

DockEM: an enhanced method for atomic-scale protein–ligand docking refinement leveraging low-to-medium resolution cryo-EM density maps

Jing Zou^{1,†}, Wenyi Zhang^{2,†}, Jun Hu^{3,*}, Xiaogen Zhou^{1,*}, Biao Zhang^{4,*,†}

¹College of Information Engineering, Zhejiang University of Technology, 288 Liuhe Road, Liuxia Street, Xihu District, Hangzhou 310023, China

²Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310024, China

³Chinese Academy of Medical Sciences Suzhou Institute of Systems Medicine, Suzhou 215123, China

⁴College of Information Engineering, Zhejiang University of Technology and Chinese Academy of Medical Sciences, Suzhou Institute of Systems Medicine, Suzhou 215123, China

*Corresponding authors. Biao Zhang, College of Information Engineering, Zhejiang University of Technology and Chinese Academy of Medical Sciences, Suzhou Institute of Systems Medicine, 100 Chongwen Road, Suzhou 215123, China. E-mail: zhangb@ism.cams.cn; Xiaogen Zhou, College of Information Engineering, Zhejiang University of Technology, 288 Liuhe Road, Liuxia Street, Xihu District, Hangzhou 310023, China. Tel:15257118077; E-mail: zxcg@zjut.edu.cn; Jun Hu, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 100 Chongwen Road, Suzhou 215123, China. E-mail: hj@ism.cams.cn

†Jing Zou, Wenyi Zhang, and Biao Zhang contributed equally to this work.

Abstract

Protein–ligand docking plays a pivotal role in virtual drug screening, and recent advancements in cryo-electron microscopy (cryo-EM) technology have significantly accelerated the progress of structure-based drug discovery. However, the majority of cryo-EM density maps are of medium to low resolution (3–10 Å), which presents challenges in effectively integrating cryo-EM data into molecular docking workflows. In this study, we present an updated protein–ligand docking method, DockEM, which leverages local cryo-EM density maps and physical energy refinement to precisely dock ligands into specific protein binding sites. Tested on a dataset of 121 protein–ligand compound, our results demonstrate that DockEM outperforms other advanced docking methods. The strength of DockEM lies in its ability to incorporate cryo-EM density map information, effectively leveraging the structural information of ligands embedded within these maps. This advancement enhances the use of cryo-EM density maps in virtual drug screening, offering a more reliable framework for drug discovery.

Keywords: docking; cryo-EM; protein–ligand; REMC simulation; refinement

Introduction

Bioinformatics is a crucial field to study big data [1–4]. It plays an important role to deciphering high-resolution protein structures, drug development, and protein targeting which involved in progression of different diseases, including various types of cancers [5–7]. Many vital biological functions within organisms depend on the interactions between proteins and other intracellular molecules [8–10]. Molecular docking is a theoretical method used to study the interactions between proteins and ligands. One of the key applications of molecular docking is in the development of innovative drugs. However, new drug development is a costly and inefficient process, even with modern automated high-throughput screening technologies, developing a typical small-molecule drug can take several years and cost millions of dollars [11]. Molecular docking can rapidly and efficiently screen millions of compound molecules computationally [12]. Enhancing docking accuracy is crucial for the success of virtual drug screening and its application within the drug discovery pipeline [13].

Advances in cryo-electron microscopy (cryo-EM) technology have significantly increased its relevance in structure-based drug discovery [14]. Improvements in cryo-EM resolution have

enabled the exploration of more potential drug targets, including membrane proteins, within structure-based drug discovery, expanding our understanding of drug target identification, design, and development, particularly in areas where conventional techniques, such as X-ray crystallography, face limitations [15, 16]. While cryo-EM density maps can achieve atomic-level resolution [17], the majority of maps stored in the Electron Microscopy Data Bank (EMDB) [18] are still deposited at medium to low resolutions (above 3 Å). Therefore, integrating effective medium to low-resolution cryo-EM density map data into the molecular docking process remains a pressing challenge that needs to be addressed.

Given the significance of docking, several methods have been specifically developed for protein–peptide docking, such as MOTR [19] and DGMOEA [20]. MOTR is an innovative multi-objective metaheuristic algorithm designed for the protein–peptide docking problem, while DGMOEA is a dual-population multi-objective evolutionary algorithm. Our work focuses primarily on the docking problem between proteins and small molecules, which differ substantially from peptides in terms of flexibility, size, and binding mechanisms. Currently, numerous molecular docking computational methods have been developed for small molecules, including ChemEM [14], AutoDock Vina [21], DOCK6 [22], EDock [23],

Received: November 15, 2024. Revised: January 26, 2025. Accepted: February 18, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

CB-Dock2 [24], and EMERALD [25]. However, traditional molecular docking methods, such as AutoDock Vina and EDock, have certain fundamental limitations. These methods rely on conformation search strategies, guided by various composite energy functions, to quantify the interactions between ligand-protein compound during the molecular docking process. Designing energy functions that can accurately and efficiently describe these interactions is a daunting task. CB-Dock2, an improved version of CB-Dock [26], introduces template-based search, allowing for the specification of protein binding sites and utilizing a new curvature-based cavity detection method calculating the center and size of the cavity before docking with the popular AutoDock Vina program. ChemEM proposes a new ChemDock scoring function and utilizes differential density maps, which extract the ligand portion from the entire density map. EMERALD leverages the docking capabilities of RosettaGenFF and GA optimization, while simultaneously integrating cryo-EM density data with the physically realistic force field of RosettaGenFF. However, both methods have a limitation: the resolution of the density maps is typically low, and the resolution near the binding site may be compromised due to ligand-induced perturbations. As a result, the differential density maps and pseudo-atomic skeleton (derived from the unmodeled density map around the ligand) obtained under these conditions may lack some essential information. Despite the successes of these docking methods, achieving a balance between accuracy and efficiency in current molecular docking methods remains a challenge under certain conditions. Recently, methods specifically designed for fitting small molecules into cryo-EM density maps have emerged [25, 27]. The success of these processes depends on the initial fitting of the ligand, as positioning the ligand close to the correct site, combined with structural information from the density map, increases the chances of generating high-quality docking conformations.

In this work, we introduce DockEM, an updated molecular docking method designed for high-quality ligand position. DockEM achieves high-precision docking results using medium to low-resolution density maps. Starting from the three-dimensional structure of the protein, such as the receptor protein model generated by the structure prediction algorithm AlphaFold2 (AF2) [28] used in this study, DockEM identifies the approximate position of the ligand using binding site information through a local structure comparison method and extracts an initial local density map, as illustrated in Fig. 1A–C. The ligand structure is subsequently fitted to the initial local density map, with continuous optimization of the ligand's position and the center position of the local ligand density map, as shown in Fig. 1D. A replica exchange Monte Carlo (REMC) [29] simulation, guided by our designed energy function that incorporates cryo-EM density map data, is employed to perform rigid-body alignment of the ligand with the local density map as well as flexible docking of the ligand to the protein, as illustrated in Fig. 1E and F. Extensive conformational searches and structural refinements are performed throughout the docking process. To rigorously assess the strengths and weaknesses of DockEM, we performed benchmark tests on 121 protein–ligand targets collected from DUDE [30] and COACH [31]. The standalone DockEM predictor used in this study is freely available at <https://github.com/Hsfssfw/DockEM>.

Materials and methods

Dataset

The dataset used in our benchmark testing consists of 121 targets derived from the DUDE and COACH datasets referenced in

EDock. After acquiring the protein–ligand datasets, we generated the corresponding simulated cryo-EM density map data using the EMAN2 [32] and UCSF Chimera [33] software packages. The resolutions of all simulated cryo-EM density maps ranged from 3 to 15 Å. The resolution range was chosen as it reflects typical medium to low-resolution scenarios [34] even though the average resolution of cryo-EM density maps in the community has been steadily improving over the past few years [35]. Details regarding the resolutions of all simulated density maps can be found in [Supplementary File S1](#).

Energy field

The force field in DockEM comprises a linear combination of the following four energy terms: the density map energy term, the ligand-protein intramolecular energy contributions from van der Waals interactions and Coulombic electrostatic interactions, the ligand's internal van der Waals interactions, and the distance constraints to the local density map.

$$E_{\text{total}} = w_1 E_{\text{CC}} + w_2 E_{\text{intra}} + w_3 E_{\text{inter}} + w_4 E_{\text{dis}} \quad (1)$$

Here, w_1 , w_2 , w_3 and w_4 represents the weight of different energy terms, where $E_{\text{CC}} = 1 - \text{CC}$. The term CC refers to the correlation coefficient (CC) between the ligand structure and the local density map. The formula for calculating CC is as follows:

$$\text{CC} = \frac{\sum_{g=1}^{N_{\text{vol}}} [\rho_c(g) - \bar{\rho}_c] [\rho_e(g) - \bar{\rho}_e]}{\sqrt{\sum_{g=1}^N [\rho_c(g) - \bar{\rho}_c]^2} \sqrt{\sum_{g=1}^N [\rho_e(g) - \bar{\rho}_e]^2}} \quad (2)$$

Here, $\bar{\rho}_c$ represents the average density value of the map inferred from the ligand structure, while $\bar{\rho}_e$ denotes the average density value of the experimental density map. N_{vol} is the total number of grid points in the local density map. The term $\rho_e(g)$ refers to the density value at the g th grid point in the experimental density map, and $\rho_c(g)$ is the density value at the g th grid point in the density map inferred from the ligand structure. The value of $\rho_c(g)$ can be calculated as follows:

$$\rho_c(g) = \sum_{i=1}^{N_a} m_i^j \sqrt{\left(\frac{\pi^2}{(2.4 + 0.8R)^2}\right)^3} \exp\left(-\left(\frac{\pi}{2.4 + 0.8R}\right)^2 |x_i - x_g|^2\right) \quad (3)$$

Here, N_a represents the total number of atoms in the ligand structure, m is the atomic mass, R denotes the resolution of the experimental density map, x_i refers to the grid coordinates of the i th atom, and x_g indicates the coordinates of the g th grid point.

The second term in Equation (1) accounts for the van der Waals and Coulombic electrostatic interactions between the ligand and the receptor:

$$E_{\text{intra}} = \sum_{i \in \text{protein}} \sum_{j \in \text{ligand}} \left[w_{21} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^6} \right) + w_{22} \frac{q_i q_j}{4d_{ij}} \right] \quad (4)$$

Here, w_{21} and w_{22} are the weights of this energy term, and d_{ij} represents the distance between the i th atom in the protein and the j th atom in the ligand. Here, the protein atoms refer to those within the protein whose distance to the density map is less than half of the maximum distance between any two atoms within the ligand. $A_{ij} = \epsilon R^{12}$ and $B_{ij} = 2\epsilon R^6$ are the repulsion and attraction parameters, respectively, where $R = r_i + r_j$ and $\epsilon = \sqrt{\epsilon_i \epsilon_j}$ are associated with the van der Waals radius and well depth, with values taken from the AMBER99 force field [36]. For

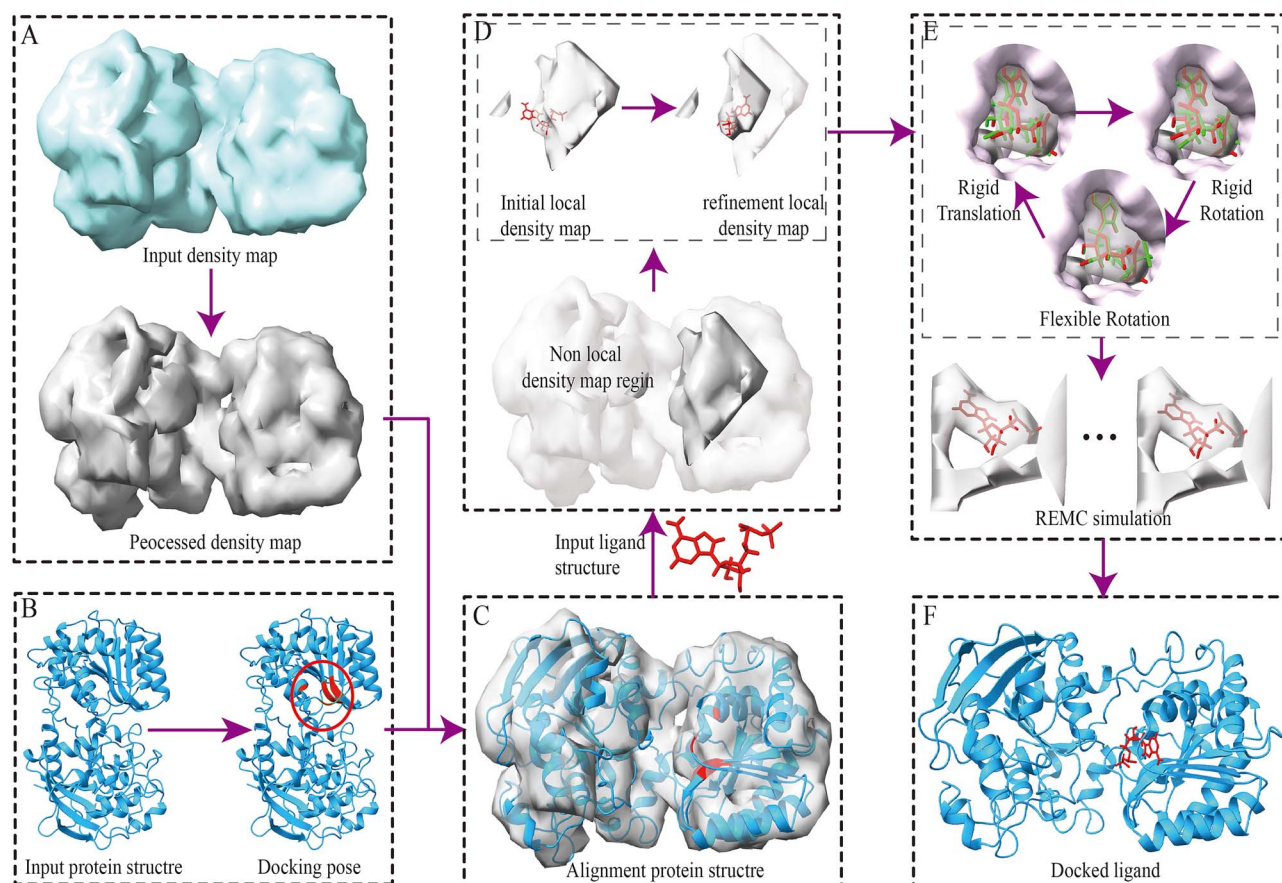


Figure 1. Workflow of protein–ligand docking using DockEM. (A) Input cryo-EM density map and processing. (B) Input structures of the protein and ligand, along with binding site information (or docking position) and cryo-EM density map. (C) Superpose the input protein structure onto the processed density map. (D) Use binding site information to locate the initial region of the local density map and update the center position of this initial local density map. (E) Perform conformational sampling through REMC simulation. (F) Select the final docked ligand structure.

the Coulomb interaction term, q_i and q_j are the charges of the atoms, determined using the Antechamber module of AMBER [37].

The third term in Equation (1) represents the van der Waals interactions within the ligand:

$$E_{\text{inter}} = \sum_{i,j \in \text{ligand}, i \neq j} \left(\frac{A_{ij}}{s_{ij}^{12}} - \frac{B_{ij}}{s_{ij}^6} \right) \quad (5)$$

Here, s_{ij} is the distance between the i th atom and the j th atom within the ligand. In the second energy term, we calculated the van der Waals forces between the protein and the ligand, which effectively prevent conflicts between the protein and ligand atoms. Here, we calculate the van der Waals forces within the ligand to prevent internal conflicts between the ligand's atoms.

The fourth term in Equation (1) represents the distance constraints to the local density map:

$$E_{\text{dis}} = \begin{cases} 0, & |c_i - c_0| < c_{\text{lim}} \\ |c_i - c_0|, & |c_i - c_0| \geq c_{\text{lim}} \end{cases} \quad (6)$$

Here, c_i is the center position of the conformation, c_0 is the center position of the local density map, and c_{lim} is half of the maximum distance between any two atoms in the ligand. The purpose of setting this energy term is to reduce the confirmation search space of the ligand, allowing it to move around the

ligand-specific region of the local density map to find a local optimum.

Rigid-body alignment and local density map positioning

The protein is aligned with the global density map using REMC simulations guided by the E_{CC} energy term (see *Energy field* section). Here, E_{CC} represents the CC calculated between the protein structure and the cryo-EM density map. Starting from the ligand–protein binding site (or docking position), we define the region of the local ligand density map. The center of the binding site serves as the initial center of the local density map. If a docking position is provided, it is used as the center of the initial local density map. The local density map represented as a cubic region within the entire density map with an edge length twice the maximum distance between any two atoms in the ligand. The ligand is aligned to the binding site center and fitted to the local density map by maximizing the CC between the ligand and this map, while continuously optimizing the map's center position. This fitting is achieved over 500 steps of Monte Carlo (MC) [38] simulation, guided by the energy function $w_1 E_{\text{CC}} + w_2 E_{\text{intra}}$ (see Equation (1)). After the simulation, the lowest-energy ligand conformation is selected, and the center of this ligand structure is set as the new center of the local density map. Based on this updated center, a cubic region with an edge length equal to the maximum ligand atom-to-atom distance is generated. The segment of the density

map within this region becomes the local density map for further docking refinement.

Ligand sampling by REMC

DockEM implements the REMC protocol, which enhances sampling efficiency and expands the sampling space. In this step, the REMC simulation includes N replicas, which are sampled in parallel. The temperature of the i th replica is given by:

$$T_i = T_{\min} \left(\frac{T_{\max}}{T_{\min}} \right)^{\frac{i-1}{N-1}} \quad (7)$$

Here, $T_{\min} = 0.01$ represents the minimum temperature, and $T_{\max} = 1$ represents the maximum temperature. Every local MC step moves, a global exchange move is attempted between two neighboring replicas i and j , with the acceptance probability given by:

$$P_{\text{global}} = \min \left\{ 1, \exp \left((E_j - E_i) \left(\frac{1}{KT_j} - \frac{1}{KT_i} \right) \right) \right\} \quad (8)$$

Here, E_j and E_i represent the energies of the j th and i th replicas, respectively, while T_j and T_i are the temperatures of the j th and i th replicas, respectively. K is a parameter. This movement between neighboring replicas allows the conformation of a low-temperature replica to be exchanged with that of a high-temperature replica, thereby helping the low-temperature simulation escape from local energy minima.

Refinement by flexible ligand-protein docking

Before performing flexible docking, we conducted an initial rigid protein-ligand docking, utilizing REMC simulations with 20 replicas. Each replica starts by positioning the ligand to the center of the local density map, with its initial state generated through random large-scale rotations and small-scale translations. Every 500 local MC moves, a global exchange move is attempted between two neighboring replicas. Following the REMC-based rigid protein-ligand docking, we selected the lowest-energy conformation from each of the 20 replicas to proceed flexible docking for further docking refinement. The flexible docking process is also implemented through REMC simulation, guided by the total energy E_{total} (see Equation (1)). For flexible movements, single rotatable bonds are identified based on the "BOND" table in the MOL2 file. After rigid-body translation and rotation of the entire molecule, a random rotatable bond is chosen as the axis of rotation. All atoms distal to this bond are rotated by a random angle within the range of $[-180^\circ, 180^\circ]$. No translation is applied to these atoms during the process, as translating only a portion of the atoms could result in unnatural bond lengths and angles.

Docking experiment evaluation indicators

Our docking performance is primarily assessed by the root mean square deviation (RMSD) of the docked ligand conformation relative to the native structure. Accurate RMSD calculation requires that the atom order in both the predicted and experimental ligand structures be consistent. However, we found cases where the atom order of some input ligand files differed from that of the experimental ligand files. This mismatch can result in discrepancies between the docked and experimental ligands, potentially leading to incorrect RMSD values. To address this issue, we employed DockRMSD [39] to standardize the atom order between the docked and experimental ligand structures before computing the RMSD.

Results and discussion

The overall docking results

To validate the DockEM method, comparisons were made with ChemEM and EMERALD, which similarly constrains ligand-protein docking using cryo-EM density; EDock, designed specifically for predicted models; and CB-Dock2, which employs templates to guide the docking pose. For ChemEM, the centroid of the binding site was specified as the location for ligand fitting, utilizing the center of the local density map identified by our method as the input centroid. For EMERALD, the ligand structure obtained after rigid docking by our method is used as the input ligand structure for EMERALD, while the protein structure fitted with the density map obtained from our method is used as the input protein structure for EMERALD. For EDock and CB-Dock2, additional binding site information was provided to ensure a fair comparison. In EDock, graph-based predicted binding site information was replaced with manually input data. CB-Dock2, an online server, allows users to input binding site information upon submission; if no information is provided, it defaults to blind docking. These steps ensured that the docking outcomes were based on consistent input data, allowing for an unbiased comparison among the five methods.

The protein structures in our dataset were predicted using AF2, encompassing 121 targets with an average TM-score of 0.983, as detailed in [Supplementary File S2](#). These results indicate that AF2 can reliably produce correctly folded models for most protein receptors. It also demonstrates that for a given binding site (or docking position), the docking center calculated from the predicted protein structure is close to the true center, which facilitates subsequent iterative optimization and docking site searches. [Table 1](#) summarizes the docking results. Overall, DockEM outperformed EMERALD, ChemEM, CB-Dock2, and EDock, achieving an average RMSD of 1.87 Å across the dataset. This represents a 10.1% reduction compared to EMERALD (2.06 Å), a 50.1% reduction compared to ChemEM (3.75 Å), a 35.1% reduction compared to CB-Dock2 (2.88 Å), and a 53.5% reduction compared to EDock (4.02 Å). Paired Student's t -tests revealed statistically significant differences in ligand structures predicted by DockEM compared to those from EMERALD, ChemEM, CB-Dock2, and EDock, with P -values of $1.21\text{E-}2$ for EMERALD, $1.24\text{E-}6$ for CB-Dock2, $1.95\text{E-}23$ for ChemEM, and $2.27\text{E-}23$ for EDock.

To provide a more comprehensive comparison of the five docking methods, we used RDKit [40] to calculate the electrostatic (ELC), van der Waals (vdW), hydrogen bond (HB), and solvation (SLV) energies for the structures generated by each method. These energy-based metrics were used to further assess the protein-ligand interactions captured by the docking poses. The analysis highlights the comparative strengths of our method against four other methods (EDock, ChemEM, EMERALD, and CB-Dock2). As shown in [Fig. 2A](#), in the ELC energy term, our method (-0.61) consistently ranked first, with only a marginal difference compared to EDock (-0.52), outperforming EMERALD (-0.47), CB-Dock2 (-0.14) and ChemEM (-0.21). For the vdW and SLV energy terms, the differences among the five methods were relatively small; however, our method ranked second in both cases, as shown in [Fig. 2B and C](#). Regarding the HB energy term, RDKit counts the number of HBs formed between the protein and ligand. The HB count cannot be directly used to assess the quality of the HB energy. While a higher count generally indicates stronger and more stable interactions between the ligand and protein, an excessively high count may also suggest potential clashes between them. This is well illustrated in EDock, as shown in

Table 1. Summary of whole-protein docking results with known ligand binding sites

Method	DockEM	EMERALD	CB-Dock2	ChemEM	EDock
RMSD(Å)	1.87	2.06	2.88	3.75	3.99
Center distance(Å)	0.94	1.09	1.85	1.71	2.69
RMSD < 3 Å	110	95	71	40	45
Success rat	90.9%	78.5%	58.7%	33.1%	37.2%
P-value	–	1.21E–2	1.24E–6	1.95E–23	2.27E–23

The RMSD and Center Distance values represent the averages for all docked ligands in the dataset. Center distance represents the distance between the ligand center obtained using different methods and the native ligand center. The success rate is defined as the proportion of docked ligand structures exhibiting an RMSD of <3 Å in comparison to the native ligand structure.

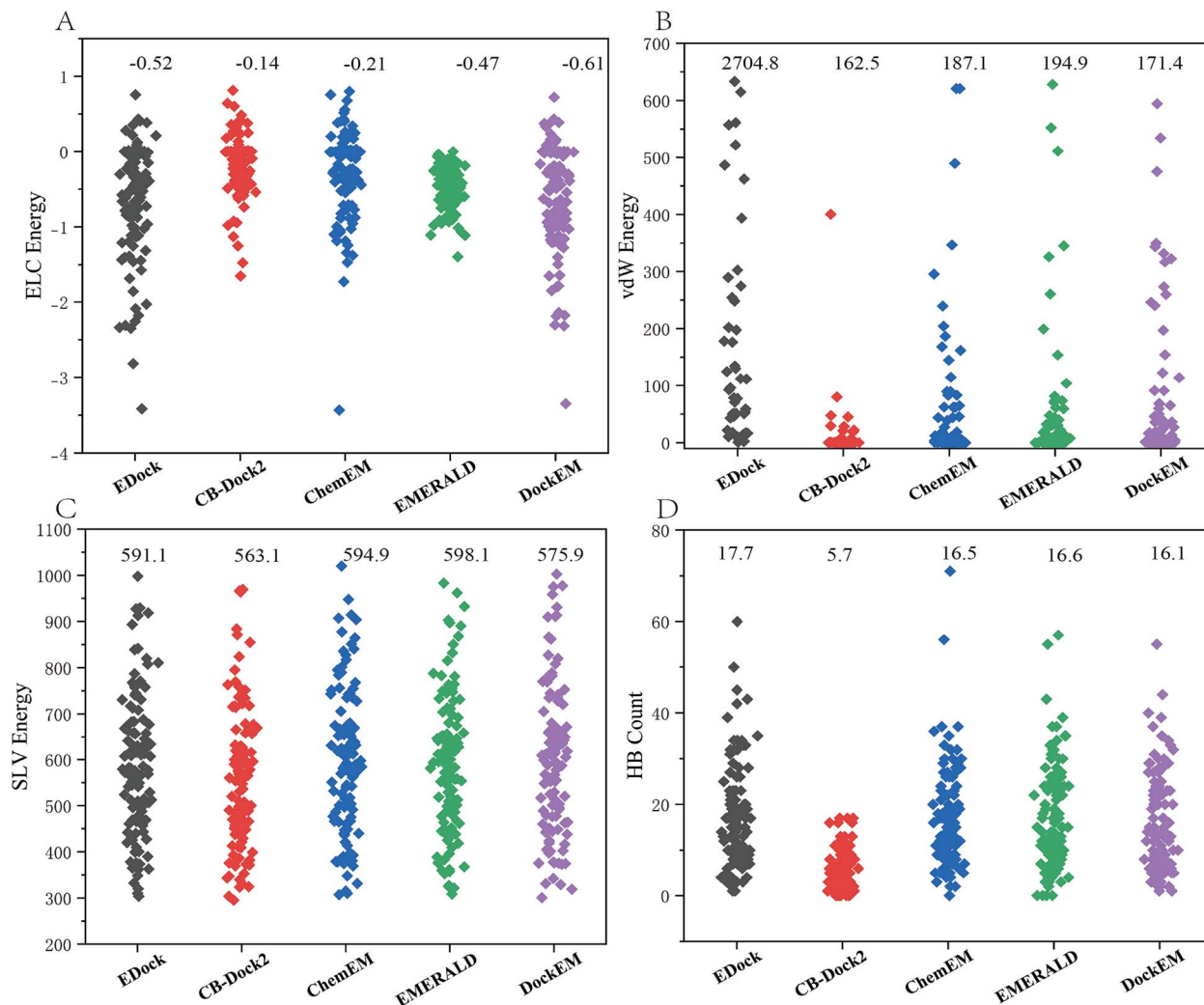


Figure 2. Comparison of five docking methods in four types of energies. All four energies are calculated between the protein and ligand structures. (A) Comparison of five docking methods in the electrostatic (ELC) energy term. (B) Comparison of five docking methods in the van der Waals (vdW) energy term. (C) Comparison of five docking methods in the solvation (SLV) energy term. (D) Comparison of five docking methods in the HB count. The numbers labeled in the figure represent the average value of each energy term.

Fig. 2D, where the HB count of 17.7 is the highest among the five methods and EDock also exhibits the largest vdW energy of 2704.8. In contrast, CB-Dock2 has the smallest HB count of 5.7, with the lowest vdW energy of 162.5. For EMERALD, the HB count is 16.6 and the vdW energy is 194.9. Our method, with an HB count of 16.1 and a vdW energy of 171.4, demonstrates a balance between the HB and vdW energy terms. These energy results indicate that our method is competitive with the other approaches, as it not only produces docking poses with lower RMSD but also achieves relatively favorable energy values across the four physical energy terms. This dual advantage underscores the effectiveness and

robustness of our method in generating high-quality protein–ligand docking models.

Performance of alignment based on local ligand density map

In molecular docking, accurately positioning the ligand within the docking site is crucial for obtaining high-quality ligand structures in subsequent docking steps. DockEM extracts the local ligand density map after aligning the protein to the global density map using rigid-body superposition. This approach leverages the local density map and rigid-body alignment to precisely determine

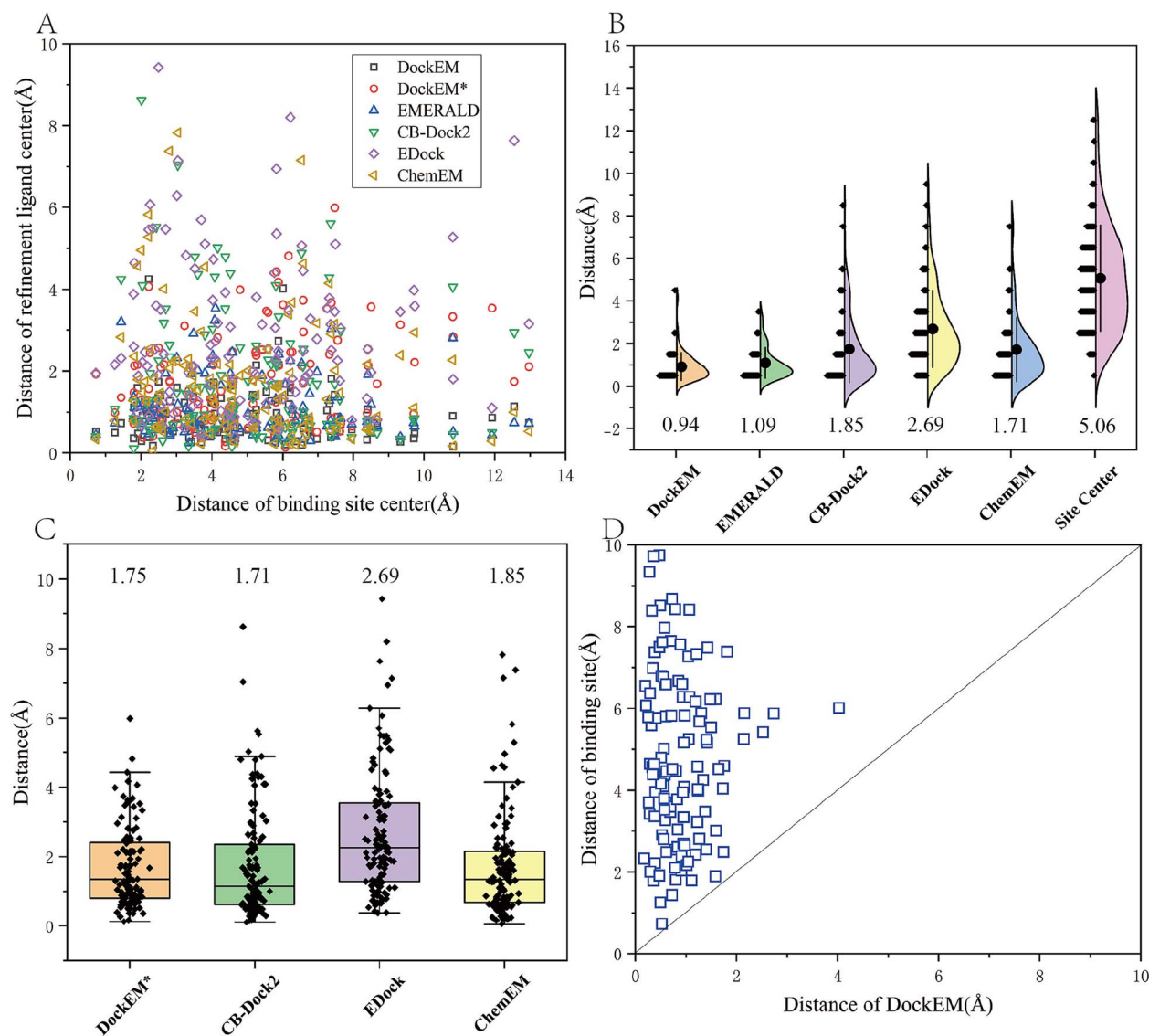


Figure 3. The statistic result of ligand and binding site in different methods. Here, DockEM and DockEM* refers to finally refinement ligand structure (flexible result) and the lowest-energy ligand structure (rigid docking result), respectively, in the DockEM method. (A) The relationship overview of distance of binding site center and distance of refinement ligand center for all methods. The x-axis represents distance of binding site center with native ligand center. The y-axis represents distance of refinement ligand center with native ligand center. All methods represented by different markers. (B) The statistic result of the distance between the refinement ligand center (or binding site center) and native ligand center for DockEM, EMERALD, CB-Dock2, EDock, ChemEM, and binding site. (C) The scatter plot of the distance between the ligand center of rigid docking and native ligand center obtained by the four methods: DockEM*, CB-Dock2, ChemEM, and EDock. (D) Comparison of the distances between the ligand center obtained by DockEM and the center of the binding site. The x-axis represents the distance between the DockEM-docked ligand center and the native ligand center; the y-axis represents the distance between the binding site center and the native ligand center.

the docking position of the ligand on the protein. To assess the effectiveness of refining the ligand docking position using the local density map, we compared the distance of predicted binding site center and the distance of aligned ligand center with the distance of the ligand's native structure center, as shown in the Fig. 3A and B.

The average distance between the ligand's center after rigid docking and the native structure's center was 1.75 Å, reflecting a 65.4% reduction (from an initial 5.06 Å) when compared to the average distance between the predicted binding site's center and the native structure's center. This significant discrepancy between the distance of predicted binding site center and the distance of native structure's center can be attributed to limitations in binding site prediction accuracy and variations in ligand size

and shape. When binding site predictions are accurate, smaller ligand molecules generally have their native structure centers closer to the predicted site center. However, even with accurate predictions, larger ligands may still have centers that deviate from the binding site center due to their size and complexity. We also compared the distance of ligand structures center which obtained using three other methods—ChemEM, CB-Dock2, and EDock—with the distance of native structure's center. As shown in Table 1 and Fig. 3C, the average distance reductions were modest at -2.33% (ChemEM: 1.71 Å), 5% (CB-Dock2: 1.85 Å), and 34.9% (EDock: 2.69 Å). The reason EMERALD is not compared here is that the ligand structure after rigid docking was directly input into EMERALD. As shown in Fig. 3D and Fig. 4A, after completing flexible docking, our method demonstrated further performance improvements

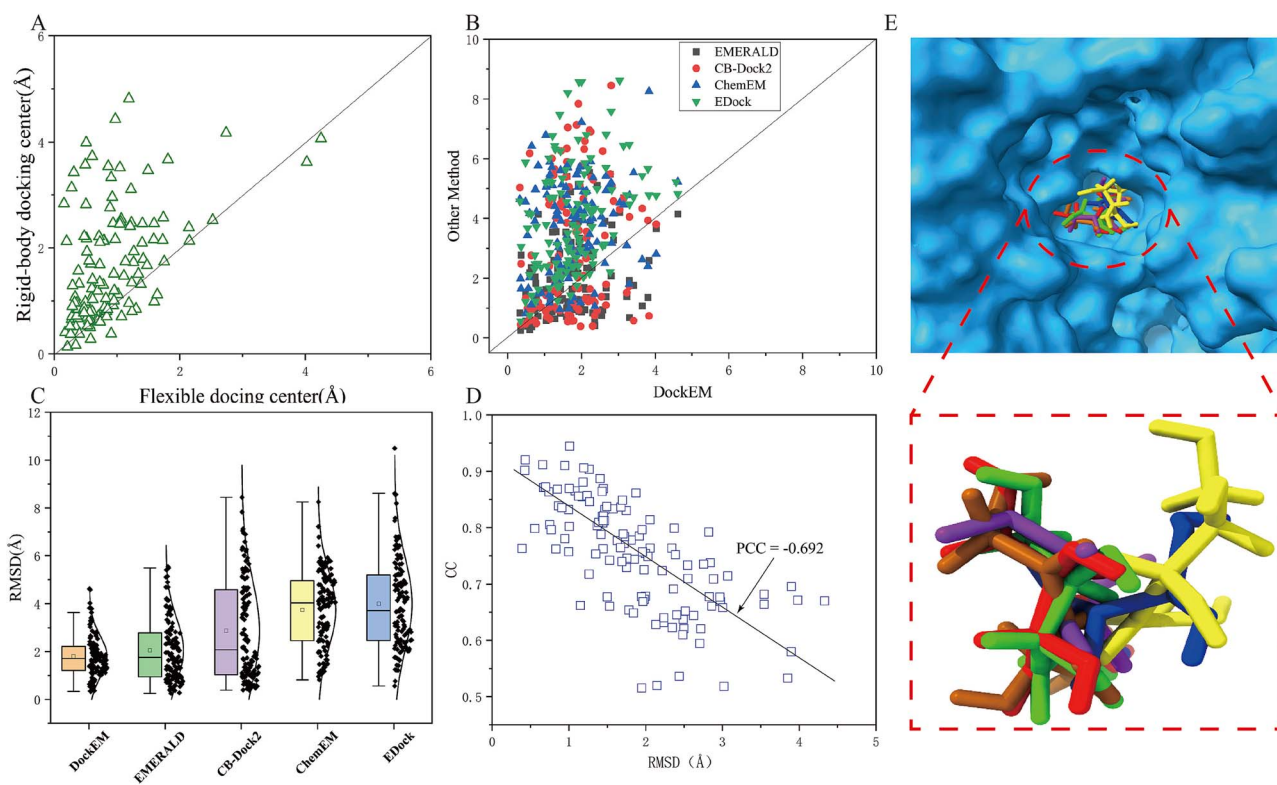


Figure 4. Results for docked ligand using 121 protein–ligand targets. (A) Comparison of the distances between the ligand centers after rigid docking and flexible docking. The x-axis represents the distance between the ligand center and the native center after flexible docking, while the y-axis represents the distance between the ligand center and the native center after rigid docking. (B) Scatter plot comparison of ligand RMSD values obtained by five methods: EMERALD, CB-Dock2, ChemEM, and EDock. (C) Box plot comparison of ligand structure RMSD values obtained by five methods: DockEM, EMERALD, CB-Dock2, ChemEM, and EDock. The y-axis represents RMSD values. (D) Schematic representation of ligand docking for the 3II3A_BS03_XUL complex based on the AF2 predicted model. The RMSD values of the ligand structures obtained by DockEM, EMERALD, ChemEM, CB-Dock2, and EDock relative to the native ligand structure are 0.36 Å, 2.76 Å, 1.65 Å, 2.48 Å, and 3.05 Å, respectively. The background shows the predicted protein model with a TM-score of 0.99. (E) RMSD values of protein–ligand complexes versus CC values from the local density map in the dataset. The RMSD values on the x-axis and CC values on the y-axis. The solid line shows the linear regression fit for all data points, with a Pearson correlation coefficient (PCC) of -0.692 .

over other methods on this metric. The average distance between the ligand center and the native structure center was 0.94 Å. Compared to binding site, the center distance with the native structure center was reduced by 81.4%; compared to EMERALD, by 15.96%; compared to ChemEM, by 44.25%; compared to CB-Dock2, by 49.19%; and compared to EDock, by 65.06%. These results demonstrate that the rigid fitting component of our method is effective in accurately locating the ligand docking position, even before flexible docking is applied. This initial docking position advantage partly explains why our method achieves superior RMSD performance in the flexible docking stage. Detailed information on distances of the binding site center, the ligand's center after rigid docking, the ligand's center after flexible docking, and the native structure's center for all dataset examples can be found in [Supplementary File S3](#).

In a dataset of 121 protein–ligand complexes, DockEM successfully docked 110 ligands, yielding a success rate of 90.9%. Since the protein structures used in our study were predicted rather than experimentally determined, we relaxed the success criteria, defining successful docking as a ligand RMSD of <3 Å. DockEM demonstrated a substantially higher docking success rate than the other four methods, with improvements of 12.4% over EMERALD (success rate of 78.5%), 57.8% over ChemEM (success rate of 33.1%), 32.2% over CB-Dock2 (success rate of 58.7%), and 53.7% over EDock (success rate of 37.2%). [Fig. 4B](#) presents a scatter plot of ligand RMSD values obtained by DockEM, EMERALD, ChemEM,

CB-Dock2, and EDock. Most blue triangles and green inverted triangle, representing ChemEM and EDock, are positioned above the diagonal line, indicating that DockEM significantly outperforms these methods. The black squares and red circles, representing EMERALD and CB-Dock2, exhibit improved performance in some instances, likely due to CB-Dock2's ability to identify suitable templates for those specific cases and the pseudo-atomic skeleton constructed by EMERALD from the unmodeled density around the ligand provides valuable information, which leads to enhanced ligand docking structures. However, in the absence of a suitable template and lower-resolution density map, Ligand structures from CB-Dock2 and EMERALD deviate further from the native structure. In contrast, DockEM, which integrates density map information, achieves favorable results even in the absence of templates and when using lower-resolution density maps. As shown in [Fig. 4C](#), under the strict criterion of RMSD <2 Å, DockEM identified 77 examples meeting this standard, compared to only 65 cases for EMERALD and 59 cases for CB-Dock2. Although EMERALD and ChemEM uses density map information to refine docking poses, its effectiveness is dependent on the resolution of the density maps. At lower resolutions (>6 Å), capturing ligand details accurately becomes challenging, leading to the loss of crucial ligand density data during the differencing process.

We present an example of a protein–ligand complex (PDB ID: 3II3A_BS03_XUL), shown in [Fig. 4D](#). The predicted receptor protein achieved a TM-score of 0.99, with a density map resolution of

3.20 Å for both the protein and ligand. By incorporating density map data in the docking process, our DockEM method achieved a low RMSD of 0.36 Å for the docked ligand (illustrated in green in the figure), outperforming EMERALD (2.76 Å, brown), ChemEM (1.65 Å, purple), CB-Dock2 (2.48 Å, blue), and EDock (3.05 Å, yellow). A major factor in DockEM's superior performance is its use of accurate density map information, which provides structural pose data not accessible to the other methods, with the partial exception of EMERALD and ChemEM. While the quality of this pose information relies on the map's resolution, DockEM can effectively use position and orientation data from the density map by precisely identifying the region corresponding to the ligand referred to in our approach as the local density map. This allows DockEM to achieve significantly better docking accuracy and ligand refinement.

Evaluating constructed models using CC

We calculated the CC value for the refined ligand conformations, considering only the ligand region within the local density map. We also tested other CC calculation methods but found these methods weren't sensitive to conformation changes and worse than current local CC. A detailed comparison of the different CC calculation methods can be found in [Supplementary File S4](#). To verify the effectiveness of the CC value in assessing the rigid docking between the ligand structure and the local density map, we analysis the CC values and RMSD values for all protein–ligand complexes along with their corresponding local density maps in the dataset, as shown in [Fig. 4E](#). Here, the CC value is calculated specifically between the ligand and the corresponding ligand region within the density map. We calculated the Pearson CC between the CC value and the RMSD, finding a moderate negative correlation of -0.692 . [Figure 4E](#) reveals a trend of decreasing RMSD values as CC values increase, suggesting that our CC energy term effectively directs the ligand toward the correct conformation. Furthermore, we also investigated the impact of density map resolution on the performance of DockEM and the CC, as shown in [Supplementary File S5](#). We observed that the performance of DockEM decreases as the resolution of the density map decreases, which is expected. As the resolution of the density map lowers, the amount of pose information that can be extracted from the map also diminishes. This leads to a reduction in the number of sampling points within the sampling space, ultimately lowering the probability of obtaining reasonable conformations when using low-resolution density maps.

We also test our method on density map with resolution range from 10 Å to 15 Å the results can be found in the [Supplementary File S7](#). It can be observed that the performance of our method declines when using density maps with resolutions above 10 Å, but it still comparable or better than other methods (as shown in [Supplementary File S8](#)). The average RMSD for these 10 cases was 2.39 Å for DockEM, 2.38 Å for CB-Dock2, 2.72 Å for ChemEM, 2.62 Å for EMERALD, and 2.91 Å for EDock. If the resolution is very low (>15 Å), this leads to the ligand's density map being nearly indistinguishable from noise. Even if the density map corresponds to the ligand region, the low resolution means that the local density map would not contain much effective pose information.

To further assess the fitting accuracy of ligand structures obtained from various docking methods relative to the local density map, we present a representative protein–ligand complex (PDB ID: 2br1A_BS01_PFF) in [Fig. 5](#). Here, the gray region represents the local density map generated by DockEM at an

initial resolution of 3.69 Å, while the blue region shows the predicted receptor protein model, which has a TM-score of 0.95. [Figure 5A](#) illustrates that the ligand structure generated by our DockEM method aligns closely with the local density map, achieving an RMSD of 0.49 Å and a cross-correlation (CC) value of 0.438. [Figure 5B and C](#) shows ligand structures obtained from the EMERALD and ChemEM, which fits reasonably well within the local density map, yielding a CC value of 0.422 and an RMSD of 0.57 Å for EMERALD and a CC value of 0.411 and an RMSD of 0.82 Å for ChemEM. In this case, only a small portion of the ligand extends beyond the local density map. In contrast, [Fig. 5D](#) depicts the ligand structures obtained from the EDock and CB-Dock2 methods, respectively, which display poor alignment with the local density map. Portions of these ligands fall outside the mapped density, resulting in CC and RMSD values of 0.325 and 1.33 Å for CB-Dock2, and 0.344 and 1.22 Å for EDock. Density maps are not used in either of these two methods. Here, they have included the density maps to facilitate a clearer comparison.

Application to experimental density maps

We validated our method using two experimental density maps, with the docking results presented in [Fig. 6](#). The first experimental density map, characterized by low resolution, is from the 26S proteasome WT-Ubp6-UbVS complex in the si state (EMDB ID: EMD-14083, resolution 7.0 Å; corresponding fitted structure PDB ID: 7qo4). After docking with our method, we achieved an RMSD of 0.90 Å and a CC value of 0.43. The green ligand structure in [Fig. 6A](#) illustrates the result from our docking approach. The second experimental density map, which is of medium resolution, originates from the complex of PRMT5 and MEP50 in the presence of MTA and a novel inhibitor (11-2F) (EMDB ID: EMD-27078; resolution 3.14 Å; corresponding fitted structure PDB ID: 8cyi). In this case, our method yielded an RMSD of 0.40 Å and a CC value of 0.47, with the green ligand structure depicted in [Fig. 6B](#). It is worth noting that in the second experiment's density map, the ligand we docked is a pharmacophore from virtual screen leads to a specific inhibitor (11-2F). 11-2F can synergistically inhibit PRMT5 (an oncogenic target for various cancer types) by cooperating with MTA, a co-factor analog accumulated in MTAP $^{-/-}$ cells [41]. After performing docking with our method, a successful docking was achieved, yielding a ligand structure with an RMSD of 0.40 Å. This result provides insights into the ligand's binding pose, which can be used to redesign new analogs that effectively synergize with MAT, thereby improving the catalytic efficiency of PRMT5 inhibition. This case illustrates how DockEM utilizes medium- to low-resolution density maps to dock ligands and address important real-world biological problems. These results demonstrate that our method is not only effective for simulated density maps but also performs well when applied to real experimental density maps.

Conclusion

In this study, we present DockEM, an updated method designed to refine protein–ligand docking poses by incorporating information from cryo-EM density maps into the docking process. We conducted a thorough evaluation of DockEM on a test set of 121 protein–ligand complexes. The average RMSD achieved by DockEM was 1.87 Å, with successful docking of 110 ligands, resulting in a success rate of 90.9%. We also compared DockEM to several other docking programs, including EDock, ChemEM, EMERALD, and CB-Dock2. The results show that DockEM outperforms these

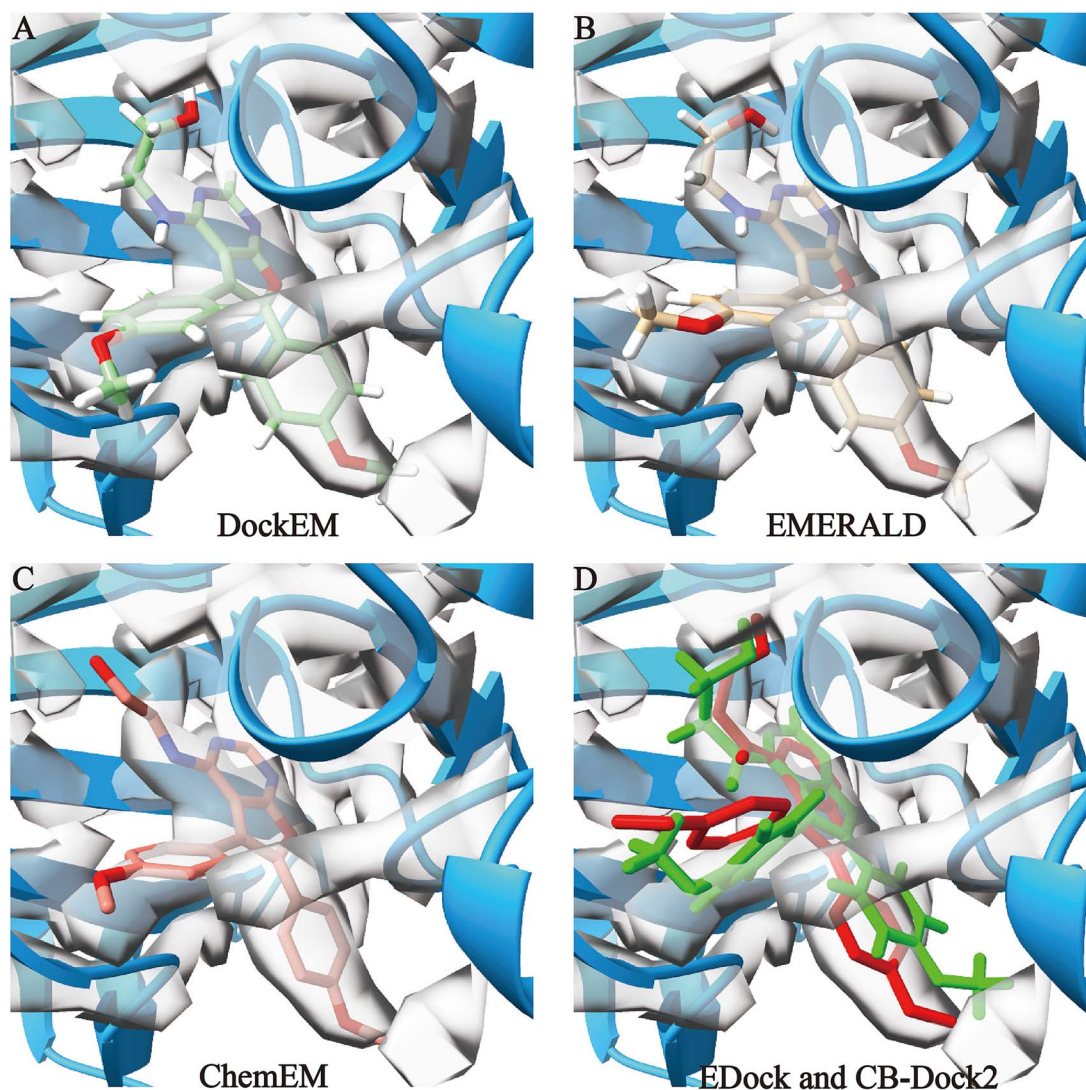


Figure 5. Comparison of ligand fitting to the local density map using five different docking methods for the protein–ligand complex (2br1A_BS01_PFF). The resolution of the local density map is 3.69 Å, while the AF2-predicted protein model with a TM-score of 0.95. (A), (B), and (C) represent the ligand structures obtained by DockEM (RMSD of 0.43 Å and a CC value of 0.438), EMERALD (RMSD of 0.57 Å and a CC value of 0.422), and ChemEM (RMSD of 0.82 Å and a CC value of 0.411). (D) Represent the ligand structures obtained by CB-Dock2 (RMSD of 1.22 Å and a CC value of 0.344) and EDock (RMSD of 1.33 Å and a CC value of 0.325).

programs in molecular docking, likely due to the advantages gained from integrating medium to low-resolution cryo-EM density map data. Furthermore, DockEM was tested on real experimental density maps, and its docking accuracy suggests promising potential for addressing key biological challenges in practical applications. However, the use of low-resolution cryo-EM density maps (with resolutions greater than 10 Å) presents a significant challenge.

Moreover, accurately localizing the ligand's docking position, selecting appropriate parameters for the energy function, and ensuring the precision of cryo-EM density maps add complexity to the global optimization of protein–ligand docking within DockEM. To improve DockEM's performance in future studies, we suggest several avenues for further investigation: (i) Developing methods to more precisely localize the ligand within the entire density map; (ii) Integrating deep learning frameworks for parameter selection to help identify optimal values; and (iii) Exploring optimization algorithms beyond REMC to accelerate the docking

process. Our current approach, which uses cryo-EM density maps to elucidate atomic-level interactions, has demonstrated potential for structure-based drug design. Additionally, this approach could provide valuable insights to support future advancements in protein–peptide docking.

Key Points

- Advanced method for protein–ligand docking using cryo-EM density maps.
- Identify the local density map containing the ligand region from the overall density map to determine the precise docking position of the ligand on the protein.
- Physics- and knowledge-based energy functions, integrated with Cryo-EM density map data, were used to optimize the ligand structure in the REMC simulation.
- Significant ability to refine ligand on the local density maps.

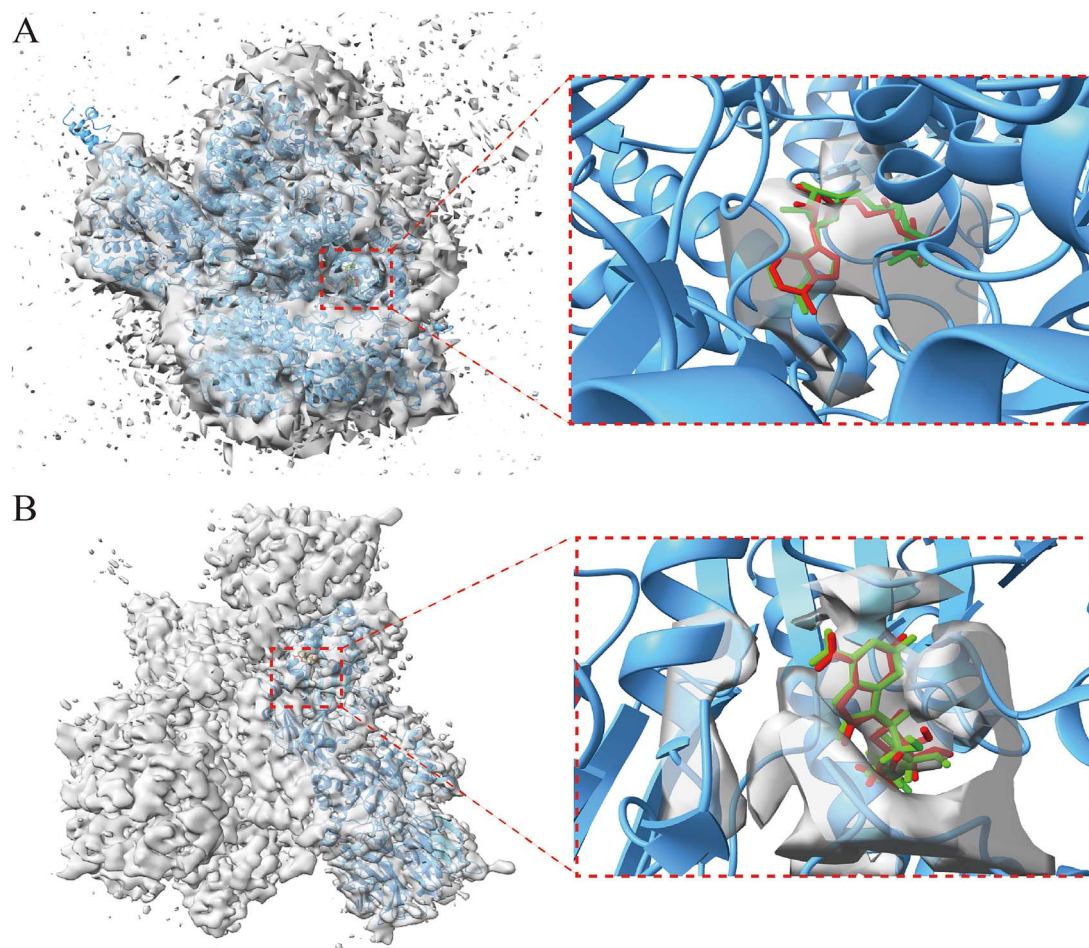


Figure 6. Docking results of ligands using DockEM in two experimental density maps. (A) Low-resolution experimental density map (ID: EMD-14083) from the EMDB, with a resolution of 7.0 Å. The RMSD of the docked ligand is 0.90 Å, and the CC value with the local density map is 0.43. (B) Medium-resolution experimental density map (ID: EMD-27078) from the EMDB, with a resolution of 3.14 Å. The RMSD of the docked ligand is 0.40 Å, and the CC value with the local density map is 0.47.

Author contributions

B.Z and W.Z designed the research; J.Z and B.Z. performed the research; H.J and X.Z participated in discussions and analyzed the data; J.H, X.Z, W.Z, and B.Z participated in discussions and proofread the manuscript; J.Z and B.Z wrote the manuscript, and all authors read and approved the final manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported in part by the National Key R&D Program of China [2022ZD0115103], the National Natural Science Foundation of China (No. 62201506, 62203389), the Youth Project of Zhejiang Provincial Natural Science Foundation of China (LQ22F020028), Postdoctoral Fellowship Program of CPSF (GZC20240159), Chinese Academy of Medical Sciences & Peking Union Medical College, Union Medical College Young Scholar Support Program, Fundamental Research Funds for the Provincial Universities of Zhejiang (RF-C2024006), the funding of Leading

Innovative and Entrepreneur Team Introduction Program of Zhejiang (2023R01006).

Data availability

All data needed to evaluate the conclusions are present in the paper and the Supplementary Materials. The additional data and code related to this paper can be downloaded from <https://github.com/Hsfswa/DockEM>. DockEM is primarily implemented in C++ and does not require the use of a GPU when applying our method. The typical runtime for a molecular docking task using DockEM is approximately 60 min, although larger ligand molecules will increase the runtime. Detailed usage instructions for DockEM can be found in the README section of the code files.

References

1. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature* 2023;**616**:673–85. <https://doi.org/10.1038/s41586-023-05905-z>
2. Gupta KK, Sharma KK, Chandra H. et al. The integrative bioinformatics approaches to predict the xanthohumol as anti-breast cancer molecule: targeting cancer cells signaling PI3K and AKT kinase pathway. *Front Oncol* 2022;**12**:950835. <https://doi.org/10.3389/fonc.2022.950835>

3. Khan S, Mosvi SN, Vohra S. et al. Implication of calcium supplementations in health and diseases with special focus on colorectal cancer. *Crit Rev Clin Lab Sci* 2024;**61**:496–509.
4. Khan S, Zakariah M, Rolfo C. et al. Prediction of *Mycoplasma hominis* proteins targeting in mitochondria and cytoplasm of host cells and their implication in prostate cancer etiology. *Oncotarget* 2016;**8**:30830–43. <https://doi.org/10.18632/oncotarget.8306>
5. Khan S, Zakariah M, Palaniappan S. Computational prediction of *Mycoplasma hominis* proteins targeting in nucleus of host cell and their implication in prostate cancer etiology. *Tumor Biol* 2016;**37**:10805–13. <https://doi.org/10.1007/s13277-016-4970-9>
6. Khan S, Imran A, Khan AA. et al. Systems biology approaches for the prediction of possible role of *Chlamydia pneumoniae* proteins in the etiology of lung cancer. *PLoS One* 2016;**11**:e0148530. <https://doi.org/10.1371/journal.pone.0165718>
7. Wang Y, Imran A, Shami A. et al. Decipher the *Helicobacter pylori* protein targeting in the nucleus of host cell and their implications in gallbladder cancer: an in silico approach. *J Cancer* 2021;**12**:7214. <https://doi.org/10.7150/jca.63517>
8. Zhou X, Li Y, Zhang C. et al. Progressive assembly of multi-domain protein structures from cryo-EM density maps. *Nat Comput Sci* 2022;**2**:265–75. <https://doi.org/10.1038/s43588-022-00232-1>
9. Zhang Z, Cai Y, Zhang B. et al. EM2: assembling protein complex structures from cryo-EM maps through intertwined chain and domain fitting. *Brief Bioinform* 2024;**25**:bbae113. <https://doi.org/10.1093/bib/bbae113>
10. Salentin S, Schreiber S, Haupt VJ. et al. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res* 2015;**43**:W443–7. <https://doi.org/10.1093/nar/gkv315>
11. Convertino M, Das J, Dokholyan NV. Pharmacological chaperones: design and development of new therapeutic strategies for the treatment of conformational diseases. *ACS Chem Biol* 2016;**11**:1471–89. <https://doi.org/10.1021/acscchembio.6b00195>
12. Ding F, Yin S, Dokholyan NV. Rapid flexible docking using a stochastic rotamer library of ligands. *J Chem Inf Model* 2010;**50**:1623–32. <https://doi.org/10.1021/ci100218t>
13. Wang J, Dokholyan NV. MedusaDock 2.0: efficient and accurate protein–ligand docking with constraints. *J Chem Inf Model* 2019;**59**:2509–15. <https://doi.org/10.1021/acs.jcim.8b00905>
14. Sweeney A, Mulvaney T, Maiorca M. et al. ChemEM: flexible docking of small molecules in cryo-EM structures. *J Med Chem* 2023;**67**:199–212. <https://doi.org/10.1021/acs.jmedchem.3c01134>
15. Wang N, Jiang X, Zhang S. et al. Structural basis of human monocarboxylate transporter 1 inhibition by anti-cancer drug candidates. *Cell* 2021;**184**:e313. <https://doi.org/10.1016/j.cell.2020.11.043>
16. Conde J, Pumroy RA, Baker C. et al. Allosteric antagonist modulation of TRPV2 by piperlongumine impairs glioblastoma progression. *ACS Central Sci* 2021;**7**:868–81. <https://doi.org/10.1021/acscentsci.1c00070>
17. Nakane T, Kotecha A, Sente A. et al. Single-particle cryo-EM at atomic resolution. *Nature* 2020;**587**:152–6. <https://doi.org/10.1038/s41586-020-2829-0>
18. Lawson CL, Patwardhan A, Baker ML. et al. EMDatabank unified data resource for 3DEM. *Nucleic Acids Res* 2016;**44**:D396–403. <https://doi.org/10.1093/nar/gkv1126>
19. Masoudi-Sobhanzadeh Y, Jafari B, Parvizpour S. et al. A novel multi-objective metaheuristic algorithm for protein-peptide docking and benchmarking on the LEADS-PEP dataset. *Comput Biol Med* 2021;**138**:104896. <https://doi.org/10.1016/j.compbio.2021.104896>
20. Cheng H, Wang G-G, Chen L. et al. A dual-population multi-objective evolutionary algorithm driven by generative adversarial networks for benchmarking and protein-peptide docking. *Comput Biol Med* 2024;**168**:107727. <https://doi.org/10.1016/j.compbio.2023.107727>
21. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61. <https://doi.org/10.1002/jcc.21334>
22. Allen WJ, Balias TE, Mukherjee S. et al. DOCK 6: impact of new features and current docking performance. *J Comput Chem* 2015;**36**:1132–56. <https://doi.org/10.1002/jcc.23905>
23. Zhang W, Bell EW, Yin M. et al. EDock: blind protein–ligand docking by replica-exchange Monte Carlo simulation. *J Chem* 2020;**12**:1–17. <https://doi.org/10.1186/s13321-020-00440-9>
24. Liu Y, Yang X, Gan J. et al. CB-Dock2: improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Res* 2022;**50**:W159–64. <https://doi.org/10.1093/nar/gkac394>
25. Muenks A, Zepeda S, Zhou G. et al. Automatic and accurate ligand structure determination guided by cryo-electron microscopy maps. *Nat Commun* 2023;**14**:1164. <https://doi.org/10.1038/s41467-023-36732-5>
26. Liu Y, Grimm M, Dai W-T. et al. CB-dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacol Sin* 2020;**41**:138–44. <https://doi.org/10.1038/s41401-019-0228-6>
27. Robertson MJ, van Zundert GC, Borrelli K. et al. GemSpot: a pipeline for robust modeling of ligands into cryo-EM maps. *Structure* 2020;**28**:e703–716.e3. <https://doi.org/10.1016/j.str.2020.04.018>
28. Jumper J, Evans R, Pritzel A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. <https://doi.org/10.1038/s41586-021-03819-2>
29. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 1999;**314**:141–51. [https://doi.org/10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9)
30. Mysinger MM, Carchia M, Irwin JJ. et al. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 2012;**55**:6582–94. <https://doi.org/10.1021/jm300687e>
31. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 2013;**29**:2588–95. <https://doi.org/10.1093/bioinformatics/btt447>
32. Tang G, Peng L, Baldwin PR. et al. EMAN2: an extensible image processing suite for electron microscopy. *J Struct Biol* 2007;**157**:38–46. <https://doi.org/10.1016/j.jsb.2006.05.009>
33. Pettersen EF, Goddard TD, Huang CC. et al. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 2004;**25**:1605–12. <https://doi.org/10.1002/jcc.20084>
34. Zhang B, Zhang X, Pearce R. et al. A new protocol for atomic-level protein structure modeling and refinement using low-to-medium resolution Cryo-EM density maps. *J Mol Biol* 2020;**432**:5365–77. <https://doi.org/10.1016/j.jmb.2020.07.027>
35. Patwardhan A. Trends in the electron microscopy data bank (EMDB). *Acta Crystallogr D Struct Biol* 2017;**73**:503–8. <https://doi.org/10.1107/S2059798317004181>
36. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?

- J Comput Chem* 2000;**21**:1049–74. [https://doi.org/10.1002/1096-987X\(200009\)21:12<1049::AID-JCC3>3.0.CO;2-F](https://doi.org/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F)
37. Wang J, Wang W, Kollman PA. et al. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 2006;**25**:247–60. <https://doi.org/10.1016/j.jmgm.2005.12.005>
 38. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo Method*. Hoboken, NJ: John Wiley & Sons, 2016. <https://doi.org/10.1002/9781118631980>.
 39. Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *J Chem* 2019;**11**:1–9.
 40. Landrum G. *Rdkit Documentation Release*, Vol. **1**, 2013. 4.
 41. Zhou W, Yadav GP, Yang X. et al. Cryo-EM structure-based selection of computed ligand poses enables design of MTA-synergic PRMT5 inhibitors of better potency. *Commun Biol* 2022;**5**:1054. <https://doi.org/10.1038/s42003-022-03991-9>