JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access
This article is licensed under CC-BY 4.0

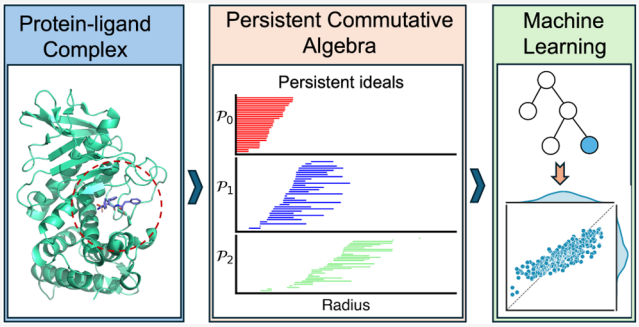
pubs.acs.org/jcim
Article

# CAML: Commutative Algebra Machine Learning—A Case Study on Protein−Ligand Binding Affinity Prediction

Hongsong Feng, Faisal Suwayyid, Mushal Zia, JunJie Wee, Yuta Hozumi, Chun-Long Chen, and Guo-Wei Wei*

Cite This: J. Chem. Inf. Model. 2025, 65, 6732−6743
Read Online
ACCESS | Metrics & More | Article Recommendations

**ABSTRACT:** Recently, Suwayyid and Wei introduced commutative algebra as an emerging paradigm for machine learning and data science. In this work, we propose commutative algebra machine learning (CAML) for the prediction of protein−ligand binding affinities. Specifically, we apply persistent Stanley−Reisner theory, a key concept in combinatorial commutative algebra, to the affinity predictions of protein−ligand binding and metalloprotein−ligand binding. We present three new algorithms, i.e., element-specific commutative algebra, category-specific commutative algebra, and commutative algebra on bipartite complexes, to tackle the complexity of data involved in (metallo) protein−ligand complexes. We show that the proposed CAML outperforms other state-of-the-art methods in (metallo) protein−ligand binding affinity predictions, indicating the great potential of commutative algebra learning.




## INTRODUCTION

Drug discovery plays a vital role in contemporary medicine, profoundly impacting global health outcomes. Conventional drug development processes, however, are time-intensive and costly, requiring more than a decade and billions of dollars to commercialize a single drug.[1] Established techniques such as molecular docking,[2−5] free energy perturbation,[6] and empirical modeling[7] have propelled advancements but face inherent constraints. These methods often suffer from inaccuracies, demand substantial computational resources for large-scale analyses, and may overlook novel binding sites or interaction dynamics, potentially missing therapeutic breakthroughs.

Machine-learning (ML) approaches are gaining traction as powerful tools in drug design,[8−11] renowned for their capacity to forecast protein structures and detect intricate patterns for enhanced predictions.[12] The adoption of deep learning, integrated with chemoinformatics and bioinformatics,[13] marks a transformative shift toward data-driven methodologies in pharmaceutical research.[14−16] Nevertheless, challenges persist, including limited data sets,[17] data imbalance,[18] intricate molecular architectures, and stereochemical complexities. Furthermore, embedding essential physical interactions, such as hydrogen bonding, van der Waals forces, hydrophobic effects, electrostatic forces, and ionic bonds, into ML algorithms for protein−ligand binding remains a significant hurdle.[19−21]

To address these limitations, researchers are employing sophisticated mathematical frameworks rooted in algebraic topology, differential geometry, and combinatorial graph theory.[22] These multiscale models, previously successful in characterizing biomolecular systems,[23−27] capture fundamental physical, chemical, and biological interactions critical to protein−ligand binding while clarifying the three-dimensional (3D) structural intricacies of these complexes. Notably, these approaches have delivered top-tier performances in the D3R Grand Challenges, a leading international competition in computer-aided drug design.[24,28] Inspired by this success, there is a continuous effort in computational biology and applied mathematics to seek advanced mathematical representations of complex biomolecules such as proteins and their interactions.

Commutative algebra is a branch of mathematics that studies commutative rings, their ideals, modules, and related structures.[29,30] It serves as a foundational framework for algebraic geometry, number theory, and many other areas of mathematics. Its key concepts include Noetherian rings, Cohen-Macaulay rings, localization theory, primary decomposition, dimension theory, and homological algebra.

Despite its importance in pure mathematics, it has hardly been applied in data science and artificial intelligence.

Received: April 27, 2025
Revised: June 1, 2025
Accepted: June 5, 2025
Published: June 16, 2025



© 2025 The Authors. Published by American Chemical Society

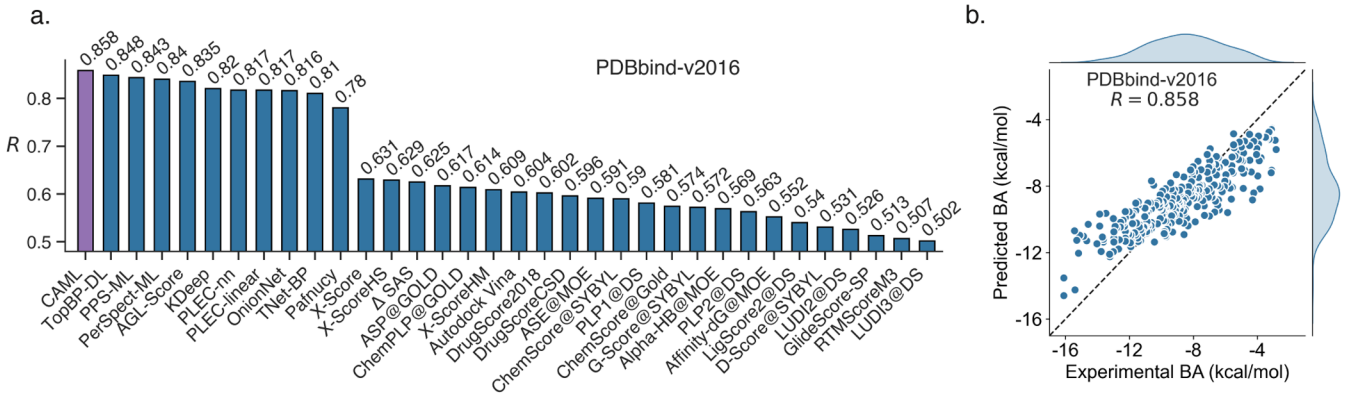



**Figure 1.** (a) Comparison of the predictions from our CAML model with other published models in terms of Pearson correlation coefficient (PCC) ($R$) on the PDBbind-v2016 data set. (b) Comparison between the experimental and predicted binding affinities (BAs) from our CAML model for the PDBbind-v2016 data set.

Recently, Suwayyid and Wei introduced persistent Stanley–Reisner theory to bridge commutative algebra, algebraic topology, machine learning, and data science.[31] Stanley–Reisner theory is the study of the commutative algebra, i.e., square-free monomial ideals in a polynomial ring, of simplicial complexes, structured sets comprising points, line segments, triangles, and their higher-dimensional counterparts.[32−34] Therefore, persistent Stanley–Reisner theory (PSRT) enables commutative algebra analysis (CAA) of point cloud data and machine-learning predictions. Specifically, PSRT examines how the Stanley–Reisner structure of a complex evolves under filtration. Many computable quantities, including persistent graded Betti numbers via Hochster's formula, persistent $f$-vectors, persistent $h$-vectors, and persistent facet ideals, have been proposed. Facet persistence barcodes, which record the birth and death of persistent facet ideals as the simplicial complex evolves, have been introduced for practical applications in data science. PSRT provides novel insights into geometry, topology, and combinatorial techniques at multiple scales. An important motivation for this development was persistent homology,[35,36] an algebraic topology tool for topological data analysis (TDA) and topological deep learning (TDL),[37] a new frontier for relational learning.[38]

The objective of this work is to explore the utility and demonstrate the potential of commutative algebra machine learning (CAML) for protein–ligand binding affinity prediction. We consider two benchmark data sets: the PDBbind-v2016 data set for protein–ligand binding interactions[39] and a metalloprotein–ligand binding data set.[40] As these data sets involve intricate three-dimensional (3D) protein–ligand complexes as well as complex physical and chemical interactions, we propose a few new CAML algorithms, i.e., element-specific commutative algebra, category-specific commutative algebra, and commutative algebra on bipartite complexes, to capture intrinsic physical and chemical interactions, such as hydrogen bonding, van der Waals forces, hydrophobic effects, electrostatic forces, and ionic bonds in 3D metalloprotein–ligand complexes. As shown in the results, the proposed CAML consistently outshines its peers, achieving state-of-the-art outcomes across benchmark data sets in protein–ligand binding affinity prediction.

The rest of this article is organized as follows. Results section is devoted to the results of CAML for protein–ligand binding and metalloprotein–ligand binding predictions. Methods are described in Methods section. This work ends with a conclusion.

## RESULTS

**Protein–Ligand Binding Affinity Predictions.** The PDBbind database[39] is a widely recognized, curated resource that systematically collects experimentally determined 3D structures of protein–ligand complexes alongside their corresponding binding affinity data, e.g., dissociation constants $K_d$, inhibition constants $K_i$, and Gibbs free energy changes $\Delta G$. It serves as a gold-standard benchmark for developing and validating computational models aimed at predicting protein–ligand binding affinities (BAs).

Leveraging our persistent Stanley-Rensiner theory (PSRT), we developed commutative algebra machine-learning (CAML) models to predict protein–ligand binding affinities. The models were benchmarked against established methods using the widely recognized PDBbind data set. Specifically, we focused on the PDBbind-v2016 data set, a rigorously curated version with clearly defined training (3768 complexes) and test sets (290 complexes). The PDBbind database[39] provides a comprehensive collection of 3D protein–ligand structures paired with binding affinity data. As shown in Figure 1a, our PSRT-guided model, CAML, outperformed existing state-of-the-art approaches, achieving superior predictive accuracy in the binding affinity estimation.

Numerous competitive models rooted in mathematical or physical frameworks,[19−21] such as persistent homology[23] and persistent spectral theories,[26,41] have been reported. These rank among the top performers in this domain (Figure 1a). On the PDBbind-v2016 test set, CAML achieved a Pearson correlation coefficient ($R$) of 0.858, significantly surpassing persistent homology-based TopBP-DL ($R$ = 0.848)[23] and persistent spectral theory-based models PerSpect-ML ($R$ = 0.843)[26] and PPS-ML ($R$ = 0.840).[41] Several competitive neural network-based models for protein–ligand binding affinity prediction have also been benchmarked on the PDBbind-v2016 test set.[42−44] We conducted extensive predictive comparisons on this data set with models from the literature, as shown in Figure 1a, which highlights the superior performance of our model in protein–ligand binding affinity prediction. These results highlight CAML's efficacy as a novel analytical tool and its ability to drive advanced predictive models for binding affinity.



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

Table 1. Modeling Performance of Various Strategies on the Test Set of PDBbind-v2016[a]

| data set | CAML(ES) | CAML(CS) | CAML(ES,CS) | transformer | CAML(ES,CS) + transformer |
|---|---|---|---|---|---|
| $R$ | 0.836 ± 0.001 | 0.834 ± 0.001 | 0.845 ± 0.001 | 0.836 ± 0.001 | **0.858 ± 0.001** |
| RMSE | 1.743 ± 0.004 | 1.745 ± 0.004 | 1.719 ± 0.003 | 1.713 ± 0.005 | **1.669 ± 0.004** |

[a]The evaluation metrics used are the pearson correlation coefficient ($R$) and the root mean square error (RMSE, in kcal/mol). Twenty independent runs with different random seeds were performed, and the average metric values are reported. The ± sign indicates the standard deviation value. CAML(ES) and CAML(CS) refer to commutative algebraic machine-learning models combined with element-specific and category-specific atom combinations, respectively. CAML(ES, CS) represents the consensus results from the CAML(ES) and CAML(CS) Models. Transformer refers to sequence-based modeling using natural language processing. CAML(ES, CS) + transformer indicates the consensus predictions from the CAML(ES, CS) and transformer models, which is defined as our final CAML model.

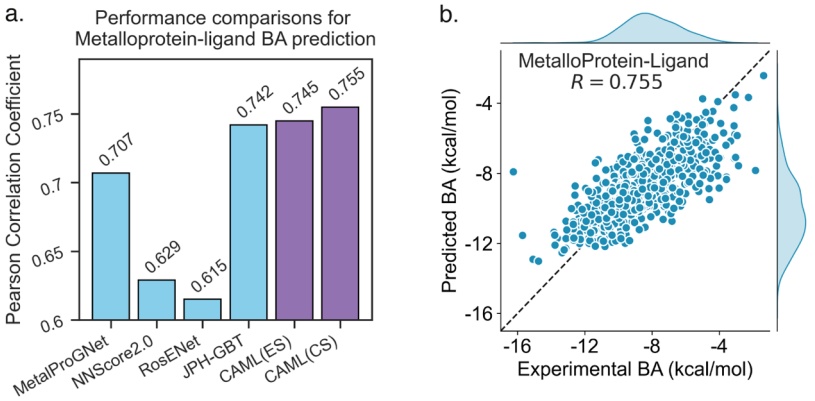



Figure 2. (a) Prediction performance of our CAML models on the metalloprotein−ligand binding affinity (BA) data set in terms of Pearson correlation coefficient ($R$). (b) Comparison between the experimental binding affinities and the predicted values from our CAML(CS) model for this data set.

CAML's reliability is further demonstrated by the strong alignment between experimental and predicted binding affinities, as visualized in Figure 1b. Its success stems from three key innovations: (1) PSRT-driven molecular data analysis, (2) element-specific (ES) and category-specific (CS) modeling of intra- and intermolecular interactions, and (3) integration of natural language processing (NLP) via a transformer architecture (details in Vectorization of Persistent Commutative Algebra section).

As detailed in Table 1, we evaluated five distinct models. The top performer (final CAML model, $R$ = 0.858) combines consensus predictions from CAML(ES, CS) (a fusion of element- and category-specific strategies) with transformer-based sequence analysis. The ES approach, a widely adopted method for dissecting atomic interactions, and the CS strategy, which categorizes interactions by atomic properties, were individually effective. Their combination (CAML(ES, CS), $R$ = 0.853) further enhanced the accuracy. Integrating these with sequence-based NLP predictions via the transformer model yielded the final CAML's performance, highlighting the synergy between structural and sequential pattern analysis.

This multifaceted approach positions CAML as a state-of-the-art tool for protein−ligand binding affinity prediction, with implications for drug discovery and molecular design.

**Metalloprotein−Ligand Binding Affinity Predictions.** As another benchmark example, we consider metalloprotein−ligand binding affinities. Metalloproteins are proteins that incorporate metal ions as integral structural components and play indispensable roles in biological processes such as cellular respiration, electron transfer, catalytic reactions, and structural stabilization.[45−47] More specifically, protein-metal-binding sites are responsible for catalyzing some of the most difficult yet important functions, such as photosynthesis, respiration, water oxidation, molecular oxygen reduction, and nitrogen fixation. Studies estimate that roughly half of all proteins in biology are metalloproteins.[48−50] The prediction of metalloprotein−ligand binding affinities represents a critical challenge in drug discovery. Deciphering the structure, such as function relationships and interaction mechanisms, of metalloproteins is pivotal for unraveling fundamental biological pathways and accelerating the design of targeted therapeutics.

Recent advancements have addressed the scarcity of specialized data sets for this task. For example, the study by ref 40 introduced the largest curated data set to date for metalloprotein−ligand binding affinity prediction, providing a robust foundation for developing and benchmarking computational models in this domain.

Using the data set from,[40] we constructed two CAML machine-learning models based on element-specific (ES) and category-specific (CS) strategies, resulting in CAML(ES) and CAML(CS) models. Figure 2a gives the comparisons between our CAML models with other published models in terms of the Pearson correlation coefficient ($R$). The previous state-of-art model is the JPH-GBT model,[51] which gave a much higher $R$ value than other models.[40,52,53] Our CAML models redefines the state-of-art for metalloprotein−ligand binding affinity predictions. Model CAML(ES) and CAML(CS) give $R$ values of 0.745 and 0.755, respectively. Figure 2b shows the comparison between the experimental binding affinity values and the predicted binding affinity using our CAML(CS) model. Table 2 gives comparisons of our models with others in terms of $R$ and RMSE metrics.

## METHODS

In this section, we first list our data sets, showing the clear separation between training and test sets. Then, we provide an



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

Table 2. Comparison of Our CAML Models with Existing Machine-Learning Approaches in Modeling Metalloprotein-Ligand Binding Affinity Data Set[a]

| machine-learning models | $R$ | RMSE |
|---|---|---|
| RosENet[40,52] | 0.615 ± 0.017 | 1.436 ± 0.011 |
| NNScore2.0[40,53] | 0.629 ± 0.002 | 1.391 ± 0.004 |
| MetalProGNet[40] | 0.703 ± 0.010 | 1.285 ± 0.020 |
| JPH-GBT[51] | 0.742 ± 0.001 | 1.205 ± 0.001 |
| CAML(ES) | 0.745 ± 0.001 | 1.202 ± 0.002 |
| CAML(CS) | **0.755 ± 0.001** | **1.185 ± 0.002** |

[a]Two evaluation metrics, including pearson correlation coefficient ($R$) and root-mean-squared error (RMSE) are used. The ± sign indicates the standard deviation value. CAML(ES) and CAML(CS) refer to models utilizing the PSRT vectorization framework combined with element-specific (ES) and category-specific (CS) atom groupings, respectively. The RMSE values are computed based on the raw p$K_d$ labels that serve as binding affinities.

overview of the persistent Stanley−Reisner theory (PSRT). Next, we describe the vectorization of persistent commutative algebra followed by natural language processing (NLP) molecular descriptors. Machine-learning models and model parameters are given. We also define the evaluation metrics.

**Data Sets.** The first benchmark data set we use is PDBbind-v2016, which is the largest protein−ligand binding affinity prediction data set with well-defined training and test sets. The second one is the metalloprotein−ligand data set compiled from PDBbind-v2020 in work.[40] This original data set consists of training, validation, and test sets with different types of metal ions. We utilize their training and test sets to benchmark the performance of PSRT-based machine-learning models. Details of the dataset sizes are listed in Table 3.

Table 3. Details of the Data Sets Utilized for Benchmark Tests in This Study

| data set | total | training set | test set |
|---|---|---|---|
| PDBbind-v2016[39] | 4057 | 3767 | 290 |
| metalloprotein−ligand[51] | 2463 | 1845 | 618 |

**Persistent Stanley−Reisner Theory.** Persistent Stanley−Reisner theory is a novel framework for analyzing the shape of data by leveraging tools from combinatorial commutative algebra.[31] It encodes point cloud data as simplicial complexes—combinatorial structures built from vertices, edges, triangles, and higher-dimensional simplices—capturing both topological and combinatorial features inherent in the data. A filtration process is then applied to these complexes to track the evolution and persistence of such features across multiple spatial or geometric scales. This approach introduces algebraic invariants such as persistent $h$-vectors, $f$-vectors, graded Betti numbers, and facet ideals, thus providing a new algebraic perspective within the broader framework of topological data analysis.

*Simplicial Complex.* A simplicial complex $\Delta$ on the finite vertex set $V = \{x_1, x_2, \cdots, x_n\}$ is a collection of subsets of $V$, referred to as faces or simplices, satisfying the following conditions:

1. If $F \in \Delta$ and $G \subseteq F$, then $G \in \Delta$.
2. For each $i = 1, \cdots, n$, the singleton $\{x_i\}$ belongs to $\Delta$. In particular, every vertex is included as a face.

A face consisting of $r + 1$ vertices is called an $r$-dimensional face. The dimension of $\Delta$ is defined as the maximum dimension among its faces. A face that is maximal with respect to inclusion is called a facet of $\Delta$, and the set of all such facets is denoted by $\mathcal{F}(\Delta)$.

*Stanley−Reisner Theory and Facet Ideals.* Let $k$ be a field, and consider the standard polynomial ring

$$S = k[x_1, x_2, ..., x_n] \tag{1}$$

endowed with the natural $\mathbb{Z}$-grading determined by $\deg(x_i) = 1$ for all $i = 1, \cdots, n$. Let $\Delta$ be a simplicial complex on the vertex set $V = \{x_1, x_2, \cdots, \text{and } x_n\}$. The Stanley−Reisner ideal associated with $\Delta$ is defined as

$$I(\Delta) = \langle x_{i_1} x_{i_2} \cdots x_{i_r} | \{x_{i_1}, x_{i_2}, ..., x_{i_r}\} \notin \Delta \rangle \tag{2}$$

that is, it is the ideal generated by all squarefree monomials corresponding to nonfaces of $\Delta$. The quotient ring

$$k[\Delta] = S/I(\Delta) \tag{3}$$

is called the Stanley−Reisner ring of $\Delta$.

It is a classical result that the Krull dimension of the Stanley−Reisner ring is given by

$$\dim(k[\Delta]) = \dim(\Delta) + 1 \tag{4}$$

which we denote by $d$. Thus, the simplicial complex $\Delta$ is said to be $(d - 1)$-dimensional.

Now, for any subset $A \subseteq V = \{x_1, x_2, \cdots, x_n\}$, we define the associated prime monomial ideal by

$$P_A := (x_i | x_i \notin A)$$

In the context of Stanley−Reisner theory, we are particularly interested in the prime monomial ideals associated with the facets of $\Delta$, referred to as the facet prime monomial ideals, or simply, facet ideals.

A fundamental property of the Stanley−Reisner ideal is that it admits a primary decomposition as the intersection of the facet ideals:

$$I(\Delta) = \bigcap_{\sigma \in \mathcal{F}(\Delta)} P_\sigma$$

where $\mathcal{F}(\Delta)$ denotes the set of all facets of the simplicial complex $\Delta$.

*Graded Betti Numbers, f-Vectors, and h-Vectors.* As a graded $S$-module, $k[\Delta]$ admits a minimal free resolution of the form

$$\cdots \to \bigoplus_j S(-j)^{\beta_{i,j}(k[\Delta])} \to \cdots \to \bigoplus_j S(-j)^{\beta_{0,j}(k[\Delta])} \to k[\Delta] \to 0 \tag{5}$$

where $S(-j)$ is the graded free module $S$ shifted in degree by $j$ and graded Betti numbers are

$$\beta_{i,j}(k[\Delta]) = \dim_k \mathrm{Tor}_i^S(k[\Delta], k)_j \tag{6}$$

with $\mathrm{Tor}_i^S(k[\Delta], k)_j$ being the Tor module, which measures how nontrivial the resolution is at homological degree $i$. For a subset $W \subseteq V$, the restriction (or induced subcomplex) of $\Delta$ to $W$ is

$$\Delta_W = \{\tau \in \Delta : \tau \subseteq W\} \tag{7}$$

Hochster's formula provides an explicit description of the $\mathbb{Z}$-graded Betti numbers $\beta_{i,j+i}(k[\Delta])$ of the Stanley−Reisner ring


https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

$k[\Delta]$ in terms of the reduced simplicial homology of induced subcomplexes. For integers $i, j \geq 0$, it states that

$$\beta_{i,j+i}(k[\Delta]) = \sum_{\substack{W \subseteq \{x_1,\ldots,x_n\} \\ |W|=j+i}} \dim_k \tilde{H}_{j-1}(\Delta_W; k) \tag{8}$$

where $\Delta_W$ denotes the subcomplex of $\Delta$ induced on the vertex set $W$, and $\tilde{H}_{j-1}(\Delta_W; k)$ is the $(j-1)$-st reduced homology group with coefficients in $k$. This formula holds for $1 \leq i \leq n-1$ and $1 \leq j \leq \min\{n-i, \dim(\Delta)+1\}$.

In particular, for $j = 1$, the formula simplifies to

$$\beta_{i,i+1}(k[\Delta]) = \sum_{\substack{W \subseteq \{x_1,\ldots,x_n\} \\ |W|=i+1}} (\beta_0(\Delta_W) - 1) \tag{9}$$

and for $j \geq 2$, it takes the form

$$\beta_{i,j+i}(k[\Delta]) = \sum_{\substack{W \subseteq \{x_1,\ldots,x_n\} \\ |W|=j+i}} \beta_{j-1}(\Delta_W) \tag{10}$$

where $\beta_{j-1}(\Delta_W)$ denotes the $(j-1)$-st Betti number of the homology of $\Delta_W$. These expressions establish a direct connection between the topological invariants of the simplicial complex $\Delta$ and the algebraic invariants of its associated Stanley–Reisner ring.

*f-Vectors and h-Vectors.* Let $\Delta$ be a simplicial complex of dimension $d-1$. The $f$-vector of $\Delta$ is defined as

$$(f_0, f_1, \ldots, f_{d-1}) \tag{11}$$

where $f_i$ denotes the number of $i$-dimensional faces of $\Delta$. By convention, we set $f_{-1} = 1$ to account for the empty face.

The Hilbert series of the Stanley–Reisner ring $k[\Delta]$, also referred to as the Hilbert series of $\Delta$, is given by

$$H_\Delta(s) = \sum_{d \geq 0} \dim_k(k[\Delta]_d) s^d \tag{12}$$

where $k[\Delta]_d$ denotes the degree-$d$ component of the $\mathbb{Z}$-graded ring $k[\Delta]$.

For a $(d-1)$-dimensional simplicial complex $\Delta$, it is a classical result that the Hilbert series can be expressed as a rational function of the form

$$H_\Delta(s) = \frac{h_0 + h_1 s + \cdots + h_d s^d}{(1-s)^d} \tag{13}$$

where $(h_0, h_1, \cdots, h_d)$ is the $h$-vector of $\Delta$, or equivalently, of its Stanley–Reisner ring.

The $f$-vector and $h$-vector are related by the identity

$$\sum_{j=0}^{d} h_j s^j = \sum_{j=0}^{d} f_{j-1}(1-s)^{d-j} s^j, \text{ with } f_{-1} = 1 \tag{14}$$

Equivalently, the entries of the $h$-vector can be expressed in terms of the $f$-vector by the relation

$$h_j = \sum_{i=0}^{j} (-1)^{j-i} \binom{d-i}{j-i} f_{i-1}, \; j = 0, 1, \ldots, d \tag{15}$$

and conversely, the $f$-vector can be recovered from the $h$-vector via

$$f_{j-1} = \sum_{i=0}^{j} \binom{d-i}{j-i} h_i, \; j = 0, 1, \ldots, d \tag{16}$$

*Filtration and Persistent Stanley–Reisner Theory (PSRT).* A fundamental limitation of using a simplicial complex $\Delta$ to model data is that it typically captures topological or combinatorial information on a single scale, thereby omitting geometric details that may vary across scales. To address this limitation, one introduces a multiscale framework through the use of filtrations, leading to the theory of persistent homology, which identifies topological features that persist across a range of scales.

Persistent Stanley–Reisner theory follows a similar philosophy, extending combinatorial and algebraic invariants to a persistent setting. Let $\Delta$ be an abstract simplicial complex on a finite vertex set $V$. Given a monotone function $g\colon \Delta \to \mathbb{R}$, that is,

$$\tau \subseteq \sigma \Rightarrow g(\tau) \leq g(\sigma)$$

we define the induced filtration of $\Delta$ by

$$\tilde{g} = (\Delta_g^r | r \in \mathbb{R}) \tag{17}$$

where each subcomplex $\Delta_g^r \subseteq \Delta$ is defined as

$$\Delta_g^r := \{\sigma \in \Delta | g(\sigma) \leq r\}$$

For notational simplicity, we may write $\Delta^r$ in place of $\Delta_g^r$ when the context is clear.

Given a subset $W \subseteq V$, then if $(\Delta^r)_{r\in\mathbb{R}}$ is a filtration of $\Delta$, the induced filtration on the subcomplex $\Delta_W$ is given by

$$\Delta_W^r := \Delta^r \cap \Delta_W \subseteq \Delta^{r'} \cap \Delta_W = \Delta_W^{r'} \text{ for all } r \leq r' \tag{18}$$

We define the persistent Stanley–Reisner graded Betti number $\beta_{i,i+j}^{r,r'}(k[\Delta])$ as

$$\beta_{i,i+j}^{r,r'}(k[\Delta]) = \sum_{\substack{W \subseteq V \\ |W|=i+j}} \dim_k(\iota_{j-1}^{r,r'}\colon \tilde{H}_{j-1}(\Delta_W^r; k) \to \tilde{H}_{j-1}(\Delta_W^{r'}; k)) \tag{19}$$

where $\iota_{j-1}^{r,r'}$ is the homomorphism induced by the inclusion $\Delta_W^r \to \Delta_W^{r'}$ on the $(j-1)$-st reduced homology, and $\beta_{j-1}(\Delta_W^r)$ is the persistent Betti number of the inclusion from $\Delta_W^r$ to $\Delta_W^{r'}$. Summing over all relevant subsets, $W$ yields a multiscale refinement of Hochster's formula, capturing the persistent homological features of the complex across varying levels of filtration.

The persistent Stanley–Reisner graded Betti numbers $\beta_{i,i+j}^{r,r'}(k[\Delta])$ generalize classical persistent Betti numbers; for example,

$$\beta_{i,|V|}^{r,r'} = \beta_{|V|-i-1}^{r,r'}$$

and further encode additional combinatorial information by tracking all monomial degrees in the resolution.

One may also extend combinatorial invariants such as $f$-vector and $h$-vector to persistent settings. The persistent $h$-vector is defined by

$$h_m^{r,r'} = \sum_{j=0}^{m} \binom{n-d+m-j-1}{m-j} \left( \sum_{i=0}^{j} (-1)^i \beta_{i,j}^{r,r'} \right) \tag{20}$$

and the corresponding persistent $f$-vector is given by

$$f_{m-1}^{r,r'} = \sum_{i=0}^{m} \binom{d-i}{m-i} h_i^{r,r'}, \; m = 0, 1, \ldots, d \tag{21}$$



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

In the case $r = r'$, the persistent $f$-vector coincides with the classical Stanley−Reisner $f$-vector; that is, $f^{r,r'} = f$. When the underlying simplicial complex is large, the filtration often yields a substantial number of faces, leading to rapid growth in the components of the $f$-vector. To normalize this growth and facilitate meaningful comparisons across scales, we define the average rate of the $f$-vector by

$$\overline{f}(r) := \frac{f^{r,r}}{r}$$

Finally, consider a filtration $(\Delta^r)_{r\in\mathbb{R}}$ of the simplicial complex $\Delta$. As $r$ increases, the corresponding Stanley−Reisner ideals $I(\Delta^r)$ evolve through the reverse inclusion

$$I(\Delta^{r'}) \subseteq I(\Delta^r) \text{ for } r \le r'$$

Let $\mathcal{P}(\Delta^r)$ denote the set of facet prime monomial ideals (or facet ideals) of $\Delta^r$. For each $i \ge 0$, define $\mathcal{P}_i(\Delta^r)$ to be the subcollection of $\mathcal{P}(\Delta^r)$ consisting of those associated with $i$-dimensional facets. Then, we have the disjoint union

$$\mathcal{P}(\Delta^r) = \bigcup_{i=0}^{\dim(\Delta^r)} \mathcal{P}_i(\Delta^r)$$

This decomposition motivates a persistent interpretation of the evolution of facet ideals over the filtration. In analogy to persistent homology, we refer to the facet ideals $P_\sigma$ of $I(\Delta^r)$ as the persistent facet ideals of $\Delta$, representing combinatorial structures that persist across filtration levels.

We define the facet persistence Betti number $\beta_i^{r,r'}$ to be the number of the persistent facet ideals in $\mathcal{P}_i(\Delta^r)$ that are persistent facet ideals in $\mathcal{P}_i(\Delta^{r'})$. Explicitly,

$$\beta_i^{r,r'} = |\mathcal{P}_i(\Delta^r) \cap \mathcal{P}_i(\Delta^{r'})|$$

Analogously to the case of persistent $f$-vectors, we define the average rate of the facet persistence Betti numbers by

$$\overline{\beta}(r) := \frac{\beta^{r,r}}{r}$$

**Vectorization of Persistent Commutative Algebra.** *Commutative Algebra on Bipartite Complexes.* In (metallo)-protein−ligand binding interactions, bipartite complexes refer to molecular assemblies composed of two distinct, non-overlapping (metallo)protein and ligand components that interact to form a functional unit. In graph theory, a bipartite complex refers to a structure where vertices (or nodes) are divided into two distinct, disjointed sets, and edges only connect vertices from different sets. This bipartite structure is adopted in our commutative algebra analysis of (metallo)-protein−ligand binding data.

*Element-Specific Commutative Algebra.* There are various types of intramolecular and intermolecular interactions, including hydrogen bonding, electrostatic forces, and hydrophobic and hydrophilic interactions. To effectively characterize these critical interactions, element-specific modeling was developed in our earlier work,[37] demonstrating notable effectiveness. In particular, molecular interactions enriched within various pairwise combinations of atom sets are captured by using persistent commutative algebra (PCA) modeling and subsequently represented through PCA-based vectorization.

For the PDBbind-v2016 data set, molecular interactions are characterized based on four commonly occurring atom types in proteins: carbon (C), nitrogen (N), oxygen (O), and sulfur (S), along with ten atom types in ligands, including carbon (C), nitrogen (N), oxygen (O), sulfur (S), phosphorus (P), fluorine (F), chlorine (Cl), bromine (Br), iodine(I), and hydrogen (H). In the case of the metalloprotein−ligand data set, additional atomic interactions involving metal ions must be considered. Based on a statistical analysis of the metal ions frequently occurring in protein−ligand binding pockets, we include seven additional atom types: zinc (Zn), magnesium (Mg), manganese (Mn), calcium (Ca), sodium (Na), iron (Fe), and nickel (Ni) due to their prevalent presence. Table 4

**Table 4. Counts of Some Frequently Occurring Metal Ions in the Training and Test Sets in the Metalloprotein-Ligand Binding Affinity Data Set**

| | elements | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| data set | Zn | Mg | Ca | Mn | Ni | Fe | Na | Cu |
| training set | 1389 | 614 | 516 | 319 | 63 | 60 | 48 | 8 |
| test set | 458 | 220 | 163 | 114 | 18 | 15 | 12 | 7 |

gives the counts of some frequently occurring metal atoms in the training and test sets in the metalloprotein−ligand binding affinity data set. Cu atoms were excluded in our machine-learning model due to their infrequent occurrence in metalloprotein−ligand complexes.

For protein−ligand complexes in the PDBbind-v2016 data set, it is sufficient to consider only protein−ligand interactions. However, for the metalloprotein−ligand binding affinity prediction, three types of interactions must be taken into account: protein−ligand interactions, metal ion−ligand interactions, and protein−metal ion interactions. Specifically, there are $4 \times 10 = 40$ atom pair combinations for protein−ligand interactions (P−L), $7 \times 4 = 28$ combinations for metal ion−protein interactions (M−P), and $7 \times 10 = 70$ combinations for metal ion−ligand interactions (M−L). Consequently, 40 types of atomic interactions are considered for protein−ligand complexes in the PDBbind-v2016 data set, while a total of 138 interaction types are incorporated for metalloprotein−ligand complexes.

For the 3D coordinate point cloud corresponding to each atom group combination, we utilize PCA to generate vectorized representations. The facet-persistent Betti numbers and their corresponding rates are employed to design a set of features for each atom group. A cutoff distance of 12 Å from the ligand is used to collect nearby protein atoms, while a cutoff distance of 15 Å from the ligand is applied to gather metal atoms. In the PCA-based featurization process, the filtration range is set from 1 Å to 12 Å or 15 Å for each case, with a step size of 0.5 Å for the filtration steps.

Rips complexes are constructed for each point cloud, while bipartite complexes are generated to capture the three types of molecular interactions. In our implementation, we focus exclusively on the facet Betti numbers and their corresponding rates for 0 and 1-simplices. The final molecular descriptor is obtained by concatenating the PCA-derived feature vectors from each point cloud. Figure 4 presents a schematic illustration of the general persistence Betti curves computed for a representative set of atoms.

*Category-Specific Commutative Algebra.* In addition to element-specific modeling, we also adopted a category-specific strategy to capture intrinsic molecular interactions characterized by amino acid types. Proteins are composed of 20



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

common amino acid residues, which can be classified into four major categories based on their side chain properties: hydrophobic ($H$), uncharged ($U$), negatively charged ($N$), and positively charged ($P$). We denote the atoms belonging to these categories as $A_H$, $A_U$, $A_N$, and $A_P$, respectively. Table 5 summarizes the four categories along with their corresponding amino acid types. Figure 3 shows the distribution of average atom numbers within different amino acid categories for the proteins in the PDBbind-v2016 and metalloprotein−ligand binding affinity data sets. Protein atoms within a 12 Å cutoff distance from the ligand were selected for atom count analysis.

**Table 5. Four Categories of Amino Acid Residues in Proteins According to Their Side Chain Properties**

| hydrophobic ($H$) | uncharged ($U$) | negatively charged ($N$) | positively charged ($P$) |
|---|---|---|---|
| glycine (Gly) | serine (Ser) | aspartic (Asp) | lysine (Lys) |
| alanine (Ala) | threonine (Thr) | glutamic (Glu) | arginine (Arg) |
| valine (Val) | asparagine (Asn) | | histidine (His) |
| leucine (Leu) | glutamine (Gln) | | |
| isoleucine (Ile) | tyrosine (Tyr) | | |
| methionine (Met) | cysteine (Cys) | | |
| proline (Pro) | | | |
| phenylalanine(Phe) | | | |
| tryptophan (Trp) | | | |

As with the element-specific atom groupings used for ligands and metal ions, we consider 10 atom groups in ligands and 7 atom groups in metal ions based on their element types. Consequently, there are 40 atom group combinations in general protein−ligand complexes and 138 combinations in metalloprotein−ligand complexes. Similar PCA-based vectorizations are applied to these atom group combinations to generate molecular descriptors.

**Natural Language Processing (NLP) Molecular Descriptors.** Natural language processing (NLP) has recently also become a popular machine-learning technique for molecular biosciences. We utilized NLP techniques to boost the performance of our PSRT-based machine-learning models for protein−ligand binding affinity predictions. Different from our PSRT theory, which analyzes molecular 3D structures, NLP extracts molecular physicochemical properties by analyzing molecular sequence patterns. For the protein−ligand complex, we have amino acid sequences for the protein and SMILES strings for the ligand. Some NLP-based molecular descriptors are designed using transformer techniques for both protein and ligand in works.[54,55] By concatenating molecular descriptors from those pretrained deep learning models for protein sequence and ligand SMILES strings, we obtained a sequence representation for the protein−ligand complex.

*Transformer-Based Protein Language Model.* The ESM-2 transformer model, introduced by Rives et al.,[54] has become one of the most widely adopted protein language models, with

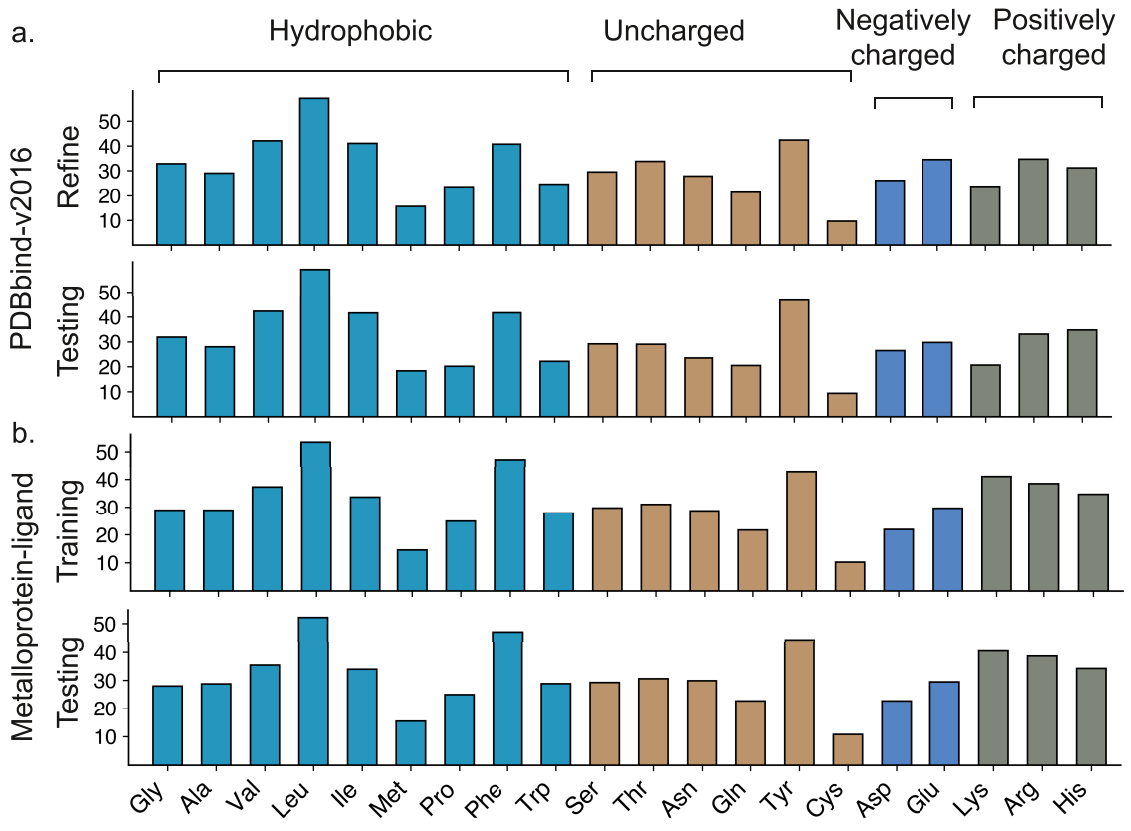



**Figure 3.** (a) Average atom counts for different amino acid types across those protein−ligand complexes in the refine and test sets of the PDBbind-v2016 data set. (b) Average atom counts for different amino acid types across those protein−ligand complexes in the training and testing sets of the metalloprotein−ligand binding affinity data sets. Protein atoms within a 12 Å cutoff distance from the ligand were selected for the atom count analysis.



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

applications in protein engineering and drug discovery. This model was trained on a data set containing 250 million amino acid sequences and employs a deep learning architecture with 34 layers and 650 million parameters. In this work, we utilized the ESM-2 model to generate sequence embeddings for proteins. At each layer, a sequence of length $L$ is encoded into a matrix of size $1280 \times L$, excluding the start and end tokens. We extracted the sequence representation from the final (34th) layer and computed the average along the sequence length axis, resulting in a 1280-dimensional feature vector.

*Transformer-Based Small Molecular Language Model.* A transformer-based deep learning framework was introduced to extract molecular representations,[55] serving as a powerful tool for machine-learning applications involving small molecules.[56] This model was trained on a collection of over 700 million SMILES strings obtained from databases such as ChEMBL, PubChem, and ZINC. Three pretrained variants were developed: model-C, model-CP, and model-CPZ. In the current study, we utilize the model-CPZ to generate molecular descriptors for ligands. For each ligand, the model produces a matrix of size $256 \times 512$, where 256 corresponds to the symbols representing the molecule and 512 is the dimension of the embedding vector for each symbol. The final molecular descriptors are obtained by the first vector among the 256 embedding vectors, resulting in a fixed-length feature vector.

**Machine Learning Modeling.** We employ the Gradient Boosting Decision Tree (GBDT) algorithm to develop our machine-learning models using the Python `scikit-learn` package (v1.3.2) for implementation. GBDT is well-regarded for its robustness against overfitting, relative insensitivity to hyperparameter settings, and ease of implementation. The algorithm creates multiple weak learners or individual trees by bootstrapping training samples and integrates their outputs to make predictions. Although weak learners are prone to making poor predictions, the ensemble approach can reduce overall errors by combining the predictions of all of the weaker learners. We input resulting PSRT molecular descriptors and transformer-based molecular descriptors into the GBDT algorithm to build regression models, respectively. The GBDT hyperparameters used for modeling are listed in Table 6.

**Table 6. Hyperparameters Used for Building Gradient Boosting Regression Models**[a]

| no. of estimators | max depth | min. sample split | learning rate |
|---|---|---|---|
| 20,000/30,000 | 7 | 5 | 0.002 |
| max features | subsample size | repetition | |
| square root | 0.8 | 20 times | |

[a]Tree numbers are set to be 20,000 and 30,000 respectively for PSRT and transformer-based molecular descriptor modeling.

**Evaluation Metrics.** To quantitatively evaluate the performance of our binding affinity prediction models, we employ the Pearson correlation coefficient (PCC), defined as

$$\mathrm{PCC}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{m=1}^{M} (y_m^e - \overline{y}^e)(y_m^p - \overline{y}^p)}{\sqrt{\sum_{m=1}^{M} (y_m^e - \overline{y}^e)^2 \sum_{m=1}^{M} (y_m^p - \overline{y}^p)^2}} \tag{22}$$

where $y_m^e$ and $y_m^p$ denote the experimental and predicted binding affinity values for the $m$th sample, respectively, and $\overline{y}^e$ and $\overline{y}^p$ are their corresponding mean values.

We also report the root mean squared error (RMSE), which is computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (y_m^e - y_m^p)^2}$$

where $y_m^e$ and $y_m^p$ represent the experimental and predicted binding affinity values for the $m$th sample, respectively.

We employ the above two metrics to assess the performance of our machine-learning models on both data sets. The original labels for these data sets are given as $pK_d$ values, which can be converted to binding free energies (in kcal/mol) by multiplying by a constant factor of 1.3633. Our models achieve low RMSE values across both data sets. For the PDBbind-v2016 data set, we convert the labels to binding energies and use them for RMSE comparisons with previously published models.

## MODEL INTERPRETABILITY

Our persistent Stanley–Reisner approach offers excellent interpretability for modeling molecular structures, motivating us to apply it in the design of molecular descriptors. As an illustrative example, we consider the vectorization of compound $C_{60}$, as shown in Figure 4. The molecule $C_{60}$ consists of 60 carbon atoms arranged in a truncated icosahedron—a polyhedral structure composed of 12 pentagonal and 20 hexagonal faces—resembling the shape of a soccer ball. In this configuration, each pentagon is surrounded by six hexagons. There are two types of bonds in the molecule. Single bonds, with a bond length of 1.453 Å, are located between pentagons and hexagons, totaling 60 in number. Double bonds with a bond length of 1.367 Å are found between pairs of hexagons. Since each hexagon is adjacent to three hexagons and three pentagons, there are $20 \times 3/2 = 30$ double bonds in total. Each atom in $C_{60}$ forms two single bonds and one double bond, connecting it to three neighboring atoms.

The figure displays the Vietoris–Rips filtration of the molecule along with the corresponding barcodes associated with its persistent facet ideals. A notable geometric feature is captured within the interval $[1.3, 1.5]$. There are 60 persistent ideal bars in the $\mathcal{P}_0$ panel corresponding to the 60 atoms. These bars vanish simultaneously around filtration parameter $r = 1.367$ Å because each vertice (atom) merges into an edge (double bond). In the $\mathcal{P}_1$ panel, 30 bar (edges) show up at $r = 1.367$ Å, corresponding to double bonds. These bars merge into triangles, which are formed due to new edges connecting a vertex $v_i$ to $v_{i+2}$ under a cyclic ordering of the hexagonal vertices. These new edges have length of 2.443 Å, which corresponds to the larger ending values of these orange bars. Such persistent facet ideals explains the existence of these 30 edges. A set of 60 pars appears at filtration $r = 1.453$ Å corresponding to these single bonds. These edges persist until new edges are created to connect vertex $v_i$ to $v_{i+2}$ under a cyclic ordering of the pentagonal vertices, where new edges have a length of 2.352 Å. This accounts for the observed variation in the birth times of persistent facet ideals corresponding to these edges, as shown in the $\mathcal{P}_1$ panel. Finally, $\mathcal{P}_2$ shows the birth and death of triangles in the system. Therefore, facet persistence


https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

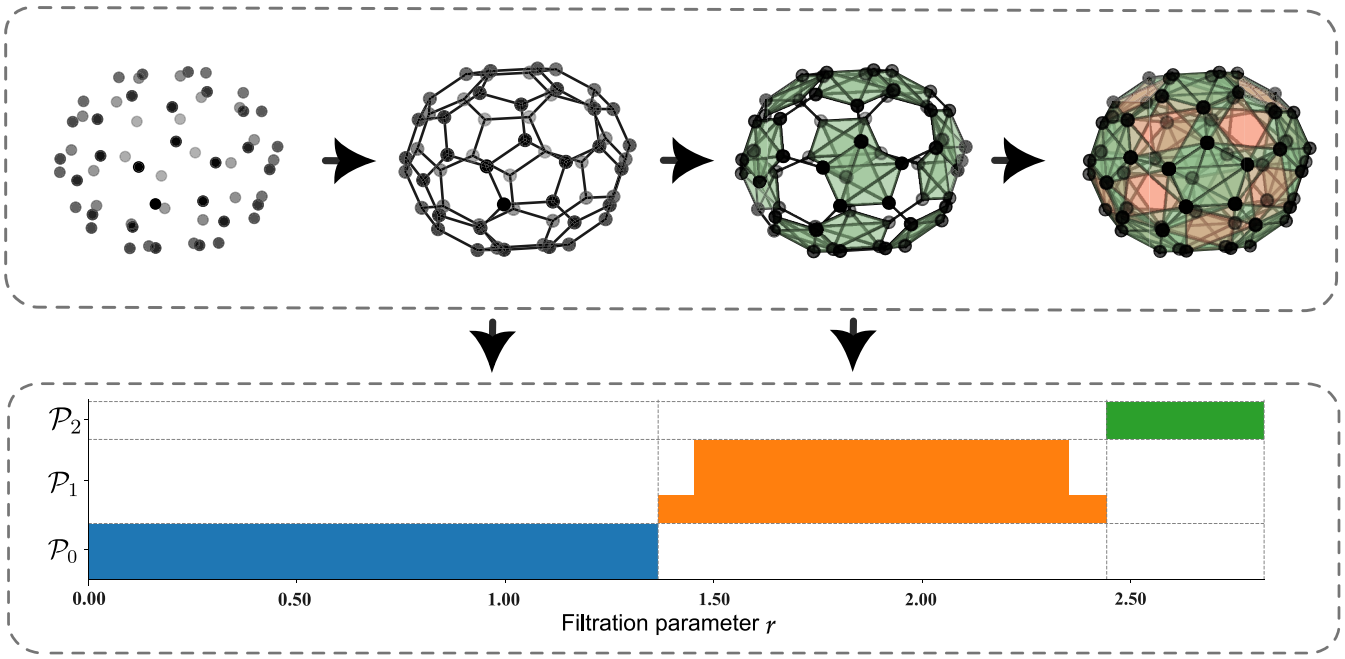



**Figure 4.** Illustration of the filtration process in the persistent commutative algebra. Given a point cloud input (a $C_{60}$ molecule), a corresponding simplicial complex with associated filtration is constructed (upper panel). Facet persistence barcodes are then computed across various dimensions (lower panel).

barcodes reveal the geometric connectivity of molecular structures and are interpretable, making commutative algebra a rational approach for machine learning.

In addition, the element- and category-specific approaches used in this work capture key intramolecular and intermolecular interactions that play crucial roles in determining the chemophysical properties of biomolecules. These essential interactions include hydrogen bonding, hydrophobic and hydrophilic interactions, van der Waals forces, electrostatic forces, and others. For example, C−C specific atoms primarily contribute to hydrophobic interactions, while N−N and N−O specific atoms are associated with hydrogen bonding, and O−C interactions often reflect electrostatic effects. In particular, these interactions exhibit different distance-dependent strengths. To effectively capture these critical interactions, we integrate persistent Stanley−Reisner complexes with various atom groupings. Altogether, this analysis offers a highly interpretable insight into the molecular structure, demonstrating the suitability of persistent facet ideals for molecular description and characterization and improving the performance of machine-learning models.

Persistent homology[35] and persistent Laplacian theory[25] have become popular and competitive approaches in machine learning modeling. While our persistent commutative algebra shares similarities with persistent homology in terms of a simplicial complex, it captures data patterns from a distinct perspective. Persistent homology focuses on the persistence of topological invariants (such as points, loops, and cavities), providing global structural information across the filtration process. In contrast, our persistent commutative algebra reveals more intrinsic local properties of simplices of various dimensions (including points, edges, and triangles) throughout the filtration. Compared to persistent Laplacian theory,[25,57] our method also demonstrates greater computational feasibility. Persistent Laplacian theory extracts data patterns based on the eigenvalues of Laplacian matrices, a process that is computationally demanding. Ultimately, each approach offers unique strengths and may be better suited to different types of data and tasks.

## CONCLUSIONS

As a foundational part of algebraic geometry and algebraic number theory, commutative algebra studies polynomial rings, their ideals, and the modules over such rings. However, commutative algebra has rarely been applied to data science and machine learning. The persistent Stanley−Reisner theory (PSRT),[31] introduced by Suwayyid and Wei, offers a new opportunity to develop commutative algebra machine learning (CAML) and commutative algebra deep learning (CADL) for data. Stanley−Reisner theory, also known as face ring theory, creates a profound connection between combinatorics and commutative algebra. PSRT integrates tools from algebra, combinatorics, and multiscale analysis (i.e., filtration) to study simplicial complexes via Stanley−Reisner rings.

This work proposes CAML for data analysis. We pair PSRT with a robust machine-learning method, gradient-boosted decision trees (GBDT), which utilize an ensemble of decision trees to make predictions. GBDT is known for its high accuracy and efficiency, particularly for relatively small data sets that are not suitable for deep learning algorithms. We consider two biomolecular data sets, i.e., a protein−ligand binding data set (PDBbind-v2016) and a metalloprotein−ligand binding data set, to validate the proposed CAML model. Due to the intricate interactions in (metallo)protein−ligand complexes, we propose new algorithms, such as commutative algebra on a bipartite complex, element-specific commutative algebra, and category-specific commutative algebra, to capture the physics and chemistry underlying the interactions. The performance of the proposed CAML model is compared to other state-of-the-art methods in the literature. We demonstrate that CAML is an extremely promising new method for protein−ligand binding predictions.



https://doi.org/10.1021/acs.jcim.5c00940
J. Chem. Inf. Model. 2025, 65, 6732−6743

Protein−ligand binding prediction serves as a case study of the proposed CAML model. CAML can easily be applied to other biomolecular data predictions and general problems in science and engineering. We believe that proposed commutative algebra learning (CAL) represents an emerging direction in machine learning and data science.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

All data and the code needed to reproduce this paper's result can be found at https://github.com/WeilabMSU/CAML. For detailed information on the metalloprotein−ligand complex data set, please refer to the reference.[40] The PDBbind-v2020 data set is available at http://pdbbind.org.cn/.

## ■ AUTHOR INFORMATION


### Corresponding Author

**Guo-Wei Wei** − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; Department of Biochemistry and Molecular Biology and Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-5781-2937; Email: weig@msu.edu

### Authors

**Hongsong Feng** − *Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223, United States;* orcid.org/0000-0001-8039-3059

**Faisal Suwayyid** − *Department of Mathematics, King Fahd University of Petroleum and Minerals, Dhahran 31261, KSA; Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

**Mushal Zia** − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-4583-5437

**JunJie Wee** − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0001-8444-3252

**Yuta Hozumi** − *Department of Physiology, Michigan State University, East Lansing, Michigan 48824, United States;* Present Address: School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0003-4674-2379

**Chun-Long Chen** − *Physical Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-5584-824X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00940


### Author Contributions

H.F. developed codes, performed the computational work, analyzed the results, and drafted the manuscript. F.S. contributed to method development, code implementation, and manuscript drafting. M.Z., J.J.W., Y.H., and C.C. analyzed data, and revised the manuscript. C.C. also obtained funding. G.-W.W. conceptualized and supervised the project, drafted the manuscript, and obtained funding.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01AI164266 and R35GM148196, NSF grant DMS-2052983, MSU Foundation, and Bristol-Myers Squibb 65109. F.S. thanks King Fahd University of Petroleum and Minerals for their support. C.-L.C. gratefully acknowledges financial support from the Defense Threat Reduction Agency (Project CB11141), and the Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (BES) under an award FWP 80124 at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated for the Department of Energy by Battelle under Contract DE-AC05-76RL01830. C.-L. C. and G.-W. W would like to also acknowledge financial support from DOE, Office of Science, BES under an award FWP 84274 for the development of machine learning models.

## ■ REFERENCES


(1) Fleming, N. Computer-calculated compounds. *Nature* **2018**, *557*, S55−S57.

(2) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224−229.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.

(4) Pinzi, L.; Rastelli, G. Molecular docking: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* **2019**, *20*, 4331.

(5) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **2017**, *9*, 91−102.

(6) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.

(7) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(8) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(9) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(10) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.

(11) Song, Y.; Wang, L. Multiobjective tree-based reinforcement learning for estimating tolerant dynamic treatment regimes. *Biometrics* **2024**, *80*, ujad017.

(12) Luo, J.; Wei, W.; Waldispühl, J.; Moitessier, N. Challenges and current status of computational methods for docking small molecules to nucleic acids. *Eur. J. Med. Chem.* **2019**, *168*, 414−425.

(13) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(14) Ben Bdira, F.; Dame, R. T. AI is a viable alternative to high throughput screening: a 318-target study. *Sci. Rep.* **2024**, *14*, No. 7526.

(15) Gómez-Sacristán, P.; Simeon, S.; Tran-Nguyen, V.-K.; Patil, S.; Ballester, P. J. Inactive-enriched machine-learning models exploiting patent data improve structure-based virtual screening for PDL1

dimerizers. *J. Adv. Res.* **2025**, *67*, 185−196, DOI: 10.1016/j.jare.2024.01.024.
(16) Hu, X.; Pang, J.; Chen, C.; Jiang, D.; Shen, C.; Chai, X.; Yang, L.; Zhang, X.; Xu, L.; Cui, S.; et al. Discovery of novel non-steroidal selective glucocorticoid receptor modulators by structure-and IGN-based virtual screening, structural optimization, and biological evaluation. *Eur. J. Med. Chem.* **2022**, *237*, 114382.
(17) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **2023**, *123*, 8736−8780.
(18) Jiang, J.; Zhang, C.; Ke, L.; Hayes, N.; Zhu, Y.; Qiu, H.; Zhang, B.; Zhou, T.; Wei, G.-W. A review of machine learning methods for imbalanced data challenges in chemistry. *Chem. Sci.* **2025**, *16*, 7637−7658, DOI: 10.1039/D5SC00270B.
(19) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(20) Wang, C.; Zhang, Y. Improving scoring-docking-screening powers of protein−ligand scoring functions using random forest. *J. Comput. Chem.* **2017**, *38*, 169−177.
(21) Li, C.; Zhang, A.; Wang, L.; Zuo, J.; Zhu, C.; Xu, J.; Wang, M.; Zhang, J. Z. Development of a polynomial scoring function P3-Score for improved scoring and ranking powers. *Chem. Phys. Lett.* **2023**, *824*, 140547.
(22) Nguyen, D. D.; Cang, Z.; Wei, G.-W. A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4343−4367.
(23) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, *14*, e1005929.
(24) Nguyen, D. D.; Cang, Z.; Wu, K.; Wang, M.; Cao, Y.; Wei, G.-W. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 71−82.
(25) Wang, R.; Nguyen, D. D.; Wei, G.-W. Persistent spectral graph. *Int. J. Numer. Methods Biomed. Eng.* **2020**, *36*, e3376.
(26) Meng, Z.; Xia, K. Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **2021**, *7*, eabc5329.
(27) Chen, D.; Liu, J.; Wei, G.-W. Multiscale topology-enabled structure-to-sequence transformer for protein−ligand interaction predictions. *Nat. Mach. Intell.* **2024**, *6*, 799−810.
(28) Nguyen, D. D.; Gao, K.; Wang, M.; Wei, G.-W. MathDL: mathematical deep learning for D3R Grand Challenge 4. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 131−147.
(29) Miller, E.; Sturmfels, B. *Combinatorial Commutative Algebra*, Graduate Texts in Mathematics; Springer, 2005; Vol. *227*.
(30) Eisenbud, D. *Commutative Algebra: With a View Toward Algebraic Geometry*; Springer Science & Business Media, 2013; Vol. *150*.
(31) Suwayyid, F.; Wei, G.-W. Persistent Stanley−Reisner Theory. 2025, arXiv:physics/2503.23482. arXiv.org e-Print archive. http://arxiv.org/abs/2503.23482.
(32) Stanley, R. P. *Combinatorics and Commutative Algebra*, 2nd ed.; Birkhäuser, 1996; Vol. *41*.
(33) Bruns, W.; Herzog, J. *Cohen-Macaulay Rings*, Cambridge Studies in Advanced Mathematics; Cambridge University Press, 1998; Vol. *39*.
(34) Hà, H. T.; Van Tuyl, A. Monomial ideals, edge ideals of hypergraphs, and their graded Betti numbers. *J. Algebr. Comb.* **2008**, *27*, 215−245.
(35) Edelsbrunner, H.; Harer, J. *Surveys on Discrete and Computational Geometry: Twenty Years Later*; American Mathematical Society, 2008; Vol. *453*, pp 257−282.
(36) Zomorodian, A.; Carlsson, G. Computing Persistent Homology. *Discret. Comput. Geom.* **2005**, *33*, 249−274.
(37) Cang, Z.; Wei, G.-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **2017**, *13*, e1005690.
(38) Papamarkou, T.; Birdal, T.; Bronstein, M.; Carlsson, G.; Curry, J.; Gao, Y.; Hajij, M.; Kwitt, R.; Lio, P.; Di Lorenzo, P. et al. Position: Topological deep learning is the new frontier for relational learning. 2024, arXiv:physics/2402.08871. arXiv.org e-Print archive. http://arxiv.org/abs/2402.08871.
(39) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405−412.
(40) Jiang, D.; Ye, Z.; Hsieh, C.-Y.; Yang, Z.; Zhang, X.; Kang, Y.; Du, H.; Wu, Z.; Wang, J.; Zeng, Y.; et al. MetalProGNet: a structure-based deep graph model for metalloprotein−ligand interaction predictions. *Chem. Sci.* **2023**, *14*, 2054−2069.
(41) Liu, R.; Liu, X.; Wu, J. Persistent Path-Spectral (PPS) Based Machine Learning for Protein−Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2023**, *63*, 1066−1075.
(42) Kyro, G. W.; Smaldone, A. M.; Shee, Y.; Xu, C.; Batista, V. S. T-ALPHA: A Hierarchical Transformer-Based Deep Neural Network for Protein−Ligand Binding Affinity Prediction with Uncertainty-Aware Self-Learning for Protein-Specific Alignment. *J. Chem. Inf. Model.* **2025**, *65*, 2395−2415.
(43) Zhang, X.; Gao, H.; Wang, H.; Chen, Z.; Zhang, Z.; Chen, X.; Li, Y.; Qi, Y.; Wang, R. Planet: a multi-objective graph neural network model for protein−ligand binding affinity prediction. *J. Chem. Inf. Model.* **2024**, *64*, 2205−2220.
(44) Shen, C.; Zhang, X.; Deng, Y.; Gao, J.; Wang, D.; Xu, L.; Pan, P.; Hou, T.; Kang, Y. Boosting protein−ligand binding pose prediction and virtual screening based on residue-atom distance likelihood potential and graph transformer. *J. Med. Chem.* **2022**, *65*, 10691−10706.
(45) Andreini, C.; Bertini, I.; Rosato, A. Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.* **2009**, *42*, 1471−1479.
(46) Banci, L.; Bertini, I. *Metallomics and the Cell*; Springer, 2012; pp 1−13.
(47) Chalkley, M. J.; Mann, S. I.; DeGrado, W. F. De novo metalloprotein design. *Nat. Rev. Chem.* **2022**, *6*, 31−50.
(48) Valasatava, Y.; Rosato, A.; Furnham, N.; Thornton, J. M.; Andreini, C. To what extent do structural changes in catalytic metal sites affect enzyme function? *J. Inorg. Biochem.* **2018**, *179*, 40−53.
(49) Hay, R. W. *Bio-inorganic Chemistry*; Ellis Horwood Chichester, 1984; Vol. *10*.
(50) Waldron, K. J.; Robinson, N. J. How do bacterial cells ensure that metalloproteins get the correct metal? *Nat. Rev. Microbiol.* **2009**, *7*, 25−35.
(51) Wang, Y.; Liu, X.; Zhang, Y.; Wang, X.; Xia, K. Join Persistent Homology (JPH)-Based Machine Learning for Metalloprotein−Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2025**, *65* (6), 2785−2793, DOI: 10.1021/acs.jcim.4c02309.
(52) Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 2791−2802.
(53) Durrant, J. D.; McCammon, J. A. NNScore 2.0: a neural-network receptor−ligand scoring function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.
(54) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2016239118.
(55) Chen, D.; Zheng, J.; Wei, G.-W.; Pan, F. Extracting predictive representations from hundreds of millions of molecules. *J. Phys. Chem. Lett.* **2021**, *12*, 10793−10801.
(56) Shen, L.; Feng, H.; Qiu, Y.; Wei, G.-W. SVSBI: sequence-based virtual screening of biomolecular interactions. *Commun. Biol.* **2023**, *6*, 536.

(57) Wei, X.; Wei, G.-W. Persistent Topological Laplacians—A Survey. *Mathematics* **2025**, *13*, 208.



https://doi.org/10.1021/acs.jcim.5c00940
*J. Chem. Inf. Model.* 2025, 65, 6732−6743