Supplementary Materials

Deep-learning with synthetic data enables automated picking of cryo-EM particle images of biological macromolecules

Ruijie Yao[1,†], Jiaqiang Qian[1,†] and Qiang Huang[1,2,*]

[1]State Key Laboratory of Genetic Engineering, MOE Engineering Research Center of Gene Technology, School of Life Sciences, Fudan University, Shanghai 200438, [2]Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 201203, China

*To whom correspondence should be addressed (Email: huangqiang@fudan.edu.cn).

[†]These two authors contributed equally.

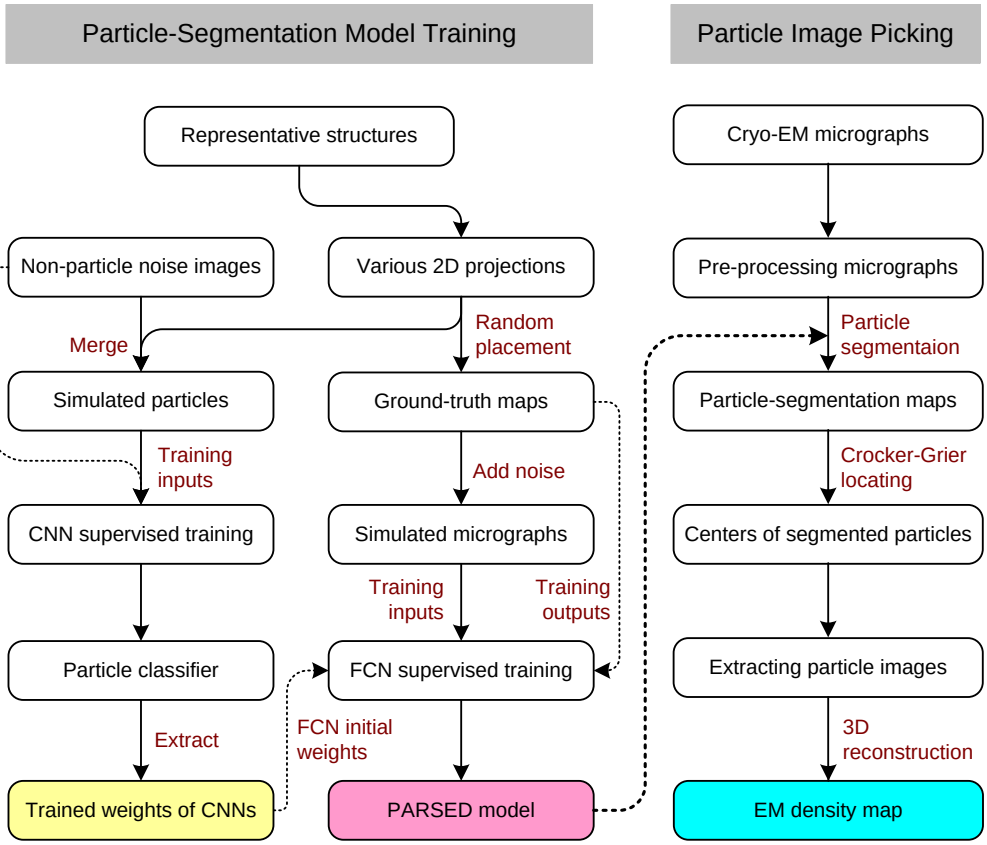

Fig. S1. The workflow for supervised training and application of our deep-learning model. Details about the workflow are described in Methods. In the phase of particle-segmentation model training, a CNN particle/non-particle classifier was firstly trained by the synthetic dataset of particles/non-particles; then, the hyperparameters of the trained CNNs were used as the initial weights of the particle-segmentation model for the supervised training with the synthetic dataset of simulated micrographs/ground-truth particle-segmentation maps. The trained model was designated as PARSED (**PAR**ticle **SE**gmentation **D**etector). In the phase of particle image picking, images of the picked particles were eventually fed into an *ab initio* 3D reconstruction program such as cryoSPARC (Punjani *et al.* 2017)

Note S1: Representative dataset of known 3D structures

To generate simulated micrographs with known 3D structures, macromolecular structures in the Protein Data Bank (PDB) were screened according to these criteria: (i) protein and protein-nucleic acid complex structures with a molecular weight > 200 kDa; (ii) structures determined by electron microscopy and with a resolution < 5.0 Å; (3) pairwise protein sequence identities < 50%. The final selected structures are listed in Supplementary Table S1.

Table S1. PDB codes of representative 3D structures for generating synthetic training datasets

2XD8	3IYJ	3IYL	3J26	3J31	3J32	3J3X	3J40	3J4U	3J6B
3J7O	3J7P	3J7Q	3J7R	3J7V	3J7Y	3J7Z	3J80	3J81	3J92
3J94	3J9B	3J9K	3J9M	3J9O	3J9P	3J9W	3J9X	3J9Y	3J9Z
3JA1	3JA7	3JAC	3JAD	3JAG	3JAH	3JAI	3JAJ	3JAM	3JAN
3JAP	3JAV	3JB5	3JB6	3JB7	3JB9	3JBM	3JBN	3JBU	3JBV
3JBW	3JC1	3JCD	3JCE	3JCF	3JCJ	3JCL	3JCM	3JCN	3JCS
3JCT	4CE4	4UG0	4UQ8	4UY8	4V19	4V1A	4V1W	4V8Y	4V91
4V92	5A22	5A2Q	5A8L	5A9Z	5ADY	5AFI	5AJ0	5AJ3	5AJ4
5AN9	5ANB	5ANC	5APN	5APO	5BK4	5FJ8	5FLM	5FLU	5FLX
5FTJ	5FYW	5G06	5G2X	5GAD	5GAE	5GAF	5GAG	5GAH	5GAK
5GAM	5GAN	5GAO	5GAP	5GAQ	5GJV	5GM6	5GMK	5GO9	5GW4
5H0R	5H0S	5H1Q	5H1S	5H3O	5H4P	5H5U	5HI9	5I08	5IMQ
5IPI	5IQR	5IRZ	5IT7	5IT9	5IYB	5IYC	5IYD	5JTE	5JU8
5JUL	5JUO	5JUP	5JUS	5JUT	5JUU	5K12	5KCR	5KCS	5KGF
5KPS	5KPV	5KPW	5KPX	5KUF	5KYH	5L35	5L3P	5LCW	5LD2
5LEG	5LI0	5LJ3	5LJV	5LKH	5LKS	5LL6	5LMN	5LMO	5LMQ
5LMR	5LMT	5LMU	5LMV	5LWG	5LZA	5LZC	5LZD	5LZE	5LZF
5LZP	5LZS	5LZT	5LZU	5LZV	5LZW	5LZX	5LZY	5LZZ	5M0Q
5M1J	5M3F	5M3L	5M5W	5M5X	5M5Y	5M64	5MBV	5MC6	5MDV
5MDW	5MDY	5MDZ	5MGP	5MKE	5MLC	5MMI	5MMJ	5MMM	5MPS
5MQ0	5MRC	5MRE	5MRF	5MUU	5MW1	5N61	5N6W	5N8N	5N8Y
5NCO	5ND1	5ND8	5ND9	5NGM	5NJT	5NP6	5NP7	5NSR	5NUG
5NV3	5NWY	5O09	5O2R	5O4U	5O5J	5O60	5O61	5O9G	5O9Z
5OA1	5OA3	5OAC	5OAF	5ODV	5OF4	5OFO	5OIK	5OJQ	5OJS
5OOL	5OOM	5OPT	5OQL	5OSG	5SZS	5T2A	5T2C	5T4D	5T5H
5T62	5T6R	5T7V	5TB2	5TC1	5TCP	5TCQ	5TCU	5TFY	5U07
5U0A	5U0P	5U1C	5U4I	5U4J	5U8T	5U9F	5U9G	5UDB	5UMD
5US7	5UU5	5UVN	5UYK	5UYL	5UYM	5UYN	5UYP	5UYQ	5UZ5
5UZ9	5V4S	5V7Q	5V8F	5V93	5VC7	5VF3	5VKQ	5VKU	5VOT
5VT0	5VY8	5VZL	5W0S	5W1R	5W5E	5W5F	5W5Y	5W64	5W65
5W66	5W68	5W7G	5WC0	5WC3	5WEK	5WEO	5WFE	5WJT	5WK1
5WK5	5WLC	5WLN	5WQ7	5WQ8	5WSG	5WSN	5WVE	5WYK	5X0X
5X0Y	5X6O	5X8P	5X8R	5X8T	5XJC	5XJY	5XLO	5XLR	5XMI
5XON	5XSY	5XTC	5XTE	5XXB	5XXU	5XY3	5XYI	5XYM	5XYU
5YFP	5YZ0	5Z10	6ALF	6ALG	6ALH	6AP1	6AZ0	6AZ1	6AZ3
6B43	6B44	6B45	6B46	6B47	6B48	6B6H	6B8H	6B9Q	6BCU
6BE1	6BFU	6BGI	6BJC	6BLL	6BLY	6BO8	6BPQ	6BPZ	6BQR
6BU8	6BUZ	6BWX	6C0W	6C1D	6C26	6C3O	6EK0	6EK5	6ELZ
6EM1	6EM3	6EM4	6EM5	6EML	6ENF	6ENJ	6ENU	6EOJ	6EU0
6EU1	6EU2	6EXN	6EZJ	6EZM	6F0L	6F40	6F41	6F44	6FBS

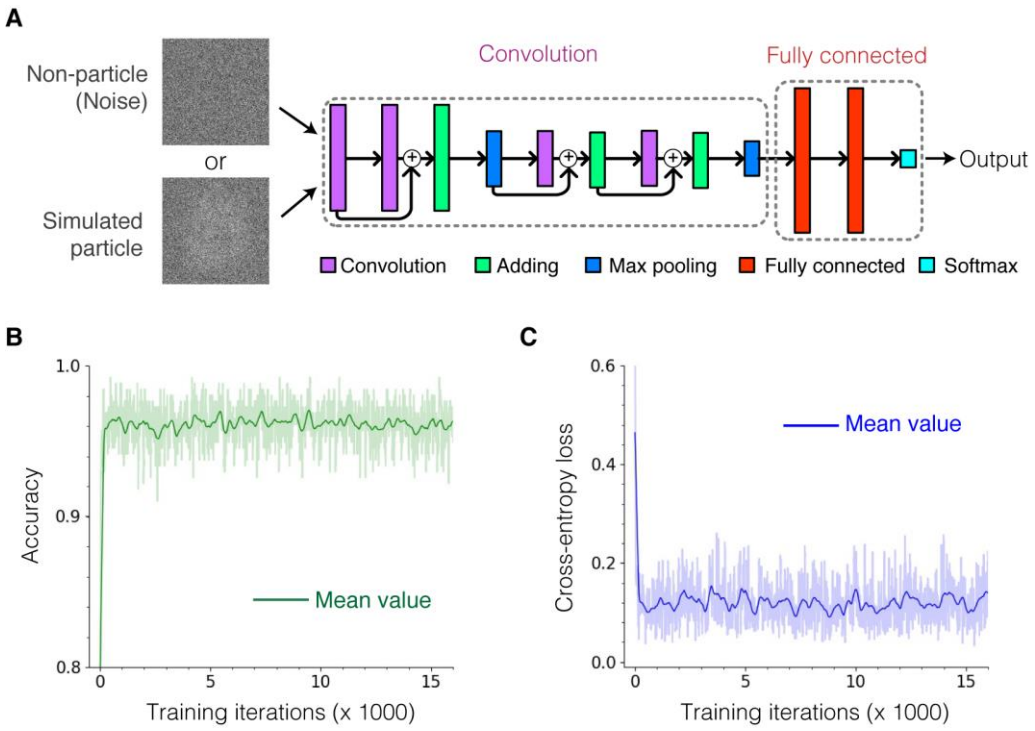

Fig. S2. Architecture and supervised training of binary particle classifier for discriminating particles and non-particles. (A) Architecture of the CNN model that classifies an input image into particle or non-particle (when outputting 1.0, particle; when outputting 0.0, non-particle). The CNN hyperparameters were used as the initial weights of the identical CNNs in PARSED (see also Supplementary Table S2). (B), (C) Accuracy and cross-entropy loss curves in the supervised learning of the CNN classifier trained with the synthetic dataset of the simulated particles/non-particles.

Table S2. Deep neural network layers of the binary particle classifier and the PARSED model

Names	Filter or node parameters	Upper layer(s)	Types	Layer distributions	
				Classifier	PARSED
Layers identical in both models					
input			input	√	√
conv1	32, (3, 3)	input	convolution	√	√
conv2_1	32, (3, 3)	conv1	convolution	√	√
conv2	-	conv1 + conv2_1	adding	√	√
pool1	(2, 2)	conv2	max pooling	√	√
conv3_1	32, (3, 3)	pool1	convolution	√	√
conv3	-	pool1 + conv3_1	adding	√	√
conv4_1	32, (3, 3)	conv3	convolution	√	√
conv4	-	conv3 + conv4_1	adding	√	√
pool2	(2, 2)	conv4	max pooling	√	√
Layers specific for CNN classifier					
ip1	128	pool2	inner product	√	
ip2	128	ip1	inner product	√	
prediction	Softmax	ip2	output	√	
Layers specific for PARSED					
dconv2d_1	32, (3, 3)	pool2	deconvolution		√
pool2_re	-	pool1 + deconv2d_1	adding		√
deconv2d_2	32, (3, 3)	pool2_re	deconvolution		√
conv2d_ip2	2, (32,32)	deconv2d_2	convolution		√
prediction	Softmax	conv2d_ip2	output		√

Note S2: Evaluation on synthetic cryo-EM data

The PARSED model was firstly evaluated on a synthetic dataset generated with the 3D structure of SpCas9-Cas9-DNA ternary complex (PDB: 5Y36). To the end, we used 5Y36 to generate 400 simulated micrographs and corresponding ground-truth particle-segmentation maps using the methods in *Simulated micrographs of a given 3D structure* (Methods)**.**

To assess the model performance, we compared the coordinates of the picked particles with those in the ground-truth maps, and then calculated true positive number (*TP*), false positive number (*FP*), false negative number (*FN*) and true negative number (*TN*) for the paticle-picking process. In the calculation, the criteria for a true positively picked particle was that the distance between the predicted location and the nearest ground-truth coordinates are less than 30% of the picking aperture diameter; and *TN* was defined as:

$$TN = \sum_k \frac{w_k h_k - \pi\varphi^2 \mathrm{N}_k}{\pi\varphi^2} \quad (1)$$

where *w* and *h* are the width and height of the *k*-th micrograph in the dataset, φ is the particle-picking aperture diameter, and N_k is the number of particles in the *k*-th micrograph. Then, usual statistical parameters, *recall* and *precision*, were calculated as:

$$recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (2)$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (3)$$

Finally, we calculated two evaluation parameters, *Accuracy* and *F1-score*, for the picking process as:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}} \quad (4)$$

$$F1\ score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \quad (5)$$

Because these two parameters depend on the mass selection threshold for selecting segmented blobs, we also generated *Accuracy* and *F1-score* curves with respect to the selection threshold, in order to determine the 'best' threshold for the picking process to get the maximum *Accuracy* and *F1-score* (e.g., Figure 2B).

Note S3: Normalization of raw micrographs

To simplify model training and particle-picking processes, all raw micrographs were firstly calculated with the following normalization method:

$$image = \frac{image - \overline{image_{\text{center}}}}{\sigma(image_{\text{center}})}$$

where *image* is the 2D array of a raw micrograph image, $\overline{image_{\text{center}}}$ is the average contrast value in the central area (1/4 sizes of both width and height) of the given micrograph, and σ is the standard deviation. To further reduce noise level, a high pass filter (500 Å) and then a low pass filter (30 Å) were applied to each calculated micrograph. Next, a down-sampling process at the Nyquist frequency (30/2 = 15 Å) (Mitra and Kuo 2006) was carried out for the filtered micrographs to output the final normalized micrographs. Examples of the normalized micrographs are given in Supplementary Figure S3.

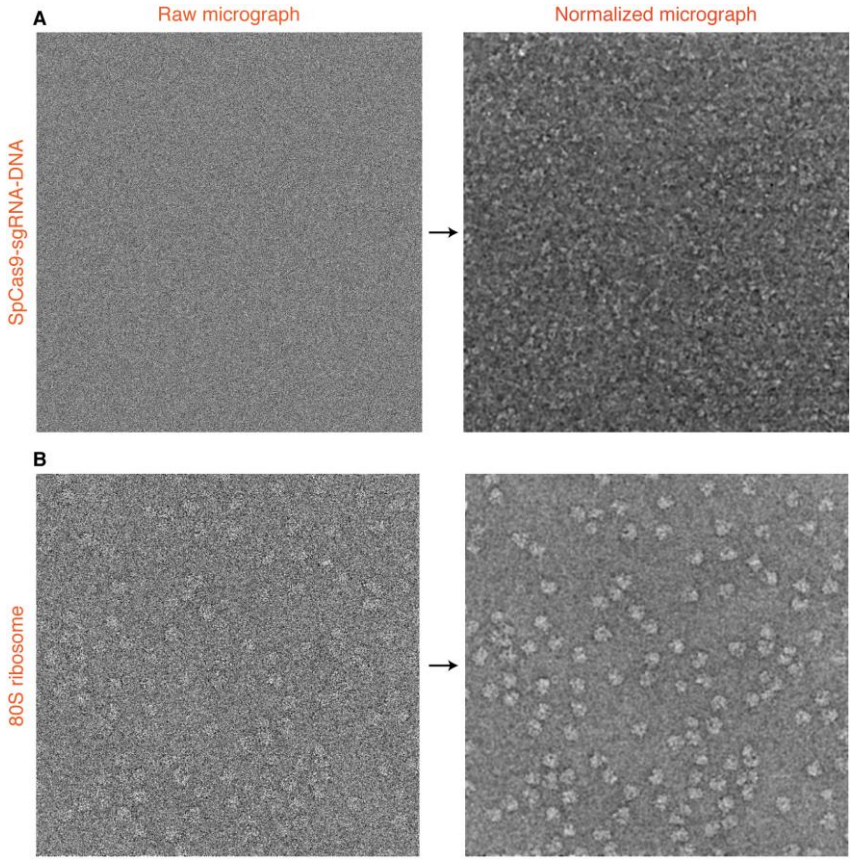

Fig. S3. Normalization of raw cryo-EM micrographs. (A) Example of a normalized micrograph of SpCas9-sgRNA-DNA ternary complex from our previous study (Huai *et al.* 2017) (see also Figure 3). (B) Example of a normalized micrograph of *Plasmodium falciparum* 80S ribosome from the dataset EMPIAR-10028. For clarity, the contrast values of the given micrographs were rescaled from 0 to 255.

Note S4: Validation on experimental cryo-EM data

Six large cryo-EM datasets were downloaded from EMPIAR to evaluate the universal particle-picking capability of the PARSED model, as listed in Supplementary Table S3. All potential particles were automatically segmented and picked by the present method, and then fed into *ab initio* structure determination program *cryoSPARC* (Punjani *et al.* 2017) to build the 3D density maps. Multiple rounds of iterative 2D classifications were carried out to select those 'good' particles for final 3D reconstructions. Using those particles selected by the 2D classifications, 3D density maps were built with standard procedure in *cryoSPARC*, and their global resolutions were calculated with the Fourier shell correlation (FSC) 0.143 cut-off criterion.

Table S3. Cryo-EM datasets from EMPIAR for validating PARSED

EMPIAR ID	Molecular names	Molecular types
10005	Transient receptor potential channel 1 (TRPV1)	Ion channel
10012	β-galactosidase proteasome	Enzyme
10028	*Plasmodium falciparum* 80S ribosome	RNA-protein-ligand complex
10049	Synaptic RAG1-RAG2 complex	DNA-protein complex
10058	*Thermoplasma acidophilum* 20S proteasome	Proteasome
10081	Human HCN1 hyperpolarization-activated channel	Ion channel

Table S4. Missing micrographs in EMPIAR-10005

IDs of the missing sum-corrected micrographs									
0105	0227	0318	0399	0461	0496	0596	0657	0756	0949
0164	0263	0320	0405	0465	0521	0606	0659	0768	0951
0166	0270	0323	0408	0467	0523	0612	0667	0796	0953
0179	0275	0339	0412	0474	0534	0623	0670	0829	0961
0188	0288	0349	0417	0476	0539	0630	0683	0831	0964
0197	0293	0353	0419	0479	0555	0634	0686	0843	
0200	0299	0357	0423	0483	0562	0637	0698	0857	
0203	0304	0362	0427	0485	0569	0641	0703	0863	
0210	0311	0375	0441	0488	0575	0649	0719	0903	
0222	0315	0377	0457	0491	0587	0654	0726	0935	

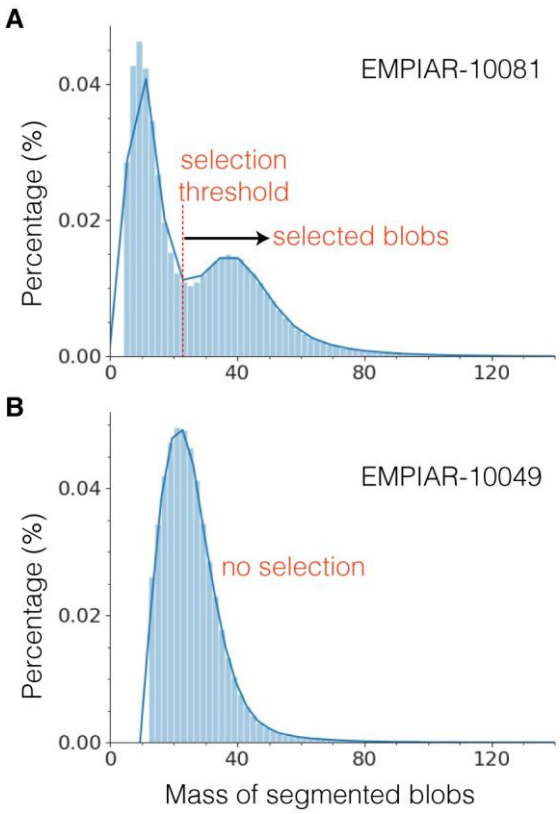

Fig. S4. Typical mass distributions of segmented blobs (particle candidates) determined by the Crocker-Grier algorithm. (A) Example of a two-peak distribution with a selection threshold at mass of ~21 in the dataset EMPIAR-10081. In this case, only those segmented blobs with mass values greater than the selection threshold are chosen as the final picked particles. (B) Example of a one-peak distribution without any selection in the dataset EMPIAR-10049. In this case, all segmented blobs are selected as the final picked particles for further investigations.

Table S5. Particle-picking results and 3D reconstructions for 6 large public cryo-EM datasets

EMPIAR ID	10005	10012	10028	10049	10058	10081
Magnified pixel size **(**Å/pixel**)**	1.2156	1.275	1.34	1.2156	1.35	1.3
Picking aperture diameter (Å)	128	128	160	128	140	128
Numbers of picked particles	268,042	191,548	178,275	260,753	105,723	180,166
Particles for 3D reconstruction	32,949	91,700	121,534	64,150	36,688	57,763
Resolution at FSC = 0.143 (Å)	3.4	3.0	3.4	3.5	3.1	3.7
EMDB maps for fitting (ID)	5778	5995	2660	6487	3348	8511
Map correlation coefficients	0.98	0.98	0.93	0.93	0.95	0.96
Atomic models for fitting (PDB IDs)	3J5P	3J7H	3J7A 3J79	3JBX	3J9I	5U6O

Table S6. Desktop computer for evaluating PARSED processing time

Hardware/Software	Description
CPU	2 × Intel Xeon E5-2683 v3
Memory	64 GB 2133 MHz ECC DDR4
GPU	2 × nVidia GeForce GTX 1080 Ti
Storage	4 TB, 7200 rpm HDD
Operating system	Linux (Canonical Ubuntu 16.04 LTS)
Cryo-EM software	RELION, XMIPP, SCIPION, cryoSPARC
Other supporting programs	TensorFlow, Keras, NumPy, OpenCV, trackpy, mrcfile

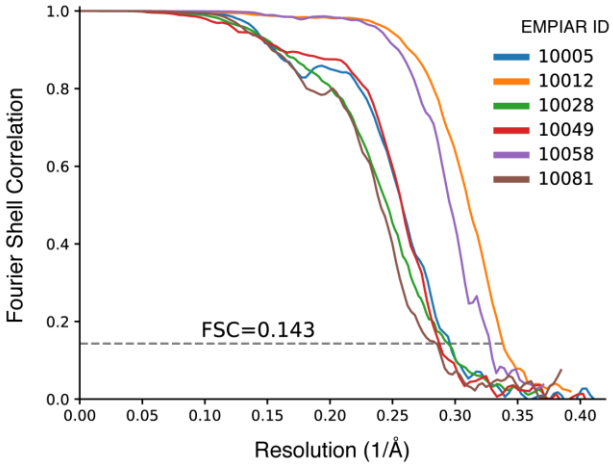

Fig. S5. Gold-standard FSC curves of EM maps for 6 cryo-EM datasets from EMPIAR. Those are FSC curves between two independently refined half-maps. All the FSC resolution curves were plotted with *cryoSPARC* (Punjani *et al.* 2017).

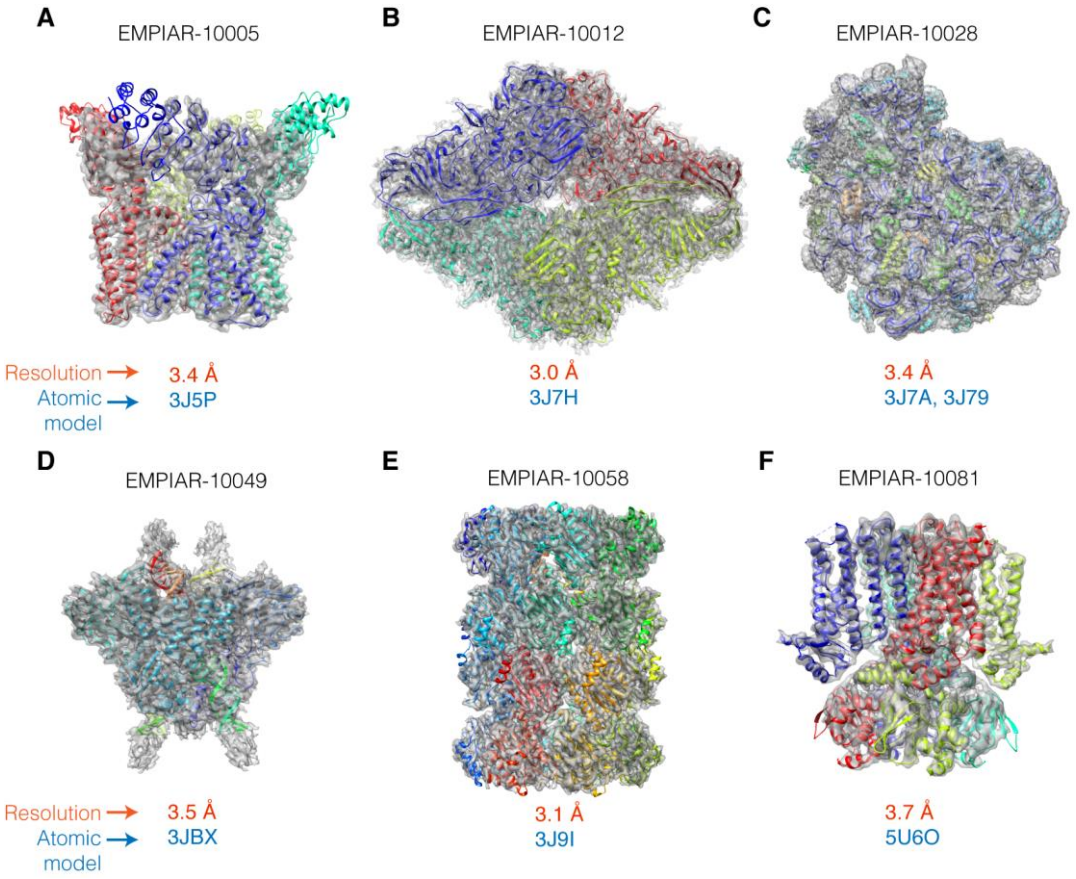

Fig. S6. Atomic model fitting into the EM density maps of 6 cryo-EM datasets from EMPIAR. (A) Transient receptor potential channel 1 (TRPV1) and atomic model (PDB: 3J5P). (B) β-galactosidase and atomic model (PDB: 3J7H). (C) *Plasmodium falciparum* 80S ribosome bound to the anti-protozoan drug emetine and atomic model (PDB: 3J7A, 3J79). (D) Synaptic RAG1-RAG2 complex and atomic model (PDB: 3JBX). (E) *Thermoplasma acidophilum* 20S proteasome and atomic model (PDB: 3J9I). (F) Human HCN1 hyperpolarization-activated channel and atomic model (PDB: 5U6O). All the EM density maps were generated with *cryoSPARC* (Punjani *et al.* 2017), and atomic model fitting was carried out with *Coot* (Emsley and Cowtan 2004).

Table S7. Command lines and parameters for the particle picking on the dataset EMPIAR-10017

Method	Descriptions

PARSED	Docker version (using -it for interactive prompt)
	`$ python3 -W ignore parsed_main.py --model=./pre_train_model.h5 --data_path=./Micrographs --output_path=./output --file_pattern=*.mrc --angpixel=1.77 --img_size=4096 --edge_cut=135 --job_suffix=autopick --core_num=100 -- aperture=100 --mass_min=4 --gpu_id=1`
	`$ python3 particle_mass.py drawmass --pick_output=./output --job_suffix=autopick --tmp_hist=tmp_hist`
	`$ python3 particle_mass.py cutoff --pick_output=./output --job_suffix=autopick --output_suffix=checked --thres=10.3`

RELION LoG picker	RELION-gpu 2.1 `$ mpirun -np 100 `which relion_autopick_mpi` --i CtfFind/job002/micrographs_ctf.star --odir AutoPick/job008/ --pickname autopick --LoG --LoG_diam_min 200 --LoG_diam_max 250 --shrink 0 --lowpass 20 --LoG_adjust_threshold 0`

APPLE	applepicker-python
	`$ python3 run.py -s 78 ./Micrographs -o ./applepicker`
	`In apple/config.py`
	`proc = 24`

DeepPicker[a]	DeepPicker-python (Docker version, using -it for interactive prompt)
	The first-round picking based on the pre-trained model
	`$ python autoPick.py --inputDir ./Micrographs --pre_trained_model ../trained_model/model_demo_type3 --particle_size 128 --mrc_number 84 --outputDir ./deeppicker --coordinate_symbol _deeppick --threshold 0.2`
	Parameters in RELION 2D classification for creating a training template set
	`--pad 2 --ctf --iter 25 --tau2_fudge 2 --particle_diameter 160 --K 10 --flatten_solvent --zero_mask --oversampling 1 --psi_step 11.25 --offset_range 5 --offset_step 2 --norm –scale`
	Additional training for the pre-trained model with selected template particles
	`$ python train.py --train_type 4 --train_inputFile ./sel_particles.star --particle_size 180 --particle_number -1 --model_save_dir './trained_model' --model_save_file model_demo_type3_2D`
	The second-round picking with the trained model
	`$ python autoPick.py --inputDir ./Micrographs --pre_trained_model ./trained_model/model_demo_type3_2D --particle_size 180 --mrc_number 84 --outputDir ./deeppicker-r2 --coordinate_symbol _deeppick --threshold 0.5`

[a] We failed to build dependencies from https://github.com/nejyeah/DeepPicker-python, due to obsolete versions of tensorflow (0.1 < 1.0) and ubuntu (14.04 < 16.04). Instead, a third-party docker version from https://github.com/thorstenwagner/docker-deeppicker was used.

Note S5: Data availability

Examples of the simulated micrographs with 5Y36, coordinate files of the picked particles by PARSED and corresponding calculated EM density maps of the 6 cryo-EM datasets are available from Dropbox at https://www.dropbox.com/sh/lzrvzv9bzqhsiws/AADQyK5wShnUjjdqGWqTzG8aa?dl=0.

References

Emsley, P. and Cowtan, K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D*, 60, 2126–2132.

Huai, C., *et al.* (2017) Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.*, 8, 1375.

Mitra, S.K. and Kuo, Y. Digital Signal Processing: A Computer-based Approach. McGraw-Hill Higher Education New York; 2006.

Punjani, A., *et al.* (2017) cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods*, 14, 290–296.