


# Improving performance of screening MM/PBSA in protein-ligand interactions via machine learning

Yuan-Qiang Chen(陈远强), Yao Xu(徐耀), Yu-Qiang Ma(马余强), and Hong-Ming Ding(丁泓铭)

**Citation:**Chin. Phys. B, 2025, 34 (**1**): 018701. DOI: 10.1088/1674-1056/ad8ecb

**What follows is a list of articles you may be interested in**

---

## Combination of density-clustering and supervised classification for event identification in single-molecule force spectroscopy data

Yongyi Yuan(袁泳怡), Jialun Liang(梁嘉伦), Chuang Tan(谭创), Xueying Yang(杨雪滢), Dongni Yang(杨东尼), and Jie Ma(马杰)

Chin. Phys. B, 2023, 32 (**10**): 108702. DOI: 10.1088/1674-1056/acf03e

## Evaluation on performance of MM/PBSA in nucleic acid-protein systems

Yuan-Qiang Chen(陈远强), Yan-Jing Sheng(盛艳静), Hong-Ming Ding(丁泓铭), and Yu-Qiang Ma(马余强)

Chin. Phys. B, 2022, 31 (**4**): 048701. DOI: 10.1088/1674-1056/ac3a5c

## Methods and applications of RNA contact prediction

Huiwen Wang(王慧雯) and Yunjie Zhao(赵蕴杰)†

Chin. Phys. B, 2020, 29 (**10**): 108708. DOI: 10.1088/1674-1056/abb7f3

## Computational prediction of RNA tertiary structures using machine learning methods

Bin Huang(黄斌), Yuanyang Du(杜渊洋), Shuai Zhang(张帅), Wenfei Li(李文飞), Jun Wang (王骏), and Jian Zhang(张建)†

Chin. Phys. B, 2020, 29 (**10**): 108704. DOI: 10.1088/1674-1056/abb303

## Individual identification using multi-metric of DTI in Alzheimer's disease and mild cognitive impairment

Ying-Teng Zhang(张应腾), Shen-Quan Liu(刘深泉)

Chin. Phys. B, 2018, 27 (**8**): 088702. DOI: 10.1088/1674-1056/27/8/088702

---

# Improving performance of screening MM/PBSA in protein–ligand interactions via machine learning

Yuan-Qiang Chen(陈远强)[1], Yao Xu(徐耀)[2], Yu-Qiang Ma(马余强)[2], and Hong-Ming Ding(丁泓铭)[1,†]

[1] *Center for Soft Condensed Matter Physics and Interdisciplinary Research, School of Physical Science and Technology, Soochow University, Suzhou* 215006, *China*

[2] *National Laboratory of Solid State Microstructures and Department of Physics, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing* 210093, *China*

(Received 29 September 2024; revised manuscript received 20 October 2024; accepted manuscript online 5 November 2024)

Accurately estimating protein–ligand binding free energy is crucial for drug design and biophysics, yet remains a challenging task. In this study, we applied the screening molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA) method in combination with various machine learning techniques to compute the binding free energies of protein–ligand interactions. Our results demonstrate that machine learning outperforms direct screening MM/PBSA calculations in predicting protein–ligand binding free energies. Notably, the random forest (RF) method exhibited the best predictive performance, with a Pearson correlation coefficient ($r_p$) of 0.702 and a mean absolute error (MAE) of 1.379 kcal/mol. Furthermore, we analyzed feature importance rankings in the gradient boosting (GB), adaptive boosting (AdaBoost), and RF methods, and found that feature selection significantly impacted predictive performance. In particular, molecular weight (MW) and van der Waals (VDW) energies played a decisive role in the prediction. Overall, this study highlights the potential of combining machine learning methods with screening MM/PBSA for accurately predicting binding free energies in biosystems.

**Keywords:** molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA), binding free energy, machine learning, protein–ligand interaction

**PACS:** 87.10.TF, 87.15.–v, 87.15.A–, 87.15.kp

**DOI:** 10.1088/1674-1056/ad8ecb **CSTR:** 32038.14.CPB.ad8ecb

## 1. Introduction

Molecular recognition plays a critical role in numerous biological processes, including self-replication, information processing, and gene expression.[1–6] Key examples of molecular recognition include protein–protein, protein–nucleic acid, and protein–ligand interactions.[1,6–8] Among these, protein–ligand interactions are ubiquitous in biological processes and are essential for drug discovery.[5,9,10] One of the primary challenges is identifying ligands that can bind with high affinity to modulate the activity of target proteins. Typically, the strength of protein–ligand binding is assessed by changes in free energy during the binding process. Therefore, accurate and efficient calculation of binding free energy is essential for the optimal design of drug molecules or ligands.

With advancements in experimental techniques, computing power, and algorithm development, research on the molecular mechanisms underlying protein structure and function at the atomic level has gained increasing attention. To accurately calculate protein–ligand binding free energies, several computational methods have been developed, such as the free energy perturbation (FEP) method,[11,12] molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA), and molecular mechanics/generalized Born surface area (MM/GBSA) methods.[13–16] Although FEP is accurate, it can be computationally expensive and difficult to converge.[17,18] In contrast, MM/PBSA and MM/GBSA are computationally efficient and relatively accurate but may perform poorly in highly charged systems.[15,16,19,20] To solve this problem, our group has developed an improved MM/PBSA method that accounts for ion shielding effects in electrostatic protein interactions, providing a reliable approach for calculating binding energies in biological systems.[21–25]

As crystallography and calorimetry data accumulate, the usage of advanced machine learning methods to predict binding energies and their thermodynamic components has become feasible.[26–29] Machine learning has gained attention for its efficiency and reliability in predicting drug–target interactions.[30–37] Typically, chemical and biological descriptors derived from system architecture are used as inputs, and models are built using various machine learning methods to predict interactions beyond the input set. Compared to traditional computational methods, machine learning offers simplicity, efficiency, and superior predictive performance, making it well-suited for large databases.[26] Therefore, this study aims to explore the integration of different machine learning approaches with the screening MM/PBSA method to enhance the accuracy and performance of protein–ligand binding energy predictions, and further to identify to most suit-

†Corresponding author. E-mail: dinghm@suda.edu.cn

© 2025 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

able machine learning method that complements the screening MM/PBSA approach.

## 2. Model and methods

### 2.1. Preparation of the datasets for protein–ligand interactions

To systematically evaluate and compare the performance of machine learning methods for protein–ligand binding free energy prediction, we selected a total of 250 protein–ligand complexes, all of which were randomly selected from the protein–ligand refined set in the PDBbind website[38] (http://www.pdbbind-cn.org). To ensure that the selected 250 protein–ligand complexes is representative of the overall protein–ligand refinement set, we used ligand molecular weight, total number of atoms, and polarized surface area as criteria for selection, and found no significant differences in the distribution of ligand properties between the two datasets (see Fig. S1 in supporting information). For the sake of comparison, we divided the data sets into two systems based on whether the ligands were charged or not. There were 125 protein–ligand complexes in the neutral-ligand system and 125 protein–ligand complexes in the charged–ligand system, as detailed in Tables 1 and 2.

**Table 1.** PDB identification, pKd value, and experimental binding free energy ($\Delta G_{exp}$) of protein–ligand (neutral) system.

| PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd | PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd | PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd |
|---|---|---|---|---|---|---|---|---|
| 1E66 | –13.49 | 9.89 | 1LGT | –8.32 | 5.92 | 2P95 | –12.78 | 9.37 |
| 186L | –6.62 | 4.85 | 1LHU | –12.04 | 4.48 | 2POG | –13.01 | 9.54 |
| 1A28 | –11.31 | 8.29 | 1LNM | –11.87 | 6.58 | 2POU | –10.12 | 7.42 |
| 1AJX | –10.79 | 7.91 | 1M2R | –8.81 | 5.10 | 2QE4 | –10.86 | 7.96 |
| 1BCU | –4.47 | 3.28 | 1MES | –10.50 | 5.94 | 2QNN | –9.75 | 7.15 |
| 1BDQ | –8.65 | 6.34 | 1MMQ | –10.26 | 10.39 | 2STD | –13.44 | 9.85 |
| 1BGQ | –11.69 | 8.57 | 1N46 | –14.35 | 7.74 | 2USN | –8.88 | 6.51 |
| 1BN1 | –12.74 | 9.34 | 1NDW | –7.13 | 6.10 | 2UWD | –11.39 | 8.35 |
| 1BNN | –13.64 | 10.00 | 1NDY | –8.42 | 8.83 | 2UWO | –11.87 | 8.70 |
| 1BNT | –13.37 | 9.80 | 1NVS | –0.67 | 8.70 | 2VES | –10.50 | 7.70 |
| 1BNW | –12.39 | 9.08 | 1OGX | –7.98 | 6.11 | 2VVV | –11.21 | 8.22 |
| 1BWA | –11.73 | 8.60 | 1OGZ | –9.34 | 6.10 | 2VW5 | –11.62 | 8.52 |
| 1D4I | –12.07 | 8.85 | 1Q7A | –9.81 | 7.70 | 2W26 | –12.82 | 9.40 |
| 1D4J | –11.40 | 8.36 | 1QBN | –7.98 | 7.52 | 2WC3 | –8.29 | 6.08 |
| 1DMP | –13.03 | 9.55 | 1QKT | –12.33 | 10.52 | 2XJ1 | –10.30 | 7.55 |
| 1E2K | –6.74 | 4.94 | 1QL9 | –7.30 | 5.23 | 2YDW | –10.01 | 7.34 |
| 1E2L | –5.85 | 4.29 | 1RBP | –9.17 | 6.17 | 3AS3 | –5.18 | 3.80 |
| 1E3G | –12.28 | 9.00 | 1RJK | –13.64 | 7.82 | 3B67 | –12.08 | 8.85 |
| 1EBW | –12.34 | 9.05 | 1RO6 | –8.81 | 6.85 | 3B68 | –11.46 | 8.40 |
| 1EC0 | –11.58 | 8.49 | 1SR7 | –13.78 | 6.85 | 3CS7 | –14.57 | 10.68 |
| 1ELD | –9.14 | 6.70 | 1SYI | –7.42 | 7.19 | 3D91 | –12.48 | 9.15 |
| 1ELE | –9.34 | 6.85 | 1TTM | –10.03 | 5.85 | 3FFG | –12.39 | 9.08 |
| 1ERB | –11.05 | 8.10 | 1UWF | –9.30 | 9.04 | 3HB4 | –12.34 | 9.05 |
| 1F3E | –9.14 | 6.70 | 1V2K | –8.44 | 5.35 | 3KDM | –9.55 | 7.00 |
| 1F4X | –7.63 | 5.59 | 1VFN | –7.64 | 6.72 | 3KMY | –6.12 | 4.49 |
| 1FJS | –13.59 | 9.96 | 1X39 | –12.58 | 10.00 | 3OAF | –12.00 | 8.80 |
| 1FKG | –10.91 | 8.00 | 1XOW | –12.66 | 6.46 | 3UDD | –9.68 | 7.10 |
| 1FKI | –9.55 | 7.00 | 2ANS | –8.08 | 5.44 | 3WK9 | –5.66 | 4.15 |
| 1FQ5 | –11.46 | 8.40 | 2BRM | –8.03 | 6.82 | 4GHI | –9.67 | 7.09 |
| 1G45 | –11.79 | 8.64 | 2F1G | –10.29 | 6.19 | 4KNM | –10.91 | 8.00 |
| 1G54 | –12.03 | 8.82 | 2G94 | –12.99 | 5.60 | 4LM4 | –4.67 | 3.42 |
| 1GX8 | –8.69 | 6.37 | 2J2U | –10.00 | 8.80 | 4Q7V | –4.90 | 3.59 |
| 1I37 | –13.05 | 9.57 | 2J34 | –10.67 | 9.22 | 4TWP | –13.64 | 10.00 |
| 1IE9 | –13.90 | 10.19 | 2J4I | –12.28 | 6.96 | 4URZ | –3.44 | 2.52 |
| 1J01 | –8.83 | 6.47 | 2O4J | –13.78 | 5.92 | 4Z0Q | –11.14 | 8.17 |
| 1JAQ | –6.11 | 4.85 | 2O4R | –12.82 | 9.52 | 5MPK | –11.62 | 8.52 |
| 1JGL | –11.87 | 8.29 | 2OI0 | –10.30 | 7.54 | 5STD | –14.31 | 10.49 |
| 1K1I | –8.98 | 4.09 | 2OJG | –7.69 | 9.52 | 5U13 | –7.38 | 5.41 |
| 1KJR | –8.27 | 3.13 | 2OJJ | –11.87 | 7.33 | 5VCY | –8.96 | 6.57 |
| 1KV1 | –8.10 | 6.34 | 2P15 | –14.05 | 5.96 | 5W2S | –10.00 | 7.33 |
| 1KZK | –14.17 | 8.57 | 2P16 | –13.78 | 8.40 | 6DYU | –14.00 | 10.26 |
| 1LEE | –10.56 | 9.34 | 2P4J | –12.22 | 8.96 | | | |

**Table 2.** PDB identification, pKd value, and experimental binding free energy ($\Delta G_{exp}$) of protein–ligand (charged) system.

| PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd | PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd | PDB ID | $\Delta G_{exp}$ (kcal/mol) | pKd |
|---|---|---|---|---|---|---|---|---|
| 10GS | –8.73 | 6.40 | 1GFY | –6.93 | 5.08 | 2QBQ | –10.15 | 7.44 |
| 1ADL | –7.31 | 5.36 | 1GHV | –5.93 | 4.35 | 2QBR | –8.63 | 6.33 |
| 1AJP | –3.04 | 2.23 | 1GHZ | –6.55 | 4.80 | 2QBS | –9.11 | 6.68 |
| 1ATL | –8.57 | 6.28 | 1GI7 | –6.15 | 4.51 | 2QMG | –12.48 | 9.15 |
| 1B8Y | –10.71 | 7.85 | 1HEE | –7.28 | 5.34 | 2QPQ | –8.69 | 6.37 |
| 1BJU | –6.55 | 4.80 | 1HLK | –6.82 | 5.00 | 2RJP | –9.55 | 7.00 |
| 1BJV | –7.56 | 5.54 | 1HMR | –8.93 | 6.55 | 2V77 | –6.90 | 5.06 |
| 1BM7 | –10.26 | 7.52 | 1HMT | –7.90 | 5.79 | 2VKM | –11.92 | 8.74 |
| 1BTY | –6.47 | 4.74 | 1J14 | –6.12 | 4.49 | 2VVC | –11.62 | 8.52 |
| 1C5O | –4.77 | 3.49 | 1KSN | –12.82 | 9.40 | 2VWL | –9.58 | 7.02 |
| 1C5S | –8.18 | 6.00 | 1L8G | –8.48 | 6.22 | 2YFA | –6.92 | 5.07 |
| 1C83 | –6.62 | 4.85 | 1LPG | –9.67 | 7.09 | 2ZFT | –8.89 | 6.52 |
| 1C84 | –6.18 | 4.53 | 1N0S | –10.16 | 7.45 | 2ZGX | –9.20 | 6.74 |
| 1C86 | –6.41 | 4.70 | 1N4H | –8.93 | 6.55 | 2ZMM | –10.37 | 7.60 |
| 1C87 | –5.73 | 4.20 | 1NFY | –12.13 | 8.89 | 3BRN | –11.87 | 8.70 |
| 1CBX | –8.66 | 6.35 | 1NL9 | –8.13 | 5.96 | 3E3B | –10.86 | 7.96 |
| 1CIZ | –10.15 | 7.44 | 1NO6 | –6.02 | 4.41 | 3GY4 | –6.96 | 5.10 |
| 1CNW | –10.53 | 7.72 | 1NZ7 | –9.71 | 7.12 | 3HK1 | –8.42 | 6.17 |
| 1D3D | –12.40 | 9.09 | 1O2Q | –10.48 | 7.68 | 3P3S | –10.35 | 7.59 |
| 1D4P | –8.59 | 6.30 | 1O2Z | –8.33 | 6.11 | 3PE1 | –13.16 | 9.65 |
| 1D6V | –8.42 | 6.17 | 1O36 | –8.13 | 5.96 | 3Q6K | –11.42 | 8.37 |
| 1D9I | –12.43 | 9.11 | 1O38 | –9.30 | 6.82 | 4BAQ | –12.29 | 9.01 |
| 1DF8 | –13.23 | 9.70 | 1ONZ | –6.96 | 5.10 | 4BBG | –11.35 | 8.32 |
| 1E3V | –5.92 | 4.34 | 1OWE | –8.46 | 6.20 | 4ERF | –12.82 | 9.40 |
| 1ECV | –6.62 | 4.85 | 1QB6 | –8.27 | 6.06 | 4I8X | –3.60 | 2.64 |
| 1EJN | –7.67 | 5.62 | 1QF0 | –10.07 | 7.38 | 4JSR | –11.46 | 8.40 |
| 1ELA | –8.68 | 6.36 | 1QF1 | –9.98 | 7.32 | 4OCC | –11.87 | 8.70 |
| 1ELB | –9.75 | 7.15 | 1QF2 | –8.08 | 5.92 | 4QXO | –7.52 | 5.51 |
| 1EZQ | –12.34 | 9.05 | 1QXK | –6.89 | 5.05 | 4R4O | –7.94 | 5.82 |
| 1F0T | –8.12 | 5.95 | 1SQO | –10.18 | 7.46 | 5D24 | –7.15 | 5.24 |
| 1F0U | –9.77 | 7.16 | 1SV3 | –5.93 | 4.35 | 5E7N | –8.31 | 6.09 |
| 1F4F | –8.18 | 6.00 | 1SWG | –9.44 | 6.92 | 5IME | –11.38 | 8.34 |
| 1F4G | –8.84 | 6.48 | 1V2N | –8.05 | 5.90 | 5OGB | –10.19 | 7.47 |
| 1F73 | –3.26 | 2.39 | 1XK9 | –9.34 | 6.85 | 5OTR | –6.71 | 4.92 |
| 1FCX | –9.81 | 7.19 | 2AZR | –4.97 | 3.64 | 5U4D | –8.63 | 6.33 |
| 1FCY | –11.62 | 8.52 | 2B07 | –8.77 | 6.43 | 6CEF | –5.10 | 3.74 |
| 1FCZ | –12.58 | 9.22 | 2C80 | –7.27 | 5.33 | 6CKW | –12.39 | 9.08 |
| 1FD0 | –11.46 | 8.40 | 2FR3 | –11.87 | 8.70 | 6DPX | –8.03 | 5.89 |
| 1FL3 | –9.28 | 6.80 | 2GKL | –7.30 | 5.35 | 6G2O | –9.23 | 6.77 |
| 1FLR | –13.64 | 10.00 | 2H4G | –8.89 | 6.52 | 6NFH | –9.75 | 7.15 |
| 1FV0 | –8.09 | 5.93 | 2H6B | –7.35 | 5.39 | 6QZ5 | –8.29 | 6.08 |
| 1G7G | –9.00 | 6.60 | 2HA2 | –6.36 | 4.66 | | | |

### 2.2. All-atom molecular dynamics (MD) simulations

In this study, the crystal structures of all complexes were obtained from the RCSB protein database (https://www.rcsb.org/).[39,40] All-atom MD simulations were carried out by using the GROMACS (version 2019.03) (http://wwwgromacs.org/) package.[41–43] To automatically process protein–ligand complexes, the tleap and antechamber modules in Amber20[44] were used to construct the topology files for the ligands, and the partial charges for each ligand were fitted using the AM1-BCC.[45,46] method. The GAFF force field[47] was employed for the small ligand molecules, and the Amber14SB[48] force field was used for the proteins. Each system (i.e., the protein–ligand complex) was solvated in TIP3P water[49] (the minimum distances from the surfaces of the box to the complex atoms were set to 15 Å), and NaCl was then added to neutralize the systems at a concentration of 0.15 M. The LINCS constraints[50] were used for all bonds involving hydrogen atoms. The periodic boundary conditions were adopted in all three directions. The particle mesh Ewald method[51,52] was used when calculating the long-range electrostatic interactions, and the Lennard–Jones (LJ) interactions were cut off at a distance of 1.0 nm.

Previous studies have demonstrated that when using MM/PBSA or MM/GBSA to predict protein–ligand binding free energy, the results derived from long time simulations

are similar to those calculated with minimized structures.[20,53] Therefore, in this study, we conducted our computational analysis based on the minimized structures. Next, during the simulations, the system was first energy-minimized using the steepest descent method for approximately $10^4$ cycles when the heavy atoms of the proteins were harmonically restrained with a spring constant of 1000 kJ·mol$^{-1}$·nm$^{-2}$. Then, the system was energy-minimized using the steepest descent method for approximately 10000 cycles by turning off all the constraints.

### 2.3. Screening MM/PBSA calculation of the protein–ligand acid binding energy

In standard MM/PBSA, the binding free energy $\Delta G_{\text{bind}}$ is calculated as follows:[15,53–55]

$$\Delta G_{\text{bind}} = \Delta E_{\text{ele}} + \Delta E_{\text{vdw}} + \Delta G_{\text{PB}} + \Delta G_{\text{SA}} - T\Delta S,$$

where $\Delta E_{\text{ele}}$ and $\Delta E_{\text{vdw}}$ are the electrostatic energy and van der Waals energy between the protein and the ligand at the gas-phase, respectively. $\Delta G_{\text{PB}}$ and $\Delta G_{\text{SA}}$ are the polar contribution and non-polar contribution of the desolvation free energy, respectively. $-T\Delta S$ is the entropy change upon binding, and is not considered here due to the high computational cost and low prediction accuracy.[56,57]

Notably, in the standard MM/PBSA, $\Delta E_{\text{ele}}$ is calculated using the Coulombic form; while in the screening MM/PBSA, an additional exponential damping factor is added to the Columbic term:[21,22]

$$E_{ij}^{\text{ele}} = \sum_i \sum_j \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{\text{in}} r_{ij}} \mathrm{e}^{-r_{ij}/\lambda_{\text{D}}},$$

where $q_i$ and $q_j$ are the charge of atom $i$ in the protein and the charge of atom $j$ in the ligand, respectively, $r_{ij}$ is the distance between atom $i$ and atom $j$, $\varepsilon_0$ is the dielectric constant in vacuum, $\varepsilon_{\text{in}}$ is the relative dielectric constant of the solute, and $\lambda_{\text{D}}$ is the electrostatic screening length. In this work, the salt (NaCl) concentration is set as 0.15 M, and the temperature is 298 K, so the electrostatic screening length is 0.8 nm.

The protein–ligand binding free energy was calculated by using the modified shell script gmx_mmpbsa.[58] The experimental binding energy was estimated as $\Delta G = k_{\text{B}} T \ln K_{\text{d}}$, where $k_{\text{B}}$ is the Boltzmann constant, $T$ is the temperature, and $K_{\text{d}}$ is the equilibrium dissociation constant.

### 2.4. Prediction of protein–ligand binding free energy by machine learning method

When employing machine learning methods to predict binding free energy, we randomly selected 25% of the protein–ligand dataset as the test set, while the remaining 75% served as the training set. During the prediction process, ten important features were identified (see Table 3). These features obtained based on the screening MM/PBSA method, which included the polar (PB) and non-polar (SA) solvation free energy components, electrostatic interaction energy (COUL), van der Waals interaction energy (VDW). We classify them as feature A (1–4). Previous studies have demonstrated that the charges of the protein (Pro_q) and ligand (Lig_q) significantly influence binding free energy.[21–24,59] Therefore, we introduced these charges as new features, classifying them as feature B (1–6). Additionally, we included four structural characteristics of ligands:[30,34,37] molecular weight (MW), total atomic number (Na), polarized surface area (PSA), and oil–water partition coefficient (LogP), which we categorized as feature C (1–10). For comparison, we also considered the VDW and MW from the above features as feature D. The specific features contained within the different feature sets are listed in Table S1 in supporting information. These feature quantities can be derived from molecular mechanics calculations and the PDBbind database.

In this paper, we selected seven different machine learning methods for model training and prediction, including linear regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), stochastic gradient descent (SGD), gradient boosting (GB), adaptive boosting (AdaBoost), and random forest (RF) methods.[26,30,31,37] These methods were implemented using the Scikit-learn library[60] in Python 3.12. For each model, we conducted a thorough grid search to optimize the hyperparameters. During the training phase, we employed cross-validation to rigorously assess the model's performance and mitigate the risk of overfitting, ensuring robustness and generalizability in our results. Finally, we used the Pearson correlation coefficient ($r_{\text{p}}$) to evaluate the predictive power of the models, while the mean absolute error (MAE) was utilized to compare the predicted binding free energies with experimental values. Both $r_{\text{p}}$ and MAE values were averaged over 1000 iterations.

**Table 3.** The name and description of features.

| Serial number | Name | Description |
|---|---|---|
| 1 | PB | the polar contribution contribution of the desolvation free energy |
| 2 | SA | the non-polar contribution of the desolvation free energy |
| 3 | COUL | electrostatic interaction energy |
| 4 | VDW | van der Waals interaction energy |
| 5 | Lig_q | net ligand charge |
| 6 | Pro_q | net protein charge |
| 7 | MW | ligand molecular weight |
| 8 | Na | ligand total atomic number |
| 9 | PSA | ligand polarized surface area |
| 10 | LogP | oil–water partition coefficient |

## 3. Results and discussion

We first calculated the binding free energies of 250 protein–ligand systems using the screening MM/PBSA method. Based on the findings of Zhu *et al.*,[23] we selected a solute dielectric constant $\varepsilon_{\text{in}}$ of 2.0 in this study. As shown

in Fig. 1, the Pearson correlation coefficient $r_p$ between the calculated and experimental values for charged systems was approximately 0.553, slightly higher than that of neutral systems, which was 0.547. The overall $r_p$ for the whole system (i.e., combined charged and neutral systems) was around 0.572. These results are consistent with those reported by Zhu *et al.*, who found the optimal Pearson correlation coefficient for protein–ligand interactions using Screening MM/PBSA to be between 0.5 and 0.6.[23]

We next applied linear regression to predict the binding free energies of protein–ligand systems. Linear regression is similar to the linear interaction energy (LIE) method,[61,62] where independent coefficients are fitted to each energy term (including VDW, ELE, PB, SA terms) to predict binding free energies. However, linear regression can incorporate a wider range of features beyond energy terms. As illustrated in Fig. 2, the linear regression fit, based on four energy terms calculated via screening MM/PBSA (feature A), showed Pearson correlation coefficients for charged, neutral, and overall systems were 0.626, 0.548, and 0.641, respectively. Although the linear regression model slightly improved prediction performance, the enhancement was not substantial. Adding new features (features B and C) did not significantly improve predictive performance, likely because some features, such as molecular weight (MW), contribute to binding free energy in a nonlinear manner. Including these features may even interfere with prediction accuracy, as seen with the slight decrease in the $r_p$ value for neutral systems.

As noted previously, linear regression is a basic machine learning method that struggles with nonlinear features. Advanced machine learning methods, such as support vector machines (SVM), *k*-nearest neighbors (KNN), stochastic gradient descent (SGD), gradient boosting (GB), adaptive boosting (AdaBoost), and random forest (RF), may better handle nonlinear relationships.

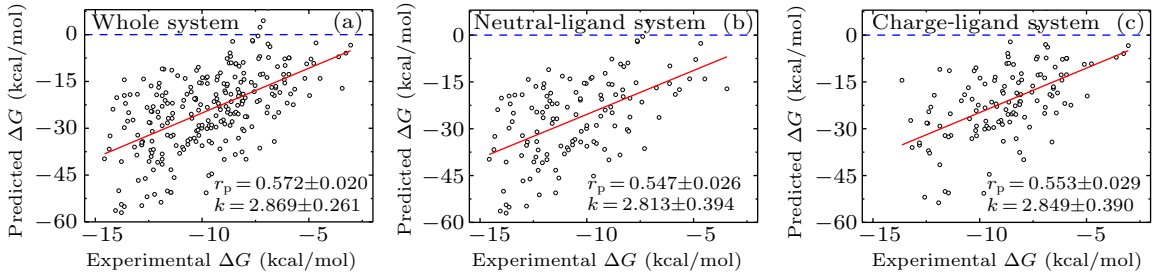



**Fig. 1.** Correlation between experimental measurements and the predicted binding free energies calculated using the screening MM/PBSA method. (a) The whole system, (b) the neutral-ligand system, (c) the charged-ligand system. The uncertainties of $r_p$ and $k$ are estimated by the bootstrapping algorithm (i.e., randomly selecting 80% of each dataset with the repeat times equaling the number of complexes in each dataset).

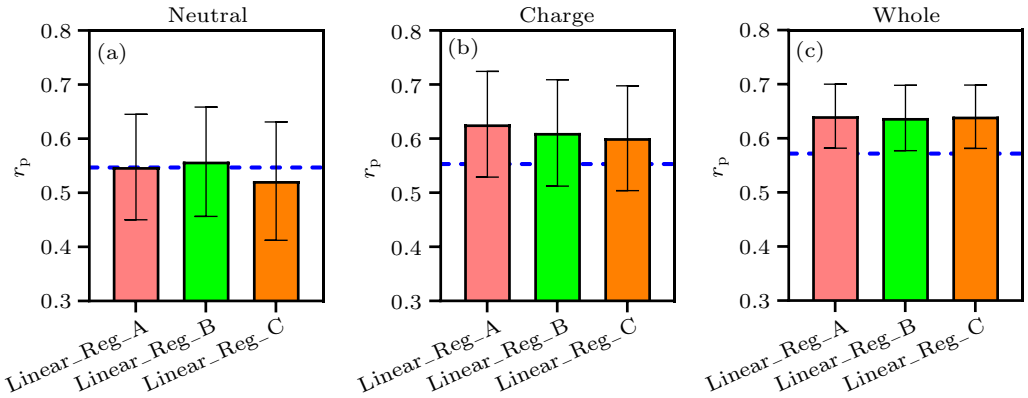



**Fig. 2.** Comparison of predicted values and experimental values of $r_p$ from the linear regression method across different datasets with the addition of various features. (a) The neutral-ligand system; (b) the charged-ligand system; (c) the whole system. Feature A corresponds to serial numbers 1–4 in Table 3, feature B to serial numbers 1–6, and feature C to serial numbers 1–10. The uncertainty estimate for $r_p$ is the standard deviation from 1000 repeated calculations, with the blue dashed lines representing results calculated using the screening MM/PBSA method across different datasets.

We then applied the six machine learning methods to assess their ability to improve the prediction of binding free energies. For neutral ligand systems, as shown in Fig. 3, the SVM method underperformed compared to linear regression across different features, while KNN and SGD exhibited similar trends and results to linear regression. In contrast, GB, AdaBoost, and RF outperformed linear regression. Interestingly, the results for feature A and feature B were similar, but the predictive performance significantly improved with the addition of more features in feature C. This trend was not observed for SVM, KNN, SGD, or linear regression, indicating that different models exhibit distinct behaviors when handling large feature sets. Linear regression and other simpler models may suffer from overfitting when too many features are included, whereas methods like random forest can select the most relevant features for prediction. For neutral-ligand sys-

tems, RF achieved the best performance, with $r_p$ and MAE values of 0.612 and 1.558 kcal/mol (feature A), 0.614 and 1.534 kcal/mol (feature B), and 0.685 and 1.465 kcal/mol (feature C).

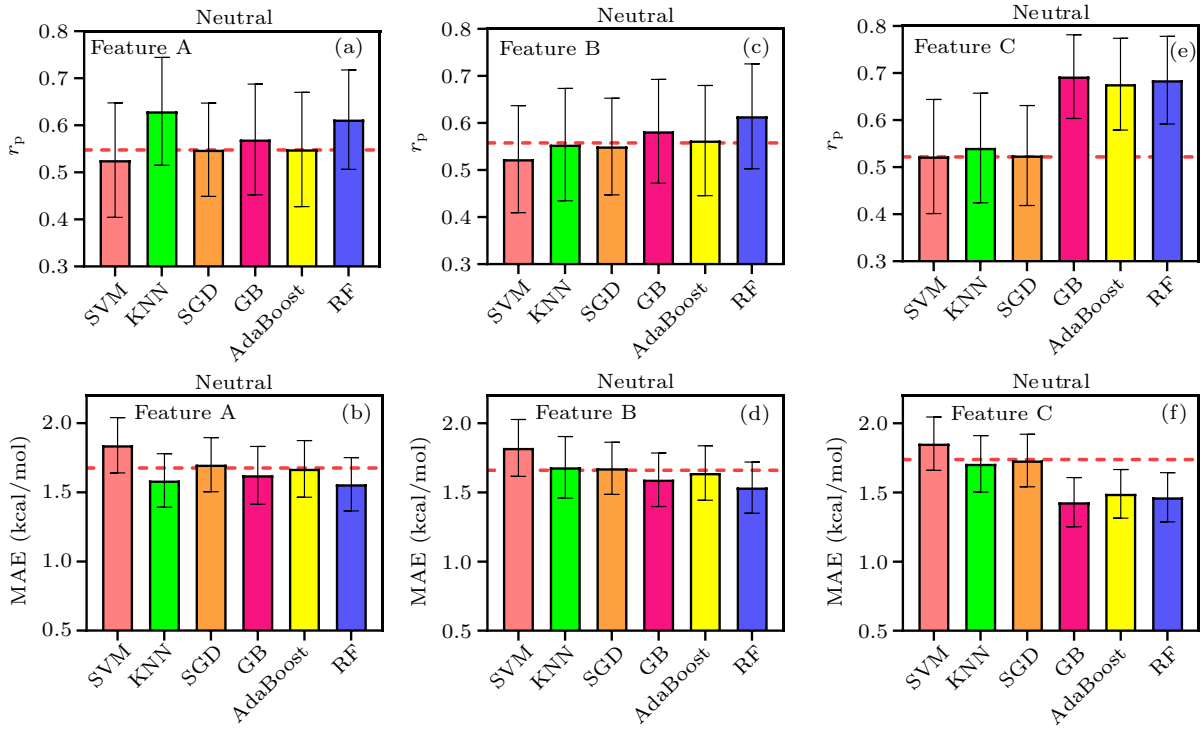



**Fig. 3.** Comparisons of predicted results from different machine learning methods when selecting various features in the neutral-ligand system. For feature set A, (a) $r_p$ and (b) MAE between predicted values and experimental values are shown; for feature set B, (c) $r_p$ and (d) MAE are presented; for feature set C, (e) $r_p$ and (f) MAE are also provided. The uncertainty estimates for $r_p$ and MAE are the standard deviations from 1000 repeated calculations, with the red dashed lines representing the predicted results from the linear regression algorithm in the neutral-ligand system.

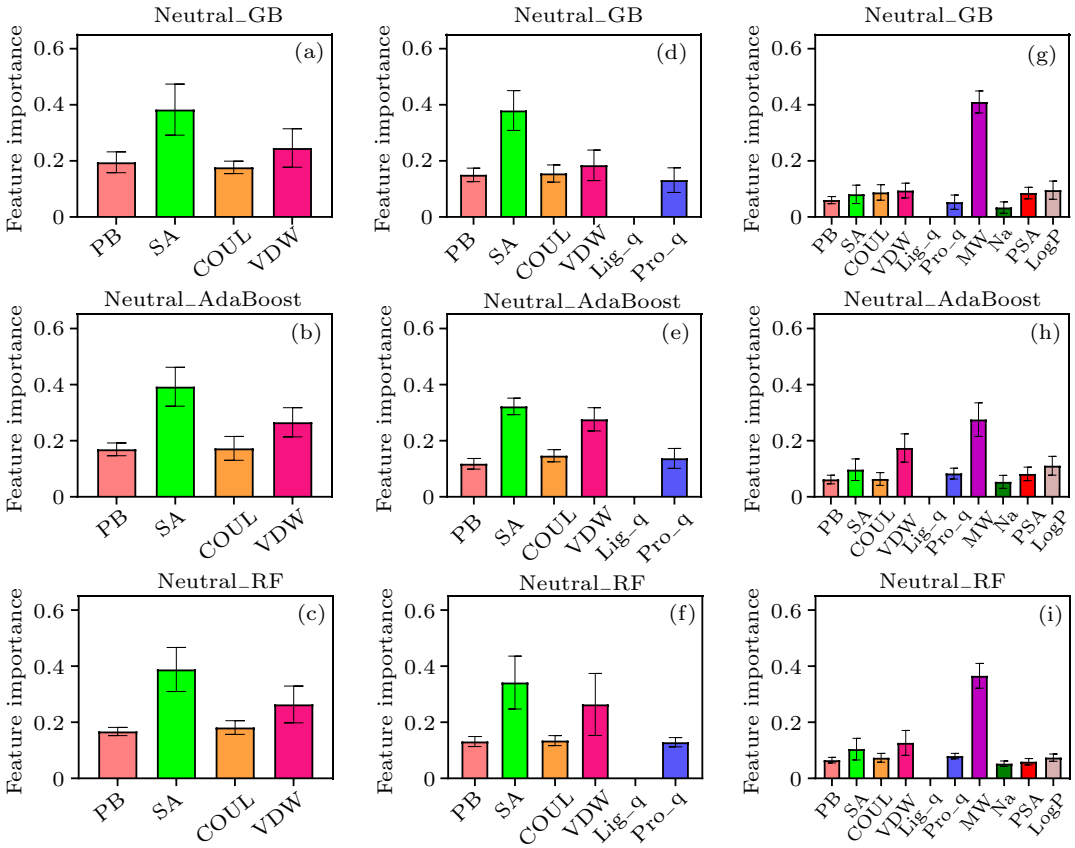



**Fig. 4.** Feature importance analysis for GB, AdaBoost, and RF methods in the neutral-ligand system under different feature sets. For feature set A, feature importance analyses are shown for (a) GB, (b) AdaBoost, and (c) RF; for feature set B, the analyses are presented for (d) GB, (e) AdaBoost, and (f) RF; for feature set C, the analyses are provided for (g) GB, (h) AdaBoost, and (i) RF.

Since GB, AdaBoost, and RF involve feature selection during prediction, we examined the importance rankings of features in these models (Fig. 4). Across all methods, surface area (SA) and van der Waals (VDW) interactions were the most significant contributors to prediction accuracy, especially when fewer features were used (Figs. 4(a)–4(f)). With the inclusion of additional features, molecular weight (MW) became dominant (Figs. 4(g)–4(i)), likely because larger neutral molecules tend to exhibit stronger interactions, leading to improved predictive performance.

For charged ligand systems, as shown in Fig. 5, only SGD performed comparably to linear regression. SVM and KNN performed slightly worse than linear regression. Interestingly, when using fewer features (Figs. 5(a)–5(d)), GB, AdaBoost, and RF underperformed compared to linear regression. However, as more features were included (Figs. 5(e)–5(f)), these methods showed improved performance, with RF achieving a lower MAE (although $r_p$ did not significantly increase), indicating that RF provides more accurate predictions. We also analyzed feature importance for charged ligand systems in the GB, AdaBoost, and RF models. As shown in Fig. 6, VDW interactions remained the most influential feature, particularly with smaller feature sets (Figs. 6(a)–6(f)). As more structural features were added, MW and Na gained importance, further enhancing predictive performance (Figs. 6(g)–6(i)). Overall, the prediction accuracy for charged systems was lower than for neutral systems, suggesting that charged systems are inherently more challenging to predict.[16,19,20,59] We hypothesize that molecular weight plays a less significant role in charged systems, as highly charged small molecules can exhibit strong binding affinities, limiting the contribution of MW to the overall binding energy.

Furthermore, we evaluated the predictive performance of different machine learning methods for the whole system, comprising both neutral and charged systems. As shown in Fig. 7, SGD performed similarly to linear regression, while SVM and KNN performed slightly worse. GB, AdaBoost, and RF exhibited increasing predictive performance with the addition of more features, with RF showing the best performance in feature C, achieving $r_p$ and MAE values of 0.702 and 1.379 kcal/mol, respectively, significantly outperforming linear regression. Feature importance analysis for the entire dataset (Fig. 8) revealed that VDW interactions contributed the most when fewer features were used (Figs. 8(a)–8(f)). With additional features (Figs. 8(g)–8(i)), MW became a key contributor, similar to the findings for neutral ligand systems.

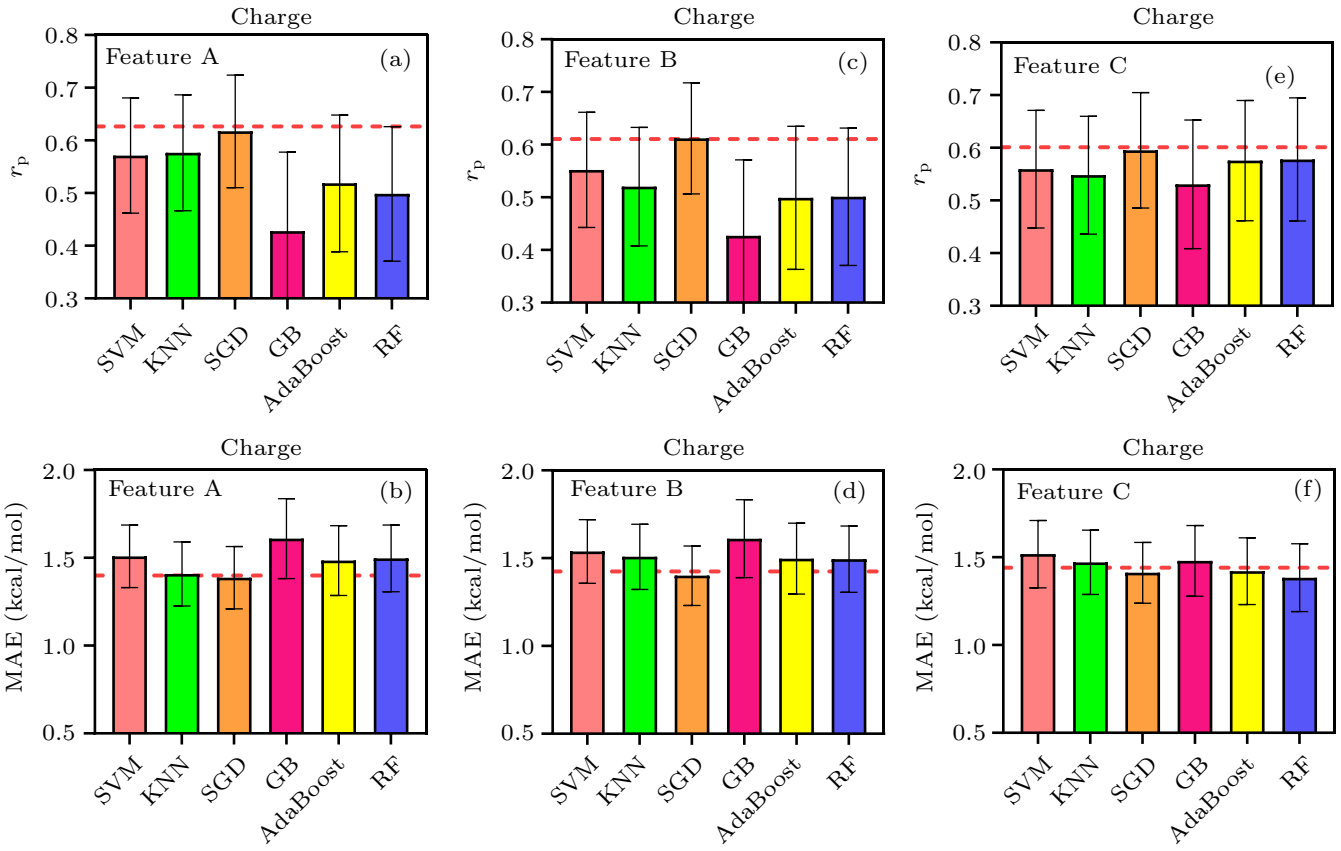



**Fig. 5.** Comparison of predicted results from different machine learning methods when selecting various features in the charge-ligand system. For feature set A, (a) $r_p$ and (b) MAE between predicted values and experimental values are shown; for feature set B, (c) $r_p$ and (d) MAE are presented; for feature set C, (e) $r_p$ and (f) MAE are also provided. The uncertainty estimates for the $r_p$ and MAE are the standard deviations from 1000 repeated calculations, with the red dashed lines representing the predicted results from the linear regression algorithm in the charge-ligand system.

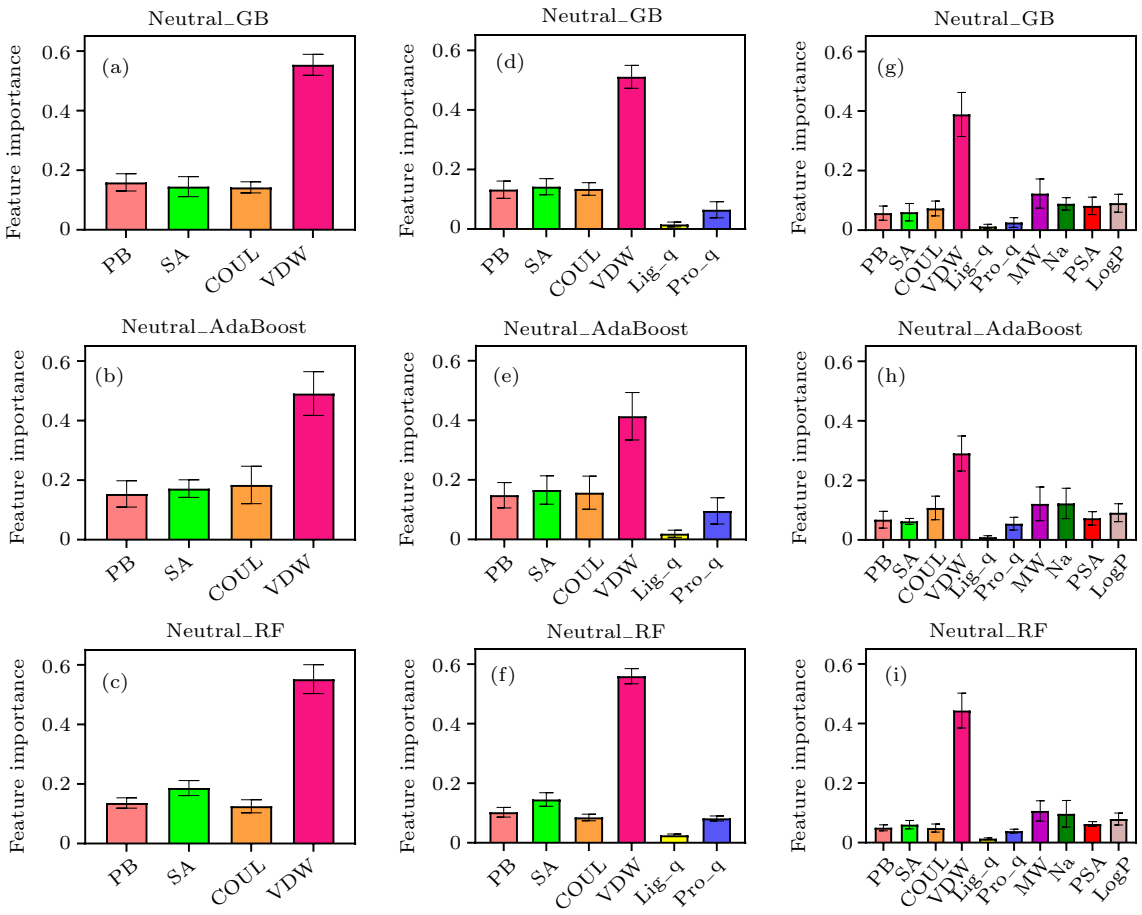



**Fig. 6.** Feature importance analyses for GB, AdaBoost, and RF methods in the charge-ligand system under different feature sets. For feature set A, feature importance analyses are shown for (a) GB, (b) AdaBoost, and (c) RF; for feature set B, the analyses are presented for (d) GB, (e) AdaBoost, and (f) RF; for feature set C, the analyses are provided for (g) GB, (h) AdaBoost, and (i) RF.

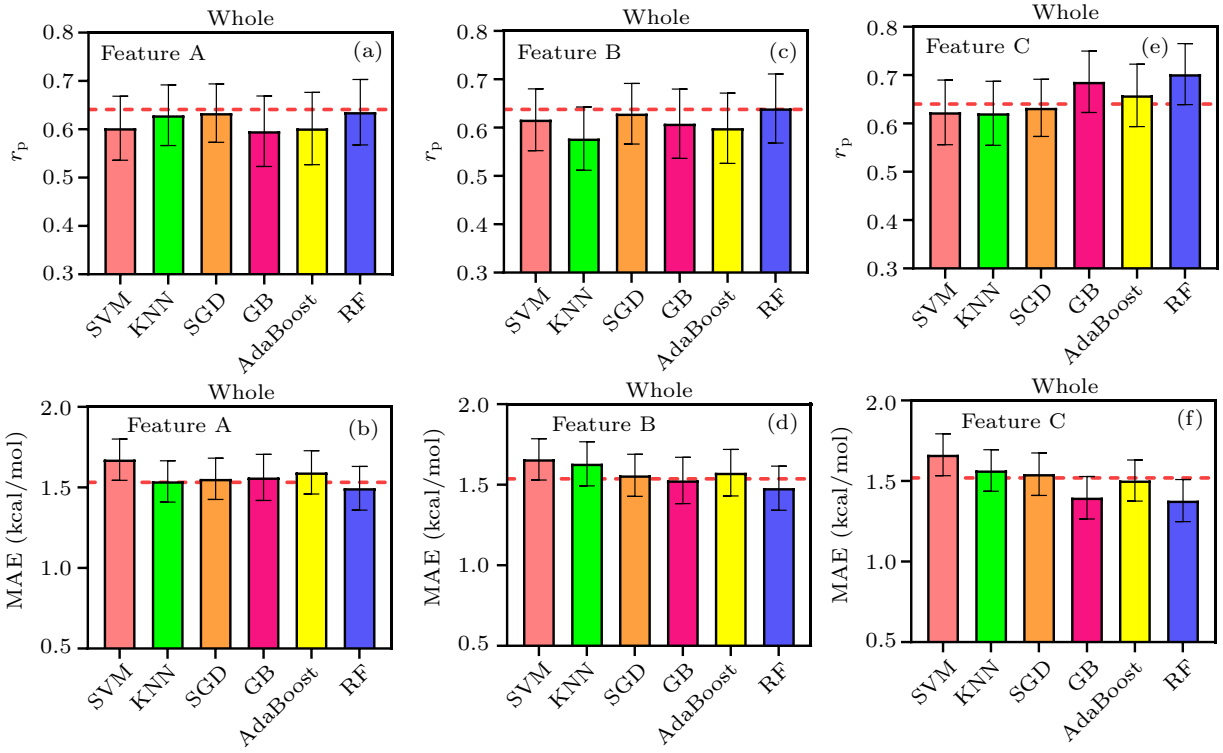



**Fig. 7.** Comparisons of predicted results from different machine learning methods when selecting various features in the whole system. For feature set A, (a) $r_p$ and (b) MAE between predicted values and experimental values are shown; for feature set B, (c) $r_p$ and (d) MAE are presented; for feature set C, (e) $r_p$ and (f) MAE are also provided. The uncertainty estimates for $r_p$ and MAE are the standard deviations from 1000 repeated calculations, with the red dashed lines representing the predicted results from the linear regression algorithm in the whole system.

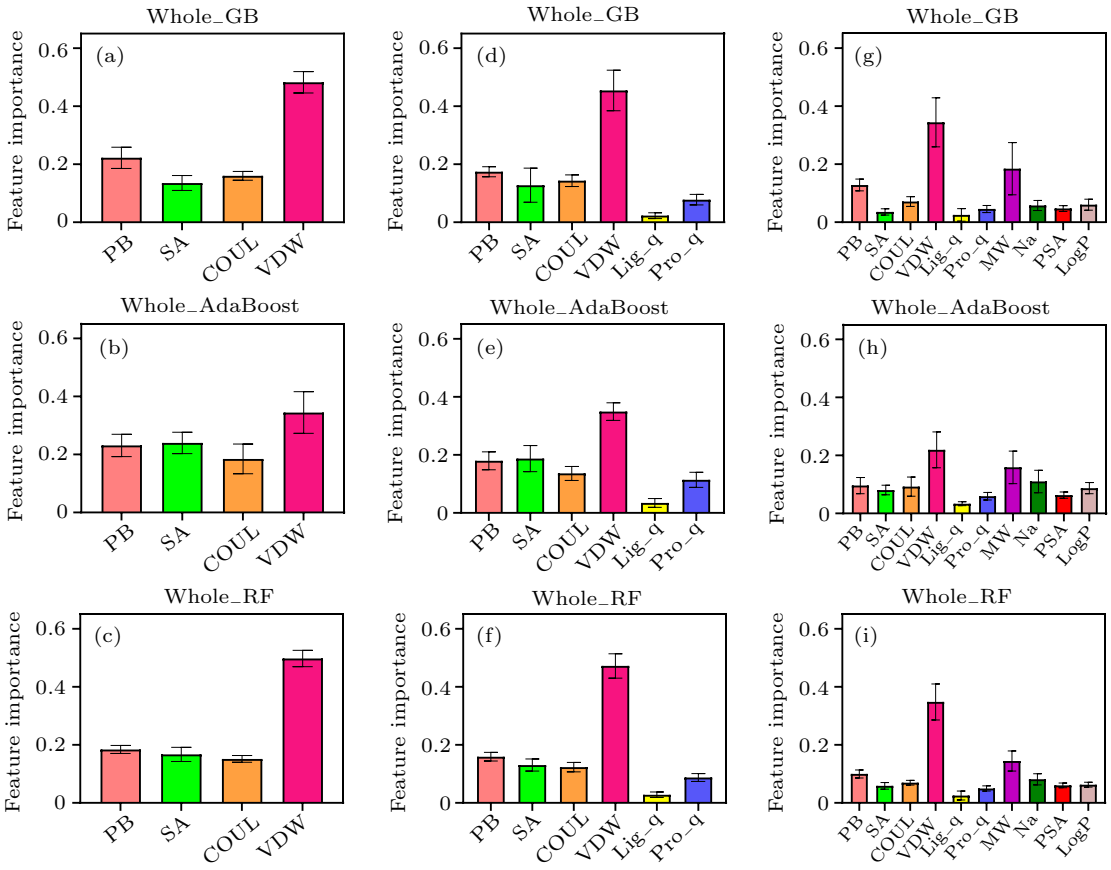



**Fig. 8.** Feature importance analyses for GB, AdaBoost, and RF methods in the whole system under different feature sets. For feature set A, feature importance analyses are shown for (a) GB, (b) AdaBoost, and (c) RF; for feature set B, the analyses are presented for (d) GB, (e) AdaBoost, and (f) RF; for feature set C, the analyses are provided for (g) GB, (h) AdaBoost, and (i) RF.

By analyzing the importance ranking of features in the prediction processes of the GB, AdaBoost, and RF methods, we found that molecular weight (MW) and van der Waals (VDW) interactions played a decisive role in the prediction results. Consequently, we used these features as feature D to reassess the prediction performance of the GB, AdaBoost, and RF methods. As shown in Fig. S2, the predictive performance of these three methods in feature D is significantly lower than that of the corresponding methods in feature C. This suggests that, although other features rank low in importance during the prediction process, they still play a crucial role in machine learning prediction methods. Furthermore, we analyzed the importance ranking within feature D (see Fig. S3 in supporting information). Our findings indicate that the contribution of MW is particularly high in the neutral ligand system, while the contribution of VDW interactions is more pronounced in the charged ligand system.

Given RF's consistently strong performance across different datasets, particularly in feature C for the whole system, we further investigated the impact of key parameters on RF's predictive accuracy. As shown in Fig. 9, increasing the number of trees ($n$) improved the $r_p$ and reduced the MAE until saturation at $n > 100$ (Figs. 9(a) and 9(b)). Similarly, increasing the maximum tree depth ($m$) improved prediction performance, with saturation occurring at $m = 8$ (Figs. 9(c) and 9(d)). Therefore, considering both efficiency and accuracy, we recommend using $n = 200$ and $m = 10$ for RF.

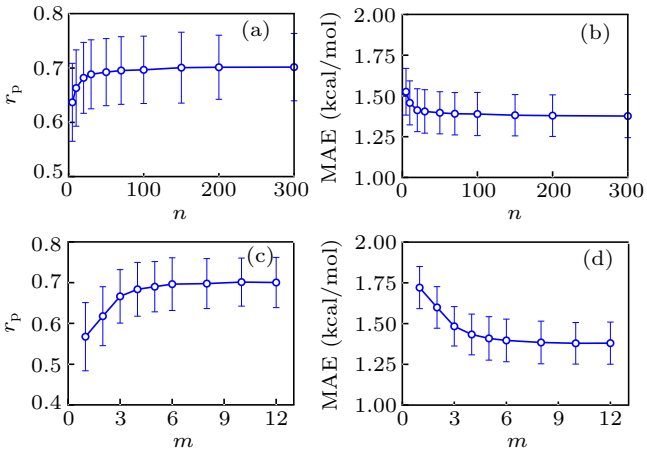



**Fig. 9.** Optimization of parameters for the random forest method. The number of trees ($n$) and its effect on the prediction results: (a) $r_p$ and (b) MAE (with the maximum tree depth fixed at 10). The maximum tree depth ($m$) and its effect on the prediction results: (c) $r_p$ and (d) MAE (with the number of trees fixed at 200). The uncertainty estimates for $r_p$ and MAE are the standard deviations from 1000 repeated calculations.

## 4. Conclusion

In summary, we have employed various machine learning methods, including linear regression (LR), support vector machines (SVM), $k$-nearest neighbors (KNN), stochastic gradient descent (SGD), gradient boosting (GB), adaptive boost-

ing (AdaBoost), and random forest (RF), to predict binding free energies. The predictions were based on features derived from the screening MM/PBSA method, which included the polar and non-polar solvation free energy components, electrostatic interaction energy, van der Waals interaction energy, and other relevant characteristics. Our results demonstrate that machine learning methods outperform the direct screening MM/PBSA calculations in predicting binding free energies. Additionally, we examined the performance of these methods across different systems, including neutral ligands, charged ligands, and the combined dataset. We found that LR, SVM, and SGD performed better for charged ligand systems, while KNN, GB, AdaBoost, and RF yielded better predictions for neutral ligand systems. Furthermore, we analyzed the importance of different features in the prediction process for GB, AdaBoost, and RF, revealing that feature selection significantly influenced their predictive performance, with molecular weight (MW) and van der Waals (VDW) interactions playing decisive roles. Overall, the random forest method exhibited the best predictive performance. When combined with the screening MM/PBSA approach, it substantially improved the accuracy of binding free energy predictions. This indicates that the method is a highly accurate, reliable, and stable tool for predicting binding free energies, making it valuable for drug screening, understanding protein–ligand interactions, and elucidating related biological processes.

Finally, it is important to note that the method proposed here could also be applicable to other biological systems, such as protein–protein and protein–nucleic acid interactions. However, the selection of features beyond those obtained from the screening MM/PBSA method poses a greater challenge compared to protein–ligand systems. The choice of overall properties of proteins or nucleic acids (e.g., charge, molecular weight, sequence length) as features is inadequate to capture the local characteristics of binding regions. Thus, identifying suitable features for protein–protein and protein–nucleic acid systems presents a challenge and is the focus of our future work.

## Data availability statement

The corresponding structure and computational code files are freely available at the website: https://doi.org/10.6084/m9.figshare.27255183.

## Acknowledgement

Project supported by the National Natural Science Foundation of China (Grant Nos. 12222506, 12347102, 12447164, and 12174184).

## References

[1] Persch E, Dumele O and Diederich F 2015 *Angew. Chem. Int. Ed.* **54** 3290
[2] Tosstorff A, Cole J C, Bartelt R and Kuhn B 2021 *ChemMedChem* **16** 3428
[3] Janin J 1995 *Prog. Biophys. Mol. Biol.* **64** 145
[4] Steinbrecher T and Labahn A 2010 *Curr. Med. Chem.* **17** 767
[5] Acuner Ozbabacan S E, Gursoy A, Keskin O and Nussinov R 2010 *Curr. Opin. Drug Discovery Dev.* **13** 527 PMID: 20812144
[6] Du X, Li Y, Xia Y L, Ai S M, Liang J, Sang P, Ji X L and Liu S Q 2016 *Int. J. Mol. Sci.* **17** 144
[7] Bartas M, Slychko K, Brázda V, Červeň J, Beaudoin C A, Blundell T L and Pečinka P 2022 *Int. J. Mol. Sci.* **23** 768
[8] Orphanides G and Reinberg D 2002 *Cell* **108** 439
[9] Chaires J B 2008 *Annu. Rev. Biophys.* **37** 135
[10] Yang X, Zhou P, Shen S, Hu Q, Tian C, Xia A, Wang Y, Yang Z, Nan J and Zhou Y 2024 *Proc. Natl. Acad. Sci. USA* **121** e2401091121
[11] Shivakumar D, Williams J, Wu Y, Damm W, Shelley J and Sherman W 2010 *J. Chem. Theory Comput* **6** 1509
[12] Schreiber G 2002 *Curr. Opin. Struct. Biol.* **12** 41
[13] Hu X and Contini A 2019 *J. Chem. Inf. Model.* **59** 2714
[14] Fu T, Wu X, Xiu Z, Wang J, Yin L and Li G 2013 *J. Theor. Comput. Chem.* **12** 1341003
[15] Wang E, Sun H, Wang J, Wang Z, Liu H, Zhang J Z H and Hou T 2019 *Chem. Rev.* **119** 9478
[16] Wang C, Greene D A, Xiao L, Qi R and Luo R 2018 *Front. Mol. Biosci.* **4** 87
[17] Abel R, Wang L, Harder E D, Berne B J and Friesner R A 2017 *Acc. Chem. Res.* **50** 1625
[18] Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren M K, Greenwood J, Romero D L, Masse C, Knight J L, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, Murcko M, Frye L, Farid R, Lin T, Mobley D L, Jorgensen W L, Berne B J, Friesner R A and Abel R 2015 *J. Am. Chem. Soc.* **137** 2695
[19] Homeyer N and Gohlke H 2012 *Mol. Inf.* **31** 114
[20] Wang E, Weng G, Sun H, Du H, Zhu F, Chen F, Wang Z and Hou T 2019 *Phys. Chem. Chem. Phys.* **21** 18958
[21] Sheng Y J, Yin Y W, Ma Y Q and Ding H M 2021 *J. Chem. Inf. Model.* **61** 2454
[22] Ding H M, Yin Y W, Sheng Y J and Ma Y Q 2021 *Chin. Phys. Lett.* **38** 018701
[23] Zhu Y X, Sheng Y J, Ma Y Q and Ding H M 2022 *J. Phys. Chem. B* **126** 1700
[24] Chen Y Q, Sheng Y J, Ding H M and Ma Y Q 2022 *Chin. Phys. B* **31** 048701
[25] Yan Z S, Xu Y, Ding H M and Ma Y Q 2022 *Chin. Phys. Lett.* **39** 108701
[26] Zhang C 2012 *Ensemble Machine Learning: Methods and Applications* (New York: Springer) pp. 167–184
[27] Chen X, Wang K, Chen J, Wu C, Mao J, Song Y, Liu Y, Shao Z and Pu X 2024 *Nat. Commun.* **15** 8130
[28] Zhang Y, Li S, Meng K and Sun S 2024 *J. Chem. Inf. Model.* **64** 1456
[29] Li J, Fu A and Zhang L 2019 *Interdiscip. Sci.: Comput. Life Sci.* **11** 320
[30] Ballester P J and Mitchell J B 2010 *Bioinformatics* **26** 1169
[31] Sato T, Honma T and Yokoyama S 2010 *J. Chem. Inf. Model.* **50** 170
[32] Deng W, Breneman C and Embrechts M J 2004 *J. Chem. Inf. Comput. Sci.* **44** 699
[33] Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K and Tylkowski B 2022 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **12** e1568
[34] Ellingson S R, Davis B and Allen J 2020 *Biochim. Biophys. Acta, Gen. Subj.* **1864** 129545
[35] S Heck G, O Pintro V, R Pereira R, Mb Levin N and F De Azevedo W 2017 *Curr. Med. Chem.* **24** 2459
[36] Colwell L J 2018 *Curr. Opin. Struct. Biol.* **49** 123
[37] Bitencourt-Ferreira G and De Azevedo W F 2018 *Biophys. Chem.* **240** 63
[38] Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y and Wang R 2018 *J. Chem. Inf. Model.* **59** 895
[39] Rose P W, Beran B, Bi C, Bluhm W F, Dimitropoulos D, Goodsell D S, Prlić A, Quesada M, Quinn G B and Westbrook J D 2010 *Nucleic Acids Res.* **39** D392
[40] Burley S K, Berman H M, Christie C, Duarte J M, Feng Z, Westbrook J, Young J and Zardecki C 2018 *Protein Sci.* **27** 316

[41] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B and Lindahl E 2015 *SoftwareX* **1** 19

[42] Berendsen H J, Van Der Spoel D and Van Drunen R 1995 *Comput. Phys. Commun.* **91** 43

[43] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark A E and Berendsen H J C 2005 *J. Comput. Chem.* **26** 1701

[44] Case D A, Aktulga H M, Belfon K, Ben-Shalom I, Brozell S R, Cerutti D S, Cheatham Iii T E, Cruzeiro V W D, Darden T A and Duke R E 2020 *Amber 2020* (University of California, San Francisco)

[45] Jakalian A, Bush B L, Jack D B and Bayly C I 2000 *J. Comput. Chem.* **21** 132

[46] Jakalian A, Jack D B and Bayly C I 2002 *J. Comput. Chem.* **23** 1623

[47] Wang J, Wolf R M, Caldwell J W, Kollman P A and Case D A 2004 *J. Comput. Chem.* **25** 1157

[48] Maier J A, Martinez C, Kasavajhala K, Wickstrom L, Hauser K E and Simmerling C 2015 *J. Chem. Theory Comput.* **11** 3696

[49] Jorgensen W L and Madura J D 1983 *J. Am. Chem. Soc.* **105** 1407

[50] Hess B, Bekker H, Berendsen H J and Fraaije J G 1997 *J. Comput. Chem.* **18** 1463

[51] Zhou R, Harder E, Xu H and Berne B 2001 *J. Chem. Phys.* **115** 2348

[52] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G 1995 *J. Chem. Phys.* **103** 8577

[53] Sun H, Duan L, Chen F, Liu H, Wang Z, Pan P, Zhu F, Zhang J Z and Hou T 2018 *Phys. Chem. Chem. Phys.* **20** 14450

[54] Genheden S and Ryde U 2015 *Expert Opin. Drug Discovery* **10** 449

[55] Rastelli G, Rio A D, Degliesposti G and Sgobba M 2010 *J. Comput. Chem.* **31** 797

[56] Sun H, Li Y, Tian S, Xu L and Hou T 2014 *Phys. Chem. Chem. Phys.* **16** 16719

[57] Chen F, Sun H, Wang J, Zhu F, Liu H, Wang Z, Lei T, Li Y and Hou T 2018 *Rna* **24** 1183

[58] https://jerkwin.github.io/gmxtool

[59] Chen Y Q, Sheng Y J, Ma Y Q and Ding H M 2022 *Phys. Chem. Chem. Phys.* **24** 14339

[60] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B and Varoquaux G 2014 *Front. Neuroinform.* **8** 71792

[61] Genheden S and Ryde U 2011 *J. Chem. Theory Comput.* **7** 3768

[62] He X, Man V H, Ji B, Xie X Q and Wang J 2019 *J. Comput.-Aided Mol. Des.* **33** 105