# Progressive Assembly of Multi-Domain Protein Structures from Cryo-EM Density Maps

## Group Meeting

Yu-Hsiang Lien 連昱翔

Article | Published: 28 April 2022

**Progressive assembly of multi-domain protein structures from cryo-EM density maps**

Xiaogen Zhou, Yang Li, Chengxin Zhang, Wei Zheng, Guijun Zhang & Yang Zhang

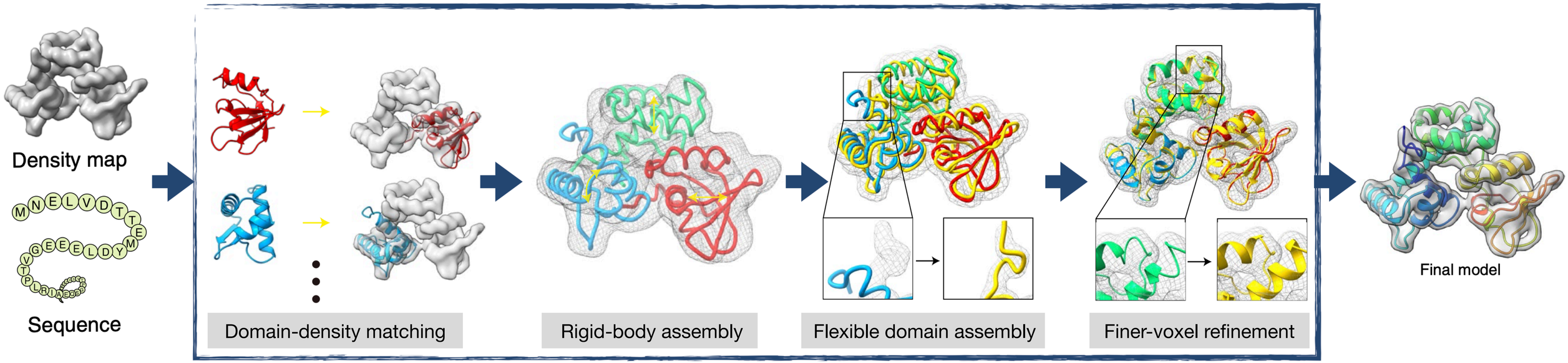
*Nature Computational Science* **2**, 265–275 (2022) | Cite this article

**2894** Accesses | **9** Citations | **2** Altmetric | Metrics

*Ref: https://doi.org/10.1038/s43588-022-00232-1*

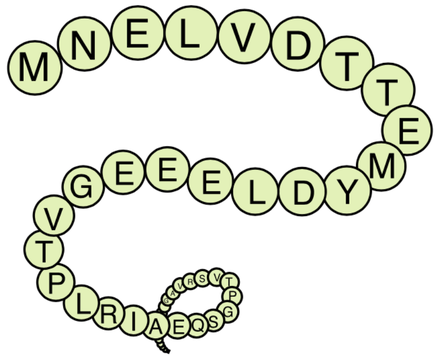
# Domain Enhanced Modeling using Cryo-EM (DEMO-EM)

- **DEMO-EM** is a hierarchical approach to **multi-domain protein structure determination** based on **cryo-EM density maps**:
  - Determining **domain boundaries** and **modeling individual domains**.
  - **Matching domain models** with a density map for the initial framework generation.
  - Rigid-body domain structure assembly for **domain position** and **orientation** optimization.
  - Flexible structure assembly and refinement simulation of full-length structural models.



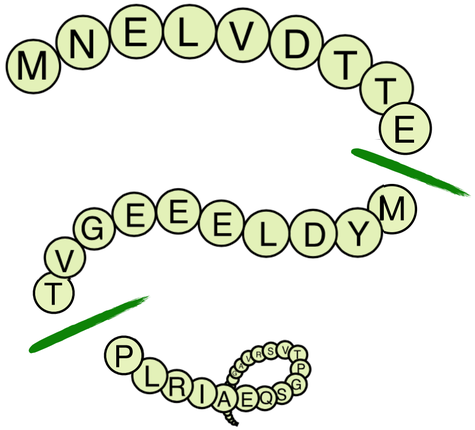
# Domain Parsing and Domain Structure Folding

**Query Sequence**



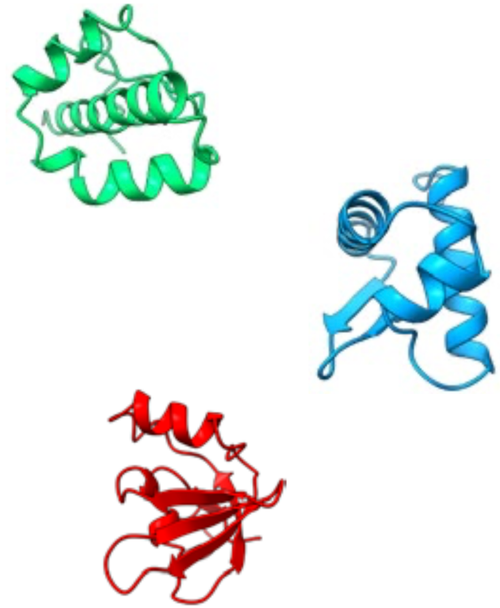
ThreaDom
or
FUpred

**Domain Boundary Prediction**



D-I-TASSER

**Domain Structure Folding**



- The DEMO-EM pipeline can start from either **experimentally determined domain structures** or **amino acid sequences**.
- When starting from amino acid sequences : Query amino acid sequence ➡ **LOMETS** (create multiple template alignments from the PDB) ➡ **ThreaDom** or **FUpred** (predict the domain boundary according to the domain conservation score) ➡ **D-I-TASSER** (generate the structural model of each domain)

# Matching of Domain Structure & Cryo-EM Density Map

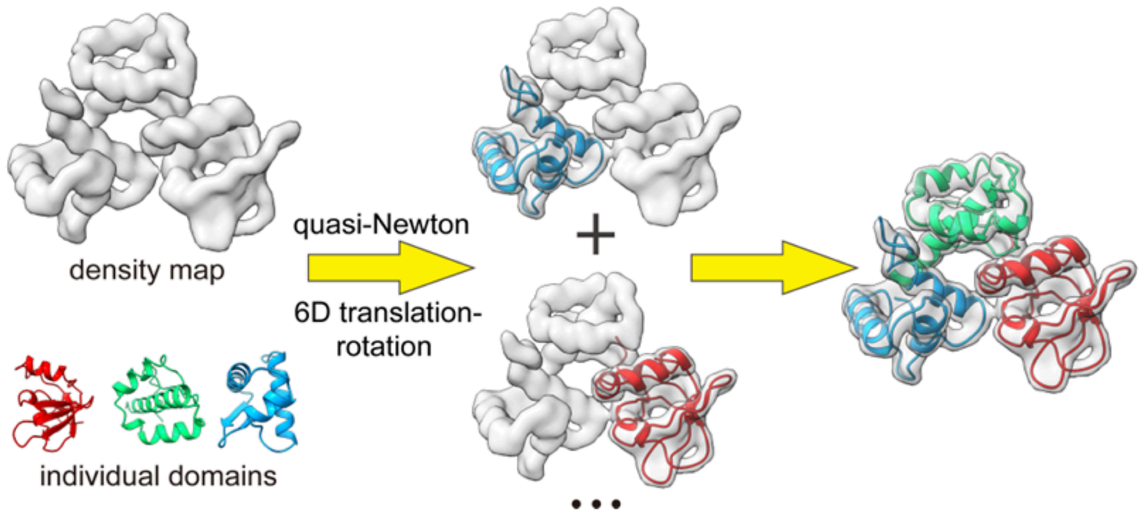
- Using Limited-memory BFGS (**L-BFGS**), a quasi-Newton optimization algorithm with 6D translation–rotation degrees of freedom, to identify the best location and orientation of each individual domain with the highest correlation with the map.
- Since L-BFGS is a local optimization method, **multiple initial positions** and **orientations** by enumerating all combinations of Euler angles with a step size of $S_{\text{rot\_ang}} = 30^{\circ}$ across the density-map space are used.
- For a given domain pose, a **Density Correlation Score** (**DCS**) calculated as following is used to guild the L-BFGS simulations:

$$E_{\text{DCS}} = 1 - \frac{\sum_{i=1}^{N_{\text{vol}}} \left(\rho_{\text{EM}}(\mathbf{v}_i) - \bar{\rho}_{\text{EM}}\right)\left(\rho_{\text{MO}}(\mathbf{v}_i) - \bar{\rho}_{\text{MO}}\right)}{\sqrt{\sum_{i=1}^{N_{\text{vol}}} \left(\rho_{\text{EM}}(\mathbf{v}_i) - \bar{\rho}_{\text{EM}}\right)^2 \sum_{i=1}^{N_{\text{vol}}} \left(\rho_{\text{MO}}(\mathbf{v}_i) - \bar{\rho}_{\text{MO}}\right)^2}}$$

- The modeling of density map for domain model:

$$\rho_{\text{MO}}(\mathbf{v}_i) = \sum_{j=1}^{L} m \sqrt[3]{\left(\frac{\pi}{(2.4+0.8\sigma)^2}\right)^2} \exp\left[-\left(\frac{\pi}{2.4+0.8\sigma}\left|\mathbf{v}_i - \mathbf{x}_j\right|^2\right)^2\right]$$

$m$ : atom mass , $\sigma$ : resolution , $\mathbf{x}_j$ : coordinate of $j$ atom




All poses for each domain with DCS < 0.5 are kept and combined (permuting the initial poses of all the domains and allows for domain overlaps) with the top poses of other domains to form the initial models of the full-length models (kept top 30 full-length models with the lowest DCS).

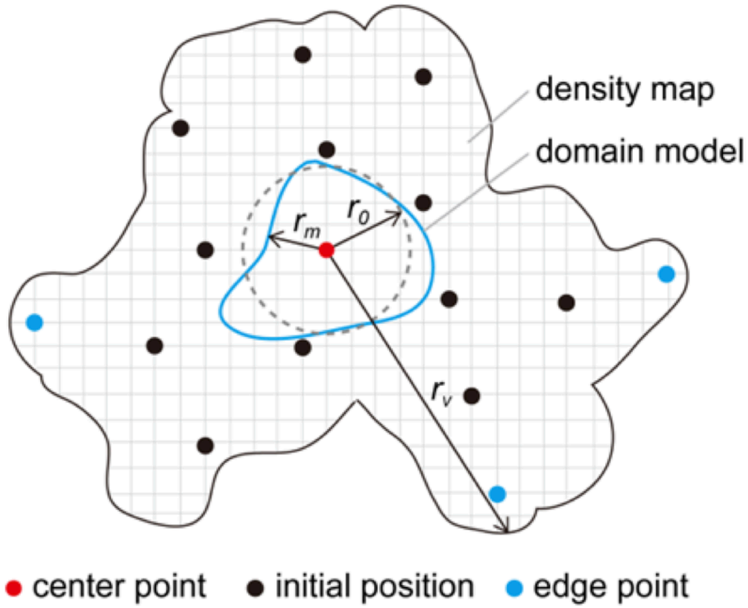
# Matching of Domain Structure & Cryo-EM Density Map

- In the enumeration of **initial position** determination for a domain model:
  - To eliminate redundant positions, the **minimum distance between two neighboring positions** is set:

$$r = \max\left(0.5r_m, r_0\right)$$

$r_m$ : the radius of gyration of the model

$r_0 = 5\text{Å}$ : the minimum distance between two initial positions




  - To remove the edge positions, the **maximum distance between each initial position and the center point of the density map** is set as:

$$r_{\text{edge}} = \min\left(\max\left(1.1(r_v - r_m), r_0\right), r_v\right)$$

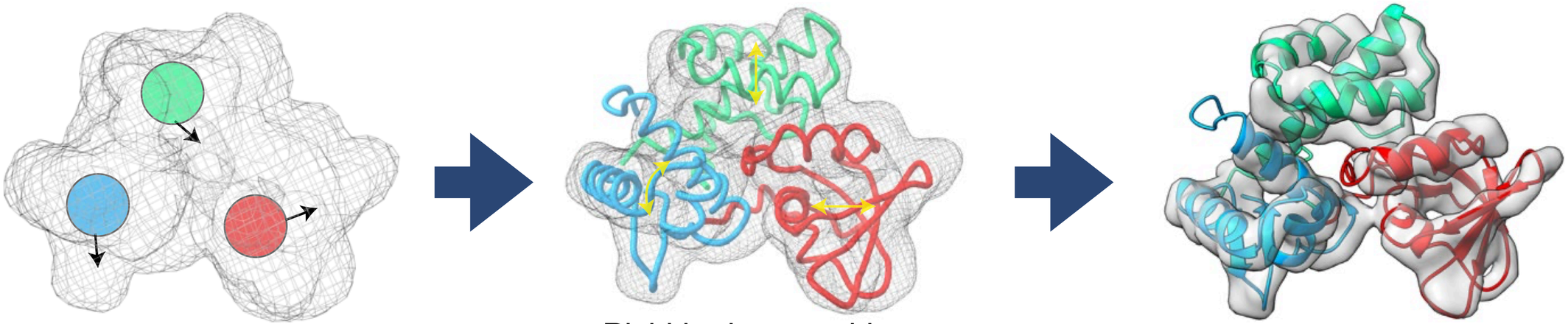
$r_v = \sqrt{\dfrac{\sum_{i=1}^{N'_{\text{vol}}}\left(\mathbf{v}_i - \mathbf{v}_{\text{center}}\right)^2}{N'_{\text{vol}}}}$ : the radius of gyration of the density map calculated by the $N'_{\text{vol}}$ voxels with density $\geq 0.05$ after normalization

$\mathbf{v}_{\text{center}} = \dfrac{\sum_{j=1}^{N'_{\text{vol}}}\mathbf{v}_j}{N'_{\text{vol}}}$ : the center point of these voxels

# Rigid-body Domain Assembly

- Two rounds of rigid-body domain assembly simulations are performed to **optimize the domain positions and orientations**:
  - **Round 1**: The domains are treated as **particles** and a quick **REMC simulation** is carried out to adjust the positions of the individual domains based on the **global model-density correlations** (e.g. DCS).
  - **Round 2**: Fine-tune the domain poses with a more **detailed energy force field**.

**REMC** (Replica-Exchange Monte Carlo) : a computer simulation method typically used to find the lowest energy state of a system of many interacting particles.



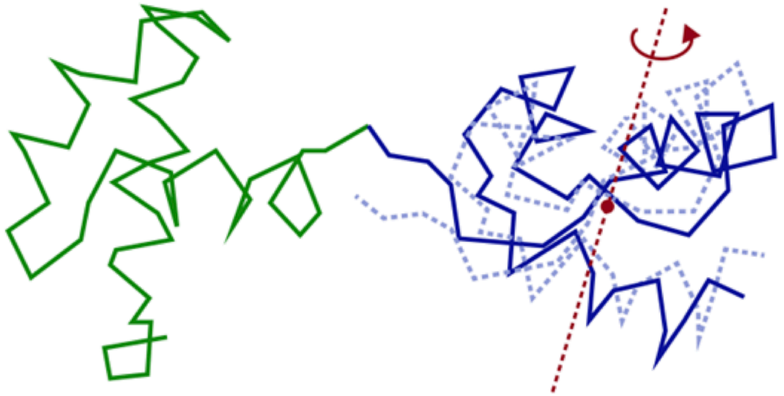
Particle movement

Rigid-body assembly

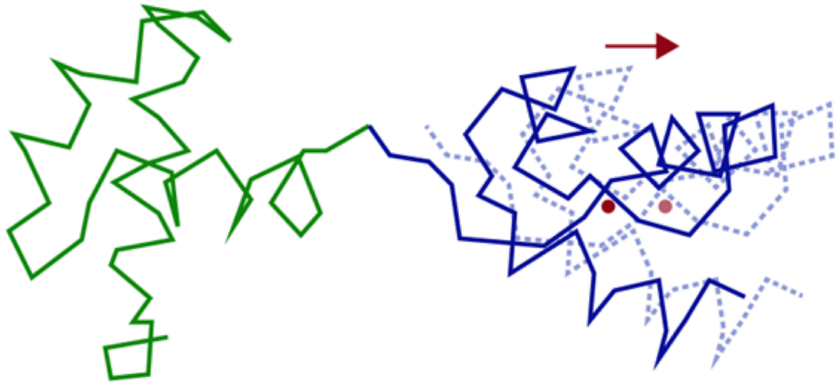
Lowest DCS model

# Rigid-body Domain Assembly — Round 1

- **Round 1**: The domains are treated as **particles** and a quick **REMC simulation** is carried out to adjust the positions of the individual domains based on the **global model-density correlations** :
  - The energy function contains only DCS $E_{\mathrm{DCS}}$ which is calculated for the full chain model.
  - The movements include rigid-body **translation** and **rotation** around each domain's centre of mass.
  - The density map with a voxel size of $3\text{Å}$ interpolated from the original map is applied.
  - The top 30 models according to the DCS are selected for the next round.



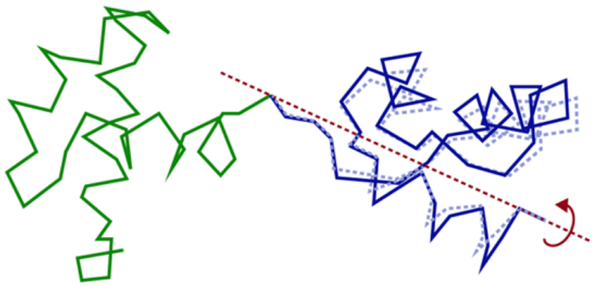
**Random rigid-body rotation around the domain's center of mass**



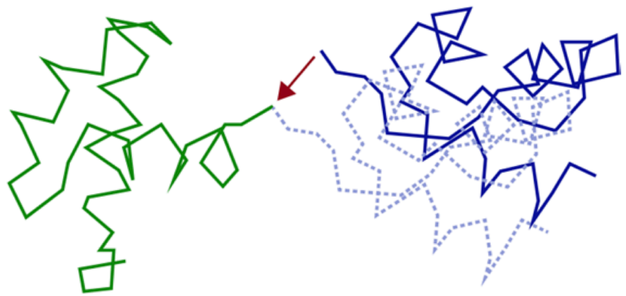
**Random rigid-body translation of the domain's center of mass**

# Rigid-body Domain Assembly — Round 2

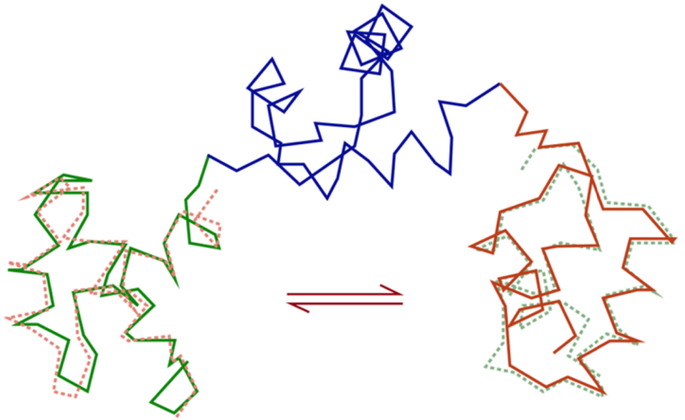
- **Round 2**: Fine-tune the domain poses with a more **detailed energy force field** :
  - **Three new movements** are added: self-rotation around the $N$-to-$C$ axis of each domain, translation along the neighboring domains in the sequence and pose exchange between two domains with similar structures.
  - Use a more elaborate density map with a voxel size of $2\text{Å}$ is interpolated from the original density map for the assembly.
  - The top 40 models according to the DCS are selected for the next step.



**Random rigid-body rotation around the axis connecting the domain's $N$- and $C$-terminal $C_\alpha$ atoms**



**Rigid-body translation along the axis connecting two domains which are neighboring in sequence**



**Pose exchange between two domains with similar structures ( TM-score $\geq 0.75$ )**

# Energy Function for Rigid-body Simulation

- Conformations in the rigid-body assembly are assessed by using an energy function with four terms:

$$E_{\text{Rigid}} = w_{\text{DCS}}E_{\text{DCS}} + w_{\text{RG}}E_{\text{RG}} + w_{\text{BC}}\sum_{m=1}^{N_{\text{dom}}-1} E_{\text{BC}}(m, m+1) + w_{\text{SC}}\sum_{m=1}^{N_{\text{dom}}}\sum_{n=m+1}^{N_{\text{dom}}} E_{\text{SC}}(m, n)$$

**Density Correlation Score**

**Radius-of-Gyration restraint**

$$E_{\text{RG}} = \begin{cases} (r_{\max} - r_{\text{decoy}})^2 & , \quad \text{if } r_{\text{decoy}} > r_{\max} \\ (r_{\text{decoy}} - r_{\min})^2 & , \quad \text{if } r_{\text{decoy}} > r_{\min} \\ 0 & , \quad \text{otherwise} \end{cases}$$

**Domain Boundary Connectivity**

$$E_{\text{BC}}(m, n) = \left(b_{mn} - b_0\right)^2$$

**Steric Clashes**

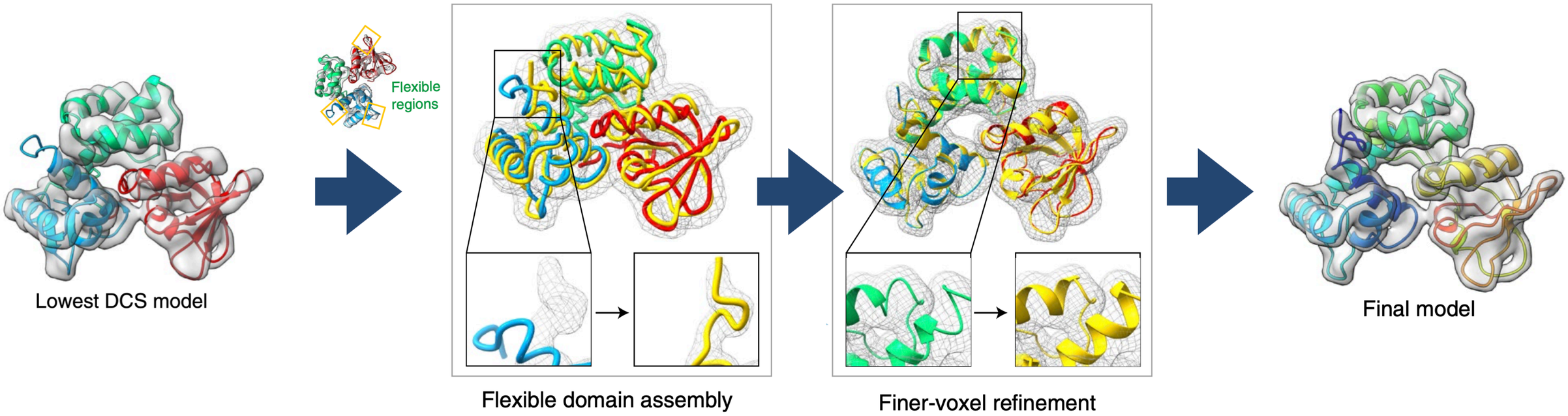
$$E_{\text{SC}}(m, n) = \sum_{i=1}^{L_m}\sum_{j=1}^{L_n} \begin{cases} \frac{1}{d_{ij}^{mn}} & , \quad \text{if } d_{ij}^{mn} < d_{\text{cut}} \\ 0 & , \quad \text{otherwise} \end{cases}$$

- The weighting factors are optimized based on a training set of 425 proteins by maximizing the correlation between the total energy and RMSD of the decoy models with respect to the native using the differential evolution algorithm:

$$w_{\text{DCS}} = 300, \quad w_{\text{RG}} = 1.13, \quad w_{\text{BC}} = 0.55, \quad w_{\text{SC}} = 0.91$$

# Atomic-Level Flexible Domain Assembly & Refinement

- The process of flexible domain assembly and refinement contains two stages of simulations with progressive voxel resolutions and sampling focuses.
  - **Stage 1**: Flexible domain assembly by considering six different movements, **an atomic-level force field** is designed to guide the REMC simulation.
  - **Stage 2**: Refinement in a finer density map with the DCS computed on all atoms.



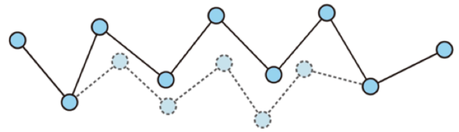
# Atomic-Level Flexible Domain Assembly — Stage 1

- **Stage 1**: Flexible domain assembly by considering six different movements, **an atomic-level force field** is designed to guide the REMC simulation.
  - To enhance the efficiency, a **nine-residue sliding window** is used to determine which region needs more aggressive conformation sampling.
  - The probability for the $i$-th residue **to be selected** for movement is set as:

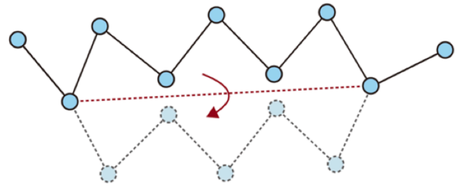
$$p_i = \begin{cases} 1 & , \quad \text{if } S_i < 0.05 \\ 0.95\left(1 - \dfrac{S_i - S_{\min}}{S_{\max} - S_{\min}}\right) & , \quad \text{if } S_i \geq 0.05 \end{cases}$$

$S_i$ : a local score for the sliding window of the centre residue $i$ is computed as the average correlation coefficient between the nine-residue fragment and the entire density map.

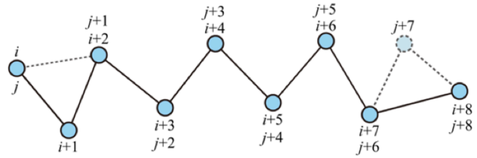
  - A density map with voxel size of $3\text{Å}$ interpolated from the original density map is applied to reduce the computation cost and the DCS is calculated based on **backbone atoms**.
  - All accepted decoys in the simulation are clustered by **SPICKER**, and the **centroid model** in the first cluster is selected as a reference model for the Stage 2.



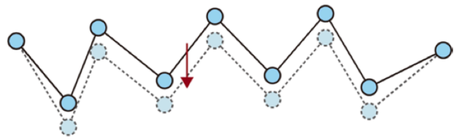
**LMPort Perturbation**



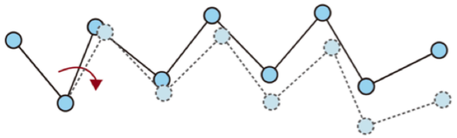
**Segment Rotation**



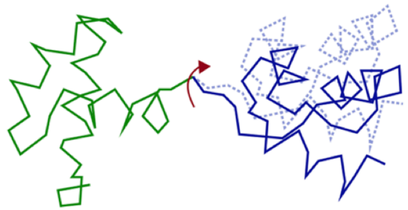

**Conformational Shift of Segments**



**Rigid-body Segment Translation**



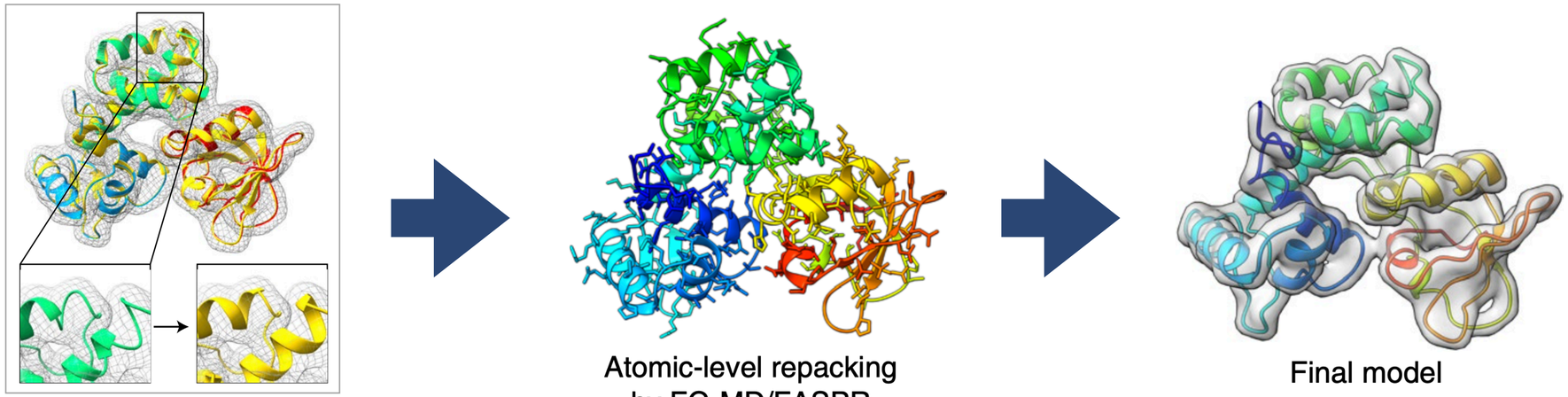
**Rigid-body Tail Rotation**



**Rigid-body Domain-level Translation and Rotation**

# Atomic-Level Flexible Domain Refinement — Stage 2

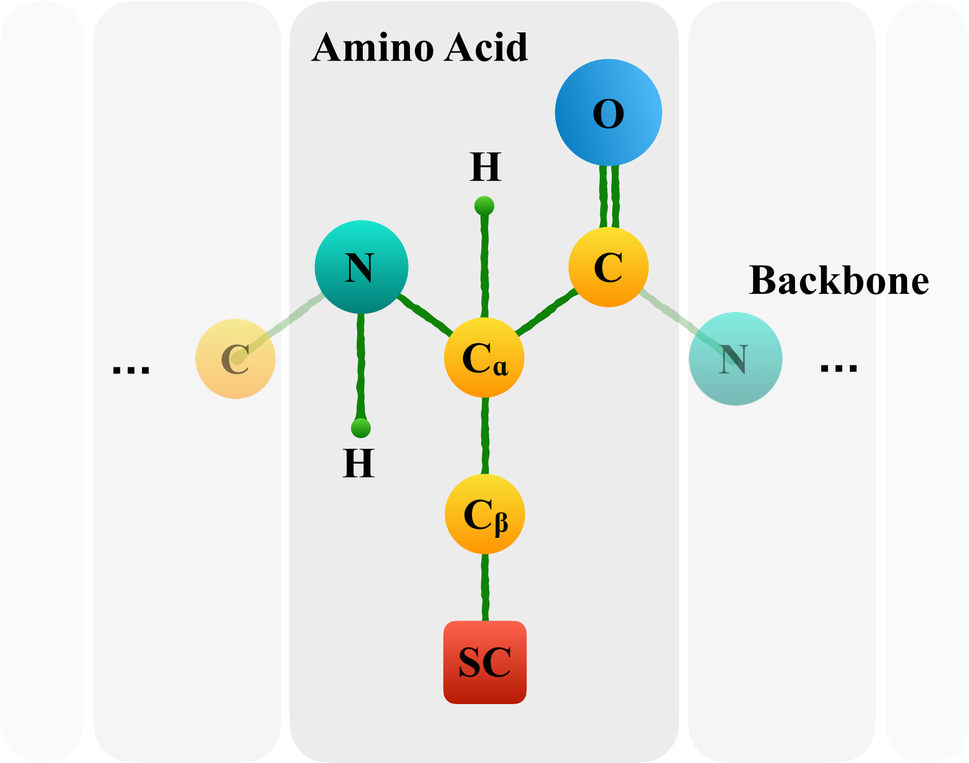
- **Stage 2**: Refinement in a finer density map with the DCS computed on all atoms.
  - A finer density map with voxel size of 2Å is implemented with the DCS computed on all atoms.
  - All residues have **equal probability** to be selected for movement and sampling.
  - The REMC simulation is guided by the same force field, but the reference model is replaced by the **centroid structure** from the Stage 1.
  - The lowest-energy decoy is selected to construct the final model, with the side-chain atoms repacked by **FASPR** followed by **FG-MD** refinement.



# DEMO-EM Force Field for Flexible Assembly Simulation

- The flexible domain assembly simulations are implemented at a **semi-atomic level**, with each **residue** represented by $N$, $C_\alpha$, $C$, $O$, $C_\beta$, $H$ and **side-chain centre of mass** (SC).

- Among the **seven modeling units**, only the **3 backbone atoms** ($N$, $C_\alpha$, $C$) have coordinates determined directly in conformation sampling. (The rest of 4 atoms determined based on their position relative to backbone atoms)



# DEMO-EM Force Field for Flexible Assembly Simulation

- The simulations are guided by a composite force field consisting of **seven** energy terms:

$$E_{\text{Flexible}} = w_{\text{DCS}}E_{\text{DCS}} + w_{\text{DT}}\sum_{m=1}^{N_{\text{dom}}}\sum_{n=m+1}^{N_{\text{dom}}}E_{\text{DT}}(m,n) + w_{\text{TA}}E_{\text{TA}} + w_{\text{DR}}E_{\text{DR}} + \sum_{i=1}^{L}\sum_{j=i+1}^{L}\left[w_{\text{EV}}E_{\text{EV}}(i,j) + w_{\text{HB}}E_{\text{HB}}(i,j,T_k) + w_{\text{GSC}}E_{\text{GSC}}(i,j)\right]$$

**Density Correlation Score**

**Inter-Domain $C_\beta$ distance map**

**Torsion Angle variation**

**Domain Structure Restrain**

**Excluded Volume interaction**

**Hydrogen Bounding**

**Generic Side-Chain-atom contact potential**

- The weighting parameters are determined by maximizing the correlation between the total energy and RMSD of the structure decoys of the 425 training proteins:

$$w_{\text{DCS}} = 320\ ,\quad w_{\text{DT}} = 0.15\ ,\quad w_{\text{TA}} = 0.3\ ,\quad w_{\text{DR}} = 1.5\ ,\quad w_{\text{EV}} = 0.1\ ,\quad w_{\text{HB}} = 0.05\ ,\quad w_{\text{GSC}} = 0.1$$

# Template Modeling Score (TM-Score)

- **TM-score** is a metric for evaluating the **topological similarity** between protein structures.

$$\text{TM score} = \max \left[ \frac{1}{L_{\text{target}}} \sum_{i=1}^{L_{\text{aligned}}} \frac{1}{1 + \left( \frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right]$$

$L_{\text{target}}$ : the amino acid sequence length of the target protein

$L_{\text{aligned}}$ : the length of the aligned residues to the reference (native) structure

$d_i$ : the distance between the $i$-th pair of aligned residues

$d_0(L_{\text{target}}) = 1.24\sqrt[3]{L_{\text{target}} - 15} - 1.8$ : normalization factor

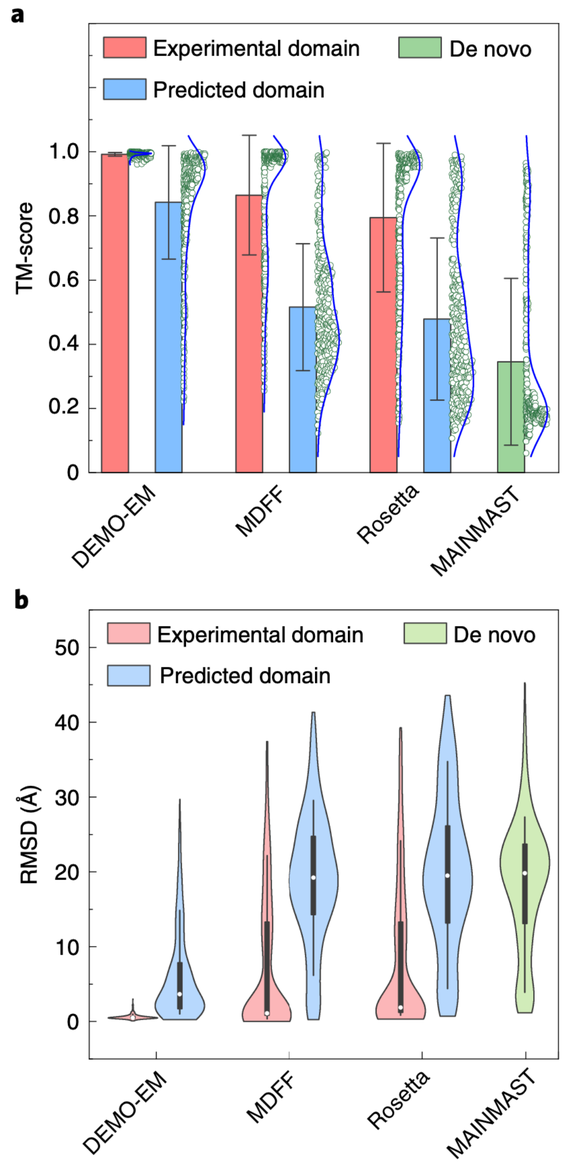
- Stringent statistics showed that **TM-score > 0.5** corresponds to a similarity with two structures having the **same fold** and/or **domain orientations**.
- TM-score is **protein length-independent** and its value is **more sensitive to the fold and domain orientation** similarities of the predicted model relative to the native, compared to **RMSD**.

# Structures Assembly from Synthesized Density Maps

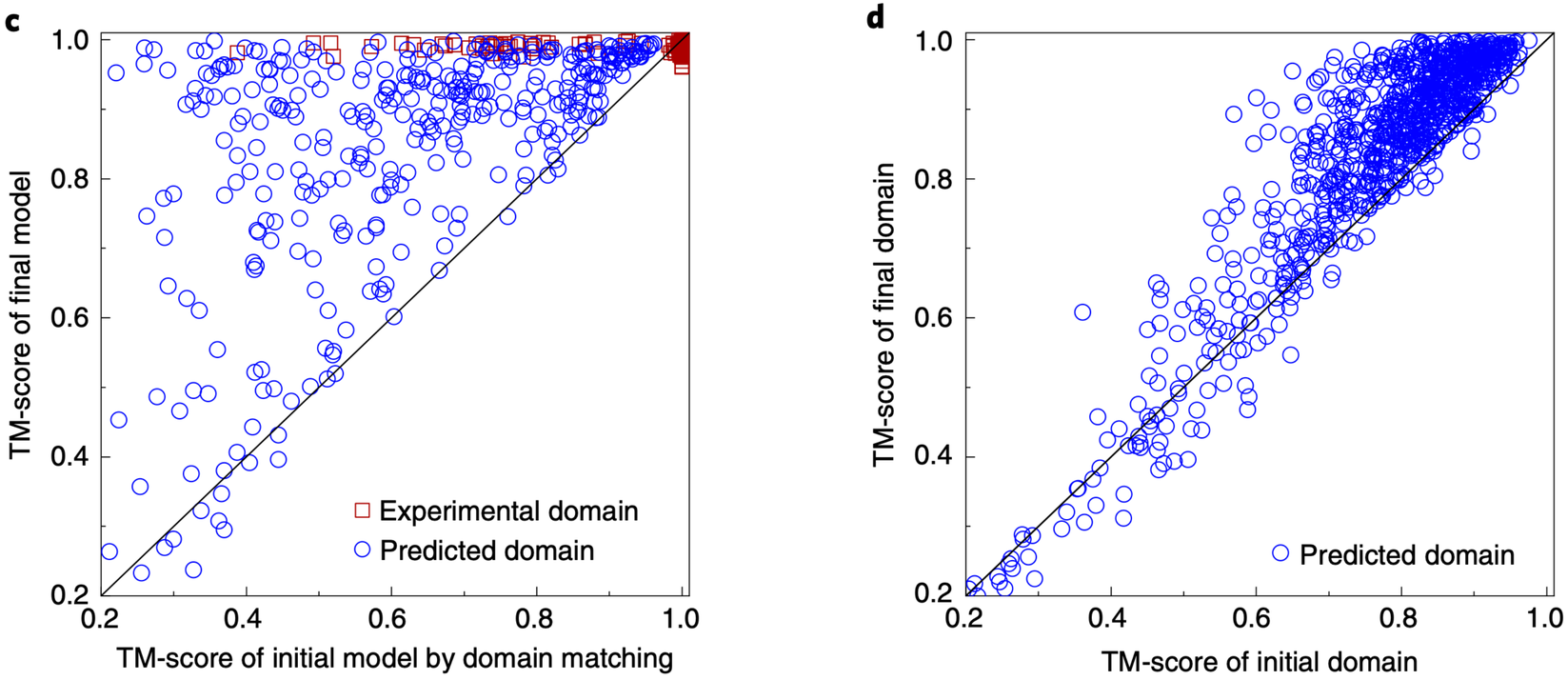
- Two separate tests are performed for evaluating DEMO-EM performance:
  - Use **experimental domain structures** extracted from the full-length target structures.
  - Use **predicted domain models** by **D-I-TASSER**.

**Table 1 | Results for the 357 test proteins using synthesized density maps**

| | MDFF | Rosetta | MAINMAST | DEMO-EM |
|---|---|---|---|---|
| Experimental domain structure assembly | | | | |
| TM-score | 0.86 (0.20) | 0.79 (0.23) | - | **0.99** (0.01) |
| RMSD (Å) | 7.1 (9.4) | 8.1 (9.9) | - | **0.6** (0.3) |
| Predicted domain model assembly | | | | |
| TM-score | 0.53 (0.22) | 0.45 (0.22) | 0.35 (0.26) | **0.85** (0.17) |
| RMSD | 16.6 (8.1) | 21.2 (10.9) | 18.3 (8.6) | **5.9** (6.4) |
| TM-score (domain)[a] | 0.63 (0.22) | 0.48 (0.26) | 0.32 (0.25) | **0.83** (0.16) |
| RMSD (domain)[b] | 5.9 (3.8) | 9.3 (7.1) | 13.7 (6.9) | **3.9** (3.8) |

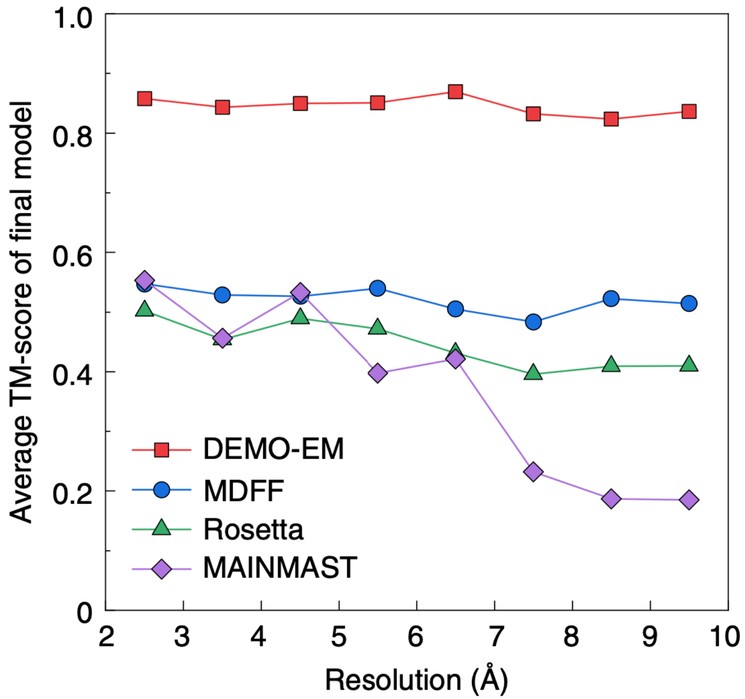


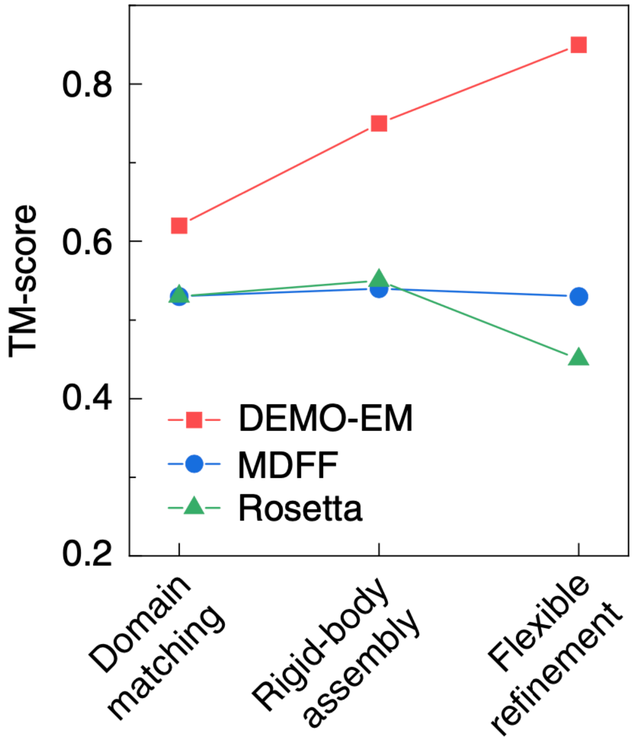
# Structures Assembly from Synthesized Density Maps




- Because a model with a **high TM-score** must achieve correct modeling of both **individual domains and inter-domain orientations**, the result indicate that DEMO-EM domain assembly simulations could significantly improve the **inter-domain orientations**.
- The domain structures are kept **flexible** in DEMO-EM, part of this increase in the TM-score of the full-length model are also result from **an improvement in the quality of individual domain structures**.

# Structures Assembly from Synthesized Density Maps

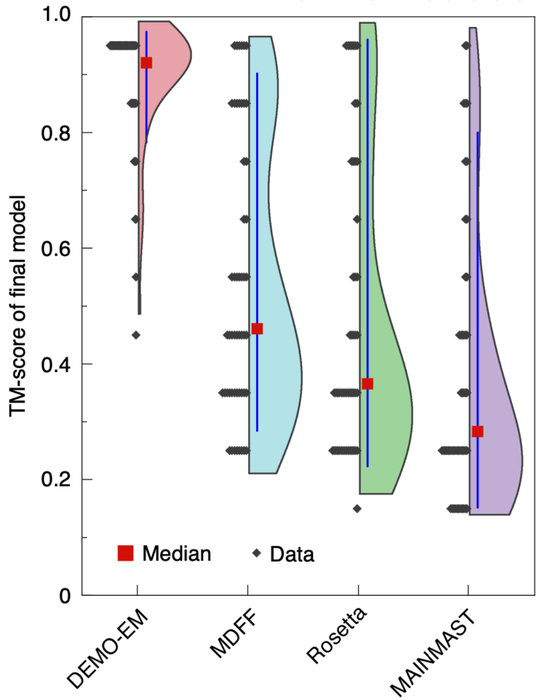



- The performance of **DEMO-EM** is not significantly affected by a reduction in the **resolution** of the density maps.
- While the performance of **MDFF** and **Rosetta** decreased slightly when using lower-resolution maps, the average TM-score of **DEMO-EM** remained significantly higher than those of **MDFF** and **Rosetta**.

# Structures Assembly from Experimental Density Maps

**From 51 cases**



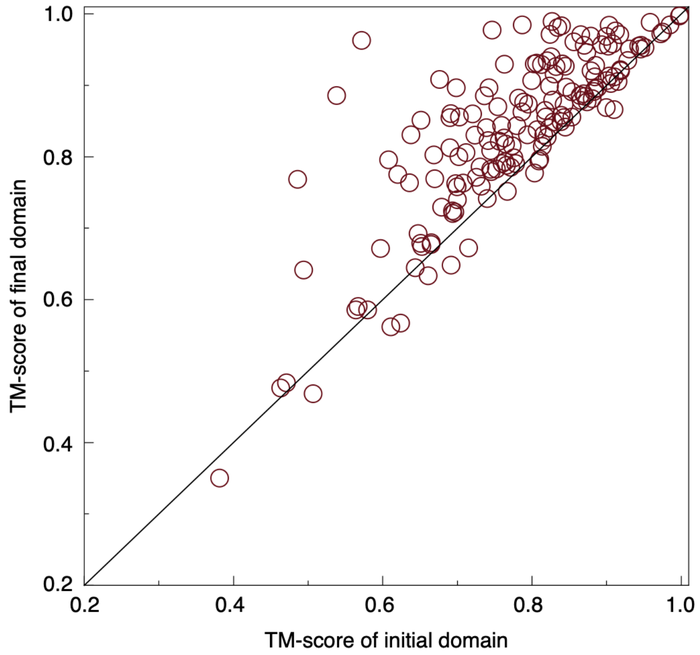
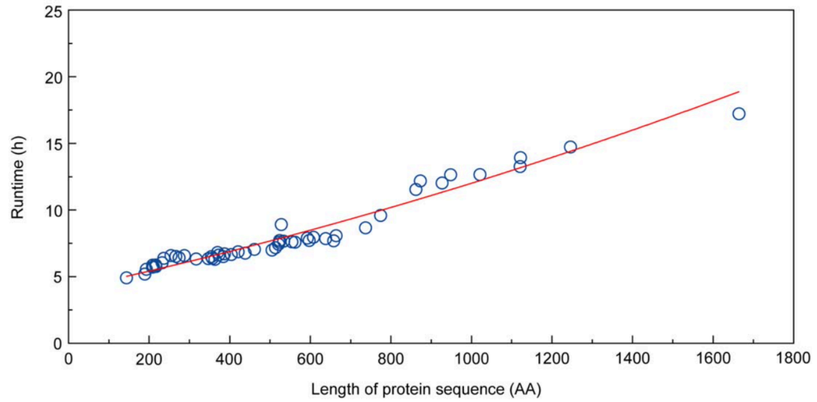




**Table 2 | Results for 51 proteins with experimental cryo-EM density maps**

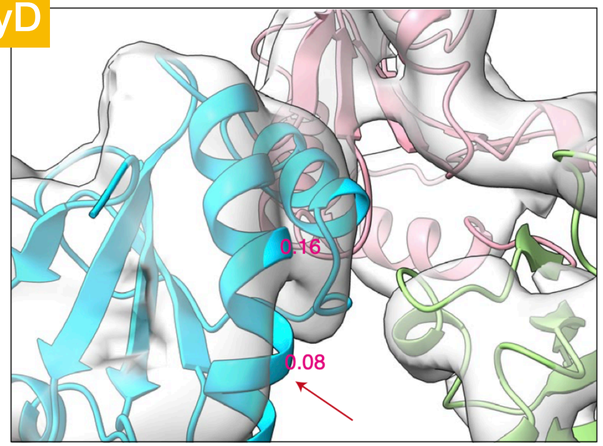
| | MDFF | Rosetta | MAINMAST | DEMO-EM |
|---|---|---|---|---|
| TM-score | 0.55 (0.24) | 0.47 (0.28) | 0.36 (0.24) | **0.88** (0.09) |
| RMSD (Å) | 17.7 (11.0) | 24.1 (16.6) | 23.0 (10.3) | **4.2** (3.2) |
| TM-score (domain)[a] | 0.56 (0.21) | 0.47 (0.26) | 0.42 (0.32) | **0.84** (0.13) |
| RMSD (domain) (Å)[b] | 7.2 (4.2) | 11.4 (10.0) | 12.0 (8.1) | **3.2** (2.8) |

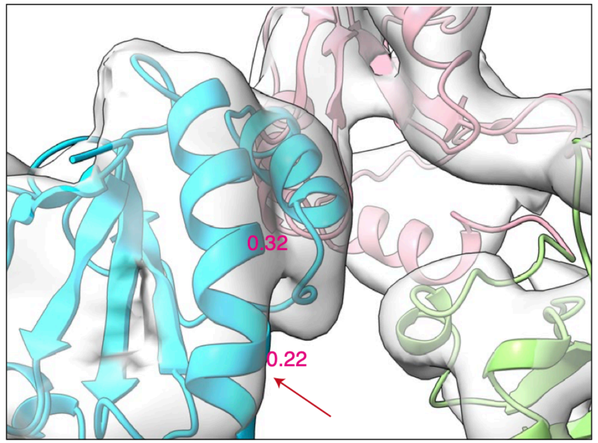



- After the DEMO-EM assembly, the full-length models achieved correct global fold (TM-score >0.5) in 98.0% of the cases.
- The average TM-score of individual domains in the full-length models was increased from 0.78 to 0.84, with 90.2% of the domains being improved
- The average runtime required by the whole pipeline of DEMO-EM for all 51 test proteins is 8.15h.

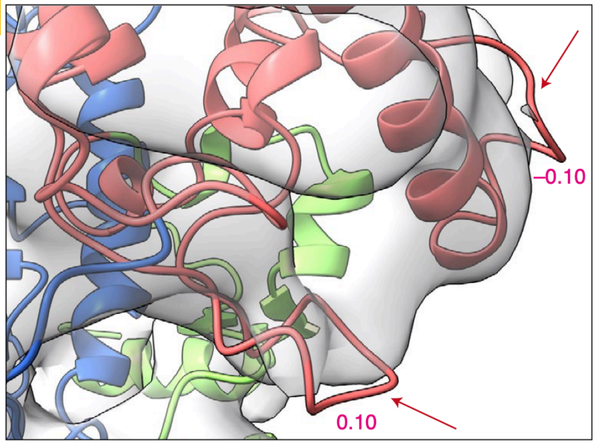
# Structures Assembly from Experimental Density Maps

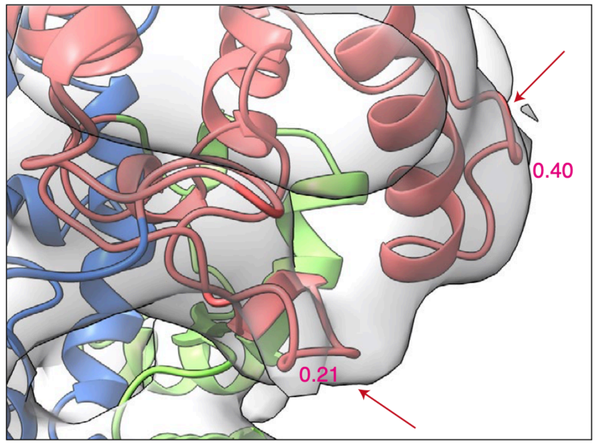
6enyD



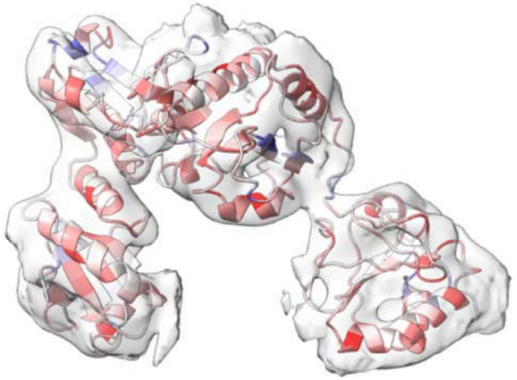

5fj6



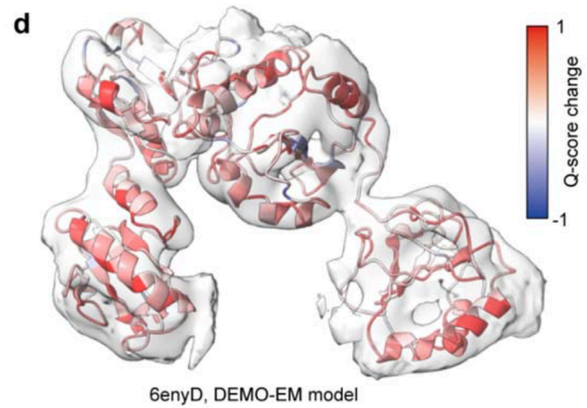




**Deposited Model from EMDB**

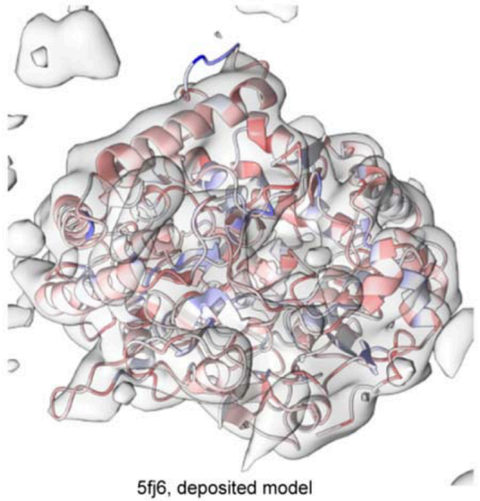
**DEMO-EM Model**



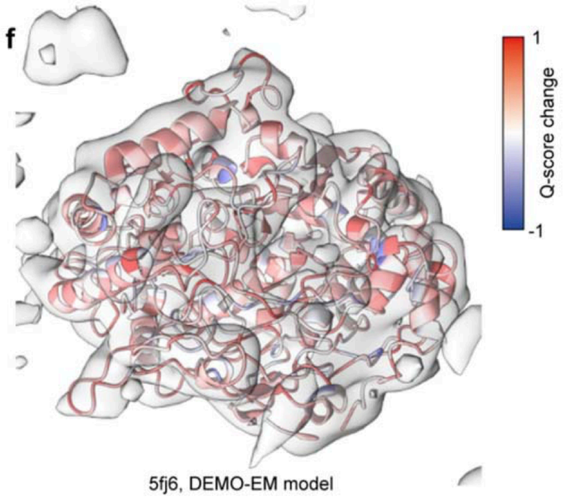
6enyD, deposited model




6enyD, DEMO-EM model



5fj6, deposited model




5fj6, DEMO-EM model

**Deposited Model from EMDB**

**DEMO-EM Model**

# Structures Assembly from Experimental Density Maps

**From 51 cases**

| Methods | Quality for full-length models | | Quality for domain-level models | |
|---|---|---|---|---|
| | TM-score | RMSD (Å) | TM-score | RMSD (Å) |
| AlphaFold2 | 0.84(0.16) | 4.4(4.2) | 0.89(0.09) | 2.0(1.3) |
| DEMO-EM | 0.88(0.09) | 4.2(3.2) | 0.84(0.13) | 3.2(2.8) |
| MDFF-AlphaFold2[a] | 0.89(0.14) | 3.6(4.5) | 0.86(0.12) | 2.4(1.3) |
| Rosetta-AlphaFold2[b] | 0.88(0.15) | 3.8(4.7) | 0.85(0.11) | 2.5(1.2) |
| DEMO-EM-AlphaFold2[c] | **0.93**(0.07) | **2.5**(1.7) | **0.90**(0.09) | **1.8**(1.1) |

- The individual domains predicted by **AlphaFold2** have a higher TM-score (0.89) than that of **DEMO-EM** (0.84), which is probably because of the lower quality of the domain models built by **D-I-TASSER**.
- The quality of the overall full-length models built by **DEMO-EM** (0.88) is better than that achieved by AlphaFold2 (0.84).
- If fed the same full-length models constructed by **AlphaFold2** into **MDFF**, **Rosetta** and **DEMO-EM** to examine the performance of the flexible assembly and refinement process, all the methods improved the initial full-length model, showing the usefulness of cryo-EM data even for the best-predicted models.

- REMC: (Replica-Exchange Monte Carlo) a computer simulation method typically used to find the lowest energy state of a system of many interacting particles.
- DomainDist : predict inter-domain distance maps.
- ResPRE is an algorithm for protein residue-residue contact-map prediction.
- ThreaDom is a template-based algorithm for protein domain boundary prediction.
- FUpred is a contact map-based domain prediction method which utilizes a recursion strategy to detect domain boundary based on predicted contact-map and secondary structure information.
- I-TASSER is a hierarchical approach to protein structure prediction and structure-based function annotation.
- D-I-TASSER : generates an initial structural model for each domain using distance-guided iterative threading assembly refinement.
- SPICKER is a clustering algorithm to identify the near-native models from a pool of protein structure decoys.
- FG-MD is a molecular dynamics (MD) based algorithm for atomic-level protein structure refinement.
- FASPR is a method for structural modeling of protein side-chain conformations.

# Flowchart of DEMO-EM