned} &J^{0, r_{t+1}}(P_n^C, Q_m^N) = \text{statistics of} \\ &\left\{ J^{0, r_{t+1}}(p_1^C, Q_m^N), J^{0, r_{t+1}}(p_2^C, Q_m^N), \cdots, J^{0, r_{t+1}}(p_n^C, Q_m^N) \right\}, \\ &J^{0, r_{t+1}}(Q_m^N, P_n^C) = \text{statistics of} \\ &\left\{ J^{0, r_{t+1}}(q_1^N, P_n^C), J^{0, r_{t+1}}(q_2^N, P_n^C), \cdots, J^{0, r_{t+1}}(q_m^N, P_n^C) \right\} \end{aligned} \qquad [\mathbf{16}]$$

We refer to the first and second approaches as mGLI-bin and mGLI-all featurization, respectively. In characterizing protein–ligand complexes, we define the scale radius set as $R = \{0, 2, 3, \cdots, 11, 12\}$ (in Å). Each of these featurization approaches results in an mGLI feature vector with a length of 40 (number of element combinations) × 2 (e-GLI fro two formulations in Eq. **15**) × 11 (scale number) × 5 (statistics for e-GLI) × 2 (statistics for a-GLI) = 8,800. Fig. 5 *E* and *F* give an illustration of protein and ligand element-specific mGLI features.

Fig. 5*F* illustrates a few cases of protein or ligand element-specific mGLI over the radius scales based on statistics of summation for two formulations in Eq. **16**. Additional cases are provided in *SI Appendix*, Figs. S2 and S3.

We investigate the potential improvements in modeling performance resulting from employing statistical measures for mGLI features. *SI Appendix*, Figs. S8–S10 demonstrate the effectiveness of utilizing various statistical measures. Comparative analysis in *SI Appendix*, section 4B validates the enhancement induced by incorporating additional statistical measures.

Adjusting the upper scale of protein-specific mGLI features could lead to an improvement in modeling performance. *SI Appendix*, Fig. S11 presents the resulting performance comparisons across various upper scales r_k, ranging from 12 to 20. Despite the increase in upper scales, the modeling performance remains consistent, indicating that an upper scale of 12 Å is adequate for ensuring optimal mGLI feature performance. The scale range and equal partitioning with an increment of 1 Å are appropriate for capturing local atomic interactions and recovering global molecular interactions.

mGLI Featurization for Small Molecules. The mGLI featurization for small molecules can utilize the same approach based on the aforementioned 10 atom groups. Two mGLI feature strategies for ligands are available: mGLI-bin-lig and mGLI-all-lig, depending on local integral scale ranges. For a ligand with atom-specific curve segmentations p_i and q_j, the atom-by-atom Gauss linking integral $L(p_i, q_j)$ is determined using median statistics, adhering to the element-specific strategy to capture more atomic interactions. For atom-specific curve segmentations p_i^C ($i = 1, 2, \cdots, n$) and q_j^N ($j = 1, 2, \cdots, m$), statistics including summation, minimum, maximum, mean, and median are applied to the multiscale element-specific Gauss linking integral in equations such as Eqs. **15** or **16**. The scale values are defined as $R = \{0, 2.0, 2.44, 2.98, 3.63, 4.43, 5.41, 6.59, 8.05, 10\}$ for characterizing small molecules. Both mGLI-bin-lig and mGLI-all-lig features have a length of 2,475. The upper scale of 10 Å is reasonable based on the 3D structure size of general small molecules as analyzed for hERG blockade molecules in *SI Appendix*, Fig. S14.

An illustration of the multiscale element-specific Gauss linking integral for a molecule is depicted in Fig. 5 *G* and *H*, with corresponding additional feature analysis provided in *SI Appendix*, Figs. S4 and S5.

Additional Molecular Descriptors and Machine Learning Algorithms. In this work, transformer and autoencoder-based NLP molecular descriptors are employed to enhance mGLI knot learning for various predictive tasks. Details about these descriptors are provided in *SI Appendix*, section 5C. Additionally, the integration of various molecular descriptors with machine learning and deep learning algorithms is discussed in *SI Appendix*, section 6.

Data, Materials, and Software Availability. All data and the code needed to reproduce this paper's result can be found at https://github.com/WeilabMSU/mGLI-KDA (61).

ACKNOWLEDGMENTS. This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, NSF grants DMS-2052983, DMS-1761320, DMS-2245903, and IIS-1900473, NASA grant 80NSSC21M0023, Michigan State University Research Foundation, Bristol-Myers Squibb 65109. The work Lei and Li was supported in part by grant (No.12331003) of National Natural Science Foundation of China and State Key Laboratory of Structural Analysis, Optimization and CAE Software for Industrial Equipment.

Author affiliations: [a]Department of Mathematics, Michigan State University, East Lansing, MI 48824; [b]School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China; [c]Beijing Institute of Mathematical Sciences and Applications, 101408, China; [d]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824; and [e]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824

1. C. C. Adams, *The Knot Book* (American Mathematical Soc., 1994).
2. R. H. Crowell, R. H. Fox, *Introduction to Knot Theory* (Springer Science & Business Media, 2012), vol. 57.
3. C. Manolescu, "An introduction to knot Floer homology" in *Physics and Mathematics of Link Homology*, S. Gukov, M. Khovanov, and J. Walcher Eds. (American Mathematical Society, 2014), vol. 680, pp. 99–135.
4. M. Khovanov, A categorification of the jones polynomial. *Duke Math. J.* **101**, 359–426 (2000).
5. T. Ohtsuki, *Quantum Invariants: A Study of Knots, 3-Manifolds, and their Sets* (World Scientific, 2002), vol. 29.
6. C. Liang, K. Mislow, Knots in proteins. *J. Am. Chem. Soc.* **116**, 11189–11190 (1994).
7. D. Sumners, *The Role of Knot Theory in DNA Research in Geometry and Topology* (CRC Press, 2020), pp. 297–318.
8. T. Schlick *et al.*, To knot or not to knot: Multiple conformations of the SARS-CoV-2 frameshifting RNA element. *J. Am. Chem. Soc.* **143**, 11404–11422 (2021).
9. K. C. Millett, E. J. Rawdon, A. Stasiak, J. I. Sułkowska, Identifying knots in proteins. *Biochem. Soc. Trans.* **41**, 533–537 (2013).
10. M. Jamroz *et al.*, Knotprot: A database of proteins with knots and slipknots. *Nucleic Acids Res.* **43**, D306–D314 (2015).
11. P. Dabrowski-Tumanski, P. Rubach, W. Niemyska, B. A. Gren, J. I. Sulkowska, Topoly: Python package to analyze topology of polymers. *Brief. Bioinform.* **22**, bbaa196 (2021).
12. E. Panagiotou, L. H. Kauffman, Knot polynomials of open and closed curves. *Proc. R. Soc. A* **476**, 20200124 (2020).
13. Q. Baldwin, B. Sumpter, E. Panagiotou, The local topological free energy of the SARS-CoV-2 spike protein. *Polymers* **14**, 3014 (2022).
14. H. Edelsbrunner *et al.*, Persistent homology-a survey. *Contemp. Math.* **453**, 257–282 (2008).
15. A. Zomorodian, G. Carlsson, "Computing persistent homology" in *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, J. Snoeyink and J. -D. Boissonnat., Eds. (Association for Computing Machinery, (2004), pp. 347–356.
16. Z. Cang, G. W. Wei, Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**, e1005690 (2017).
17. D. D. Nguyen *et al.*, Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges. *J. Comput. Aided Mol. Des.* **33**, 71–82 (2019).
18. R. Wang, J. Chen, G. W. Wei, Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America. *J. Phys. Chem. Lett.* **12**, 11850–11857 (2021).
19. J. Chen, G. W. Wei, Omicron BA.2 (B.1.1.529.2): High potential for becoming the next dominant variant. *J. Phys. Chem. Lett.* **13**, 3840–3849 (2022).
20. J. Chen, Y. Qiu, R. Wang, G. W. Wei, Persistent Laplacian projected omicron BA.4 and BA.5 to become new dominating variants. *Comput. Biol. Med.* **151**, 106262 (2022).
21. C. F. Gauss, "Integral formula for linking number" in *Zur mathematischen theorie der electrodynamische wirkungen*, C. F. Gauss and The Königlichen Gesellschaft der Wissenschaften zu Göttingen, Eds. (Springer Berlin Heidelberg, 1833), vol. 5, p. 605.
22. J. M. Cornwall, N. Graham, Sphalerons, knots, and dynamical compactification in Yang-Mills-Chern-Simons theories. *Phys. Rev. D* **66**, 065012 (2002).
23. M. A. Berger, Third-order link integrals. *J. Phys. A Math. Gen.* **23**, 2787 (1990).
24. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
25. D. Chen *et al.*, Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **12**, 3521 (2021).
26. R. Winter, F. Montanari, F. Noé, D. A. Clevert, Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).
27. R. L. Ricca, B. Nipoti, Gauss'linking number revisited. *JKTR* **20**, 1325–1343 (2011).
28. A. Rader, C. Chennubhotla, L. W. Yang, I. Bahar, "The gaussian network model: Theory and applications" in *Normal Mode Analysis*, Q. Cui, I. Bahar, Eds. (Chapman and Hall/CRC, 2005), pp. 65–88.
29. E. Eyal, L. W. Yang, I. Bahar, Anisotropic network model: Systematic evaluation and a new web interface. *Bioinformatics* **22**, 2619–2627 (2006).
30. I. Bahar, A. Rader, Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **15**, 586–592 (2005).
31. J. K. Park, R. Jernigan, Z. Wu, Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations. *Bull. Math. Biol.* **75**, 124–160 (2013).
32. K. Opron, K. Xia, G. W. Wei, Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *J. Chem. Phys.* **140**, 06B617-1 (2014).
33. D. Bramer, G. W. Wei, Atom-specific persistent homology and its application to protein flexibility analysis. *Comput. Math. Biophys.* **8**, 1–35 (2020).
34. Z. Cang, E. Munch, G. W. Wei, Evolutionary homology on coupled dynamical systems with applications to protein flexibility analysis. *J. Appl. Comput. Topol.* **4**, 481–507 (2020).
35. H. Cai *et al.*, Carsidock: A deep learning paradigm for accurate protein-ligand docking and screening based on large-scale pre-training. *Chem. Sci.* **15**, 1449–1471 (2024).
36. Q. U. Ain, A. Aleksandrova, F. D. Roessler, P. J. Ballester, Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).
37. X. Pan *et al.*, Aa-score: A new scoring function based on amino acid-specific interaction for molecular docking. *J. Chem. Inf. Model.* **62**, 2499–2509 (2022).
38. R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004).
39. Z. Liu *et al.*, PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
40. Z. Cang, L. Mu, G. W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**, e1005929 (2018).
41. R. Wang, D. D. Nguyen, G. W. Wei, Persistent spectral graph. *Int. J. Numer. Meth. Biomed. Eng.* **36**, e3376 (2020).
42. Z. Meng, K. Xia, Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **7**, eabc5329 (2021).
43. R. Liu, X. Liu, J. Wu, Persistent path-spectral (PPS) based machine learning for protein-ligand binding affinity prediction. *J. Chem. Inf. Model.* **63**, 1066–1075 (2023).
44. H. Feng, G. W. Wei, Virtual screening of DrugBank database for hERG blockers using topological Laplacian-assisted AI models. *Comput. Biol. Med.* **153**, 106491 (2023).
45. C. Zhang *et al.*, In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicol. Res.* **5**, 570–582 (2016).
46. X. Zhang, J. Mao, M. Wei, Y. Qi, J. Z. Zhang, HergSPred: Accurate classification of hERG blockers/nonblockers with machine-learning models. *J. Chem. Inf. Model.* **62**, 1830–1839 (2022).
47. X. Li, Y. Zhang, H. Li, Y. Zhao, Modeling of the hERG K+ channel blockage using online chemical database and modeling environment (OCHEM) *Mol. Inf.* **36**, 1700074 (2017).
48. C. Cai *et al.*, Deep learning-based prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model.* **59**, 1073–1084 (2019).
49. D. Chen, J. Zheng, G. W. Wei, F. Pan, Extracting predictive representations from hundreds of millions of molecules. *J. Phys. Chem. Lett.* **12**, 10793–10801 (2021).
50. D. D. Nguyen, G. W. Wei, AGL-score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.* **59**, 3291–3304 (2019).
51. K. Wu, G. W. Wei, Quantitative toxicity prediction using topology based multitask deep neural networks. *J. Chem. Inf. Model.* **58**, 520–531 (2018).
52. K. Gao *et al.*, Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **22**, 8373–8390 (2020).
53. T. Martin, "User's guide for test (version 4.2) (toxicity estimation software tool) a program to estimate toxicity from molecular structure" (Tech. Rep. EPA/600/R-16/058, US EPA office of research and development, Washington, DC, 2016).
54. N. Gügümcü, L. H. Kauffman, New invariants of knotoids. *Eur. J. Comb.* **65**, 186–229 (2017).
55. E. Panagiotou, L. H. Kauffman, Vassiliev measures of complexity of open and closed curves in 3-space. *Proc. R. Soc. A* **477**, 20210440 (2021).
56. W. Niemyska *et al.*, Complex lasso: New entangled motifs in proteins. *Sci. Rep.* **6**, 36895 (2016).
57. P. Dabrowski-Tumanski *et al.*, Knotprot 2.0: A database of proteins with knots and other entangled structures. *Nucleic Acids Res.* **47**, D367–D375 (2019).
58. J. Chen, R. Zhao, Y. Tong, G. W. Wei, Evolutionary de Rham-Hodge method. *Discret. Contin. Dyn. Syst. Ser. B* **26**, 3785 (2021).
59. K. Opron, K. Xia, G. W. Wei, Communication: Capturing protein multiscale thermal fluctuations. *J. Chem. Phys.* **142**, 06B401-1 (2015).
60. L. Shen, J. Liu, G. W. Wei, Evolutionary Khovanov homology. *AIMS Math.* **9**, 26139–26165 (2024).
61. L. Shen *et al.*, Data from "mGLI-KDA". GitHub. https://github.com/WeilabMSU/mGLI-KDA. Deposited 17 March 2024.