RE2DC: a robust and efficient 2D classifier with visualization for processing massive and heterogeneous cryo-EM data

Szu-Chi Chung[a,b], Hsin-Hung Lin[c], Kuen-Phon Wu[d], Ting-Li Chen[a],
Wei-Hau Chang[c,*], and I-Ping Tu[a,*]

[a]Institute of Statistical Science, Academia Sinica, Taiwan
[b]Department of Applied Mathematics, National Sun Yat-sen University, Taiwan
[c]Institute of Chemistry, Academia Sinica, Taiwan
[d]Institute of Biological Chemistry, Academia Sinica, Taiwan

*Correspondence: weihau@gate.sinica.edu.tw (W.H.C.), iping@stat.sinica.edu.tw (I.P.T.).

November 25, 2022

Abstract

Despite the fact that single particle cryo-EM has become a powerful method of structural biology, processing cryo-EM images are challenging due to the low SNR, high-dimension and un-label nature of the data. Selecting the best subset of particle images relies on 2D classification—a process that involves iterative image alignment and clustering. This process, however, represents a major time sink, particularly when the data is massive or overly heterogeneous. Popular approaches to this process often trade its robustness for efficiency. Here, we introduced a new unsupervised 2D classification method termed RE2DC. It is built upon a highly efficient variant of γ-SUP, a robust statistical cryo-EM clustering algorithm resistant to the attractor effect. To develop this efficient variant, we employed a tree-based approximation to reduce the computation complexity from $O(N^2)$ to $O(N)$, with N as the number of images. In addition, we exploited t-SNE visualization to unveil the process of 2D classification. Our tests of RE2DC using various datasets demonstrate it is both robust and efficient, with the potential to reveal subtle structural intermediates. Using RE2DC to curate a dataset of sub-millions of COVID-19 spike particles picked from 3,511 movies only takes 8 hours, suggesting its capability of accelerating cryo-EM structural determination. Currently, RE2DC is available with both CPU and GPU versions, where the implementation only requires modest hardware resources.

Introduction

Since the resolution revolution of cryogenic electron microscopy (cryo-EM) was ignited by advancement in direct detectors [1, 2, 3, 4] and improved image analysis algorithms [5, 6, 7], this method has transformed the landscape of structural biology. The approach of single particle analysis (SPA) [8] allows structural determination without the need for crystals, yet produces high-resolution density maps (Coulomb potential maps) comparable to those from X-ray crystals [9, 10, 11] to support de novo atomic model building for proteins and protein-nucleic acid complexes. As SPA cryo-EM records images from structures co-existing in solution [12], it permits capturing the structural intermediates for understanding molecular mechanisms [13]. The incredible power of cryo-EM is highlighted by the timely delivery of the spike protein structure of SARS-CoV-2 at its emergence [14], followed by continually elucidating the structural variants from rapidly evolved strains during this global pandemic [15, 16, 17], all together revealing the mechanisms underlying infection and antibody escape. Nonetheless, because cryo-EM images are exceedingly noisy as they are recorded with low electron dose (10-40 $e^-/\text{Å}^2$) due to biological molecules being sensitivity to high-energy electrons, 3D structural determination from cryo-EM images of biological molecules represents a challenging inverse problem, for which developing image analysis software with improved performance to streamline the processing [18, 19, 20, 21, 22, 23, 24, 25] continues to be a vibrant activity in the cryo-EM community.

Computing a 3D structure from single particle images of low signal-to-noise ratio (SNR) with unknown orientations is a challenging task. This task is achieved by a workflow that consists of multiple steps (Figure 1) [7, 18, 19, 20, 21, 22, 23, 24, 25]. A vital step in the workflow is to remove bad particles and non-particle contaminants/artifacts [26]. As the SNR of the particles is low, this step of particle cleaning relies on 2D classification in an unsupervised manner [8] (see Figure 1), a process that clusters particles of similar view into a class to give an average image of high SNR with identifiable features [27]. Currently, 2D classification may represent a major time-sink in the workflow for massive data, which is exacerbated by the high degree of heterogeneity. Prior to particle clustering, particle images must be aligned with each other to minimize the variations due to their differences in translation and in-plane rotation. Unfortunately, aligning particle images of low SNR using cross-correlation between images does not guarantee that the highest score gives the correct solution for alignment or finds the most resembling partner. To handle this issue, a maximum likelihood approach was introduced [28, 29, 30] and then adopted by popular cryo-EM packages [7, 21, 22]. This approach models each experimental image as derived from one of K representative images via in-plane rotation, translation, corrupted by microscope contrast transfer function (CTF) and noise. As a result, the 2D classification becomes

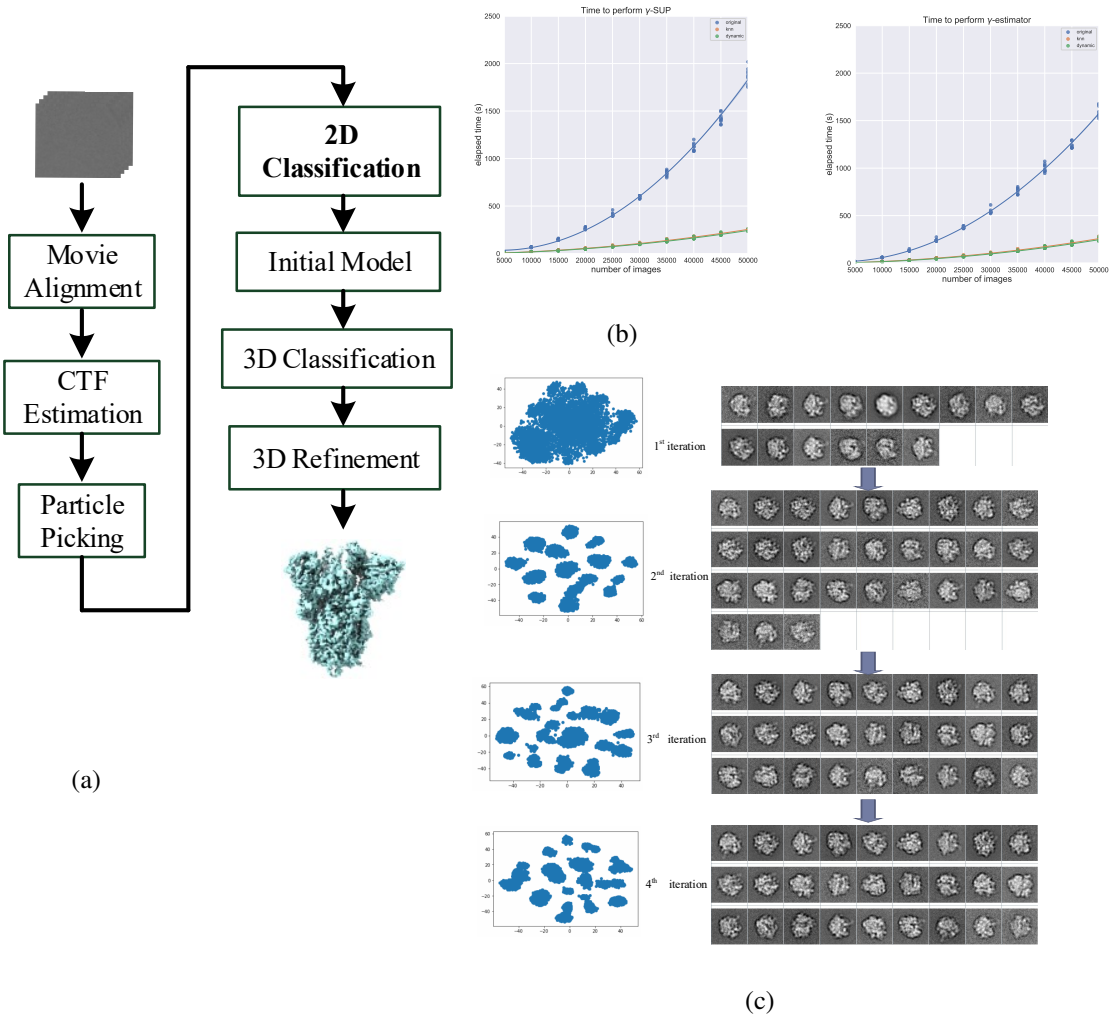


Figure 1: A workflow for processing single particle cryo-EM images. (a) Individual steps in the workflow from movie alignment (motion correction) to a final density map, as exemplified by a COVID-19 spike protein. (b) Dependence of the clustering time on the number of images, original γ-SUP versus new γ-SUP (knn γ-SUP or dynamical knn γ-SUP). (c) t-SNE plot for displaying particle distribution in high-dimension space in each iteration of a 2D classification, here exemplified by RE2DC on 70S ribosome.

a computationally demanding process that involves the tasks of enormous image alignment and solving optimization problem using expectation-maximization [30]. Although this likelihood-based approach has been proven powerful for solving high-resolution structures for difficult proteins due to extremely low SNR [31], its popularity was hindered by the need of accessing a large number of CPU cores, which was later alleviated by the advent of GPU-enabled parallelism [20]. Even so, using likelihood-based 2D classification accelerated by GPU to curate a dataset that contains sub-millions of particles may still take several days [32].

In addition to particle cleaning, 2D classification serves other functions, including exploring the heterogeneity of a dataset—this heterogeneity stems from particle orientations, conforma-

tions, compositions, and their combinations. It is well-known that any likelihood approach inevitably suffers from the attractor effect of "rich getting richer and poor getting poorer" [30, 33] due to experimental images producing higher correlation scores with those class averages of large-sized classes that exhibit high SNR. As a result, applying a likelihood approach to highly heterogeneous datasets might not give a satisfactory outcome. To date, a better 2D classification method for heterogeneous data is represented by "iterative and stable alignment and clustering" (ISAC) [34], which is now integrated into the SPHIRE package [35], and recently tested to be compatible with other packages [36] with the aid of the SCIPION [19]. ISAC circumvents the above-mentioned class imbalance issue rooted in K-means clustering by "equalizing" the size of each class, and thereby more sensitive to data heterogeneity, and thereby more sensitive to data heterogeneity. Although ISAC uses cross-correlation scores to determine the best alignment, it manages to offer highly reliable outcome by eliminating impurities/outliers in each class using repeated reference-free alignment to validate each image member, which is extremely time-consuming. Thus, in this age of speed, ISAC is deemed best for ab initio or exploratory data analysis [37].

Here, we introduced RE2DC, a new 2D classification method that is robust and efficient. We chose γ–SUP clustering algorithm [38] to build RE2DC because it is proven to be highly robust against noise or outliers, with particular strengths in resisting the attractor effect rooted in K-means clustering [38]. However, the original γ–SUP is impractical for clustering a large number of images ($N > 10^4$) because its computation complexity for calculating the pairwise weight matrix scales quadratically in N, the number of data points (particle images). Another shortcoming of γ–SUP is that it requires human intervention for choosing a good hyper-parameter, which is critical for determining the number of clusters. In the present work, we eliminated those drawbacks by major algorithmic development with novel solutions. In brief, we constructed a sparse approximation for the weight matrix of γ–SUP by employing the well-known k-d-tree method (see Methods). This has decreased the computation complexity for calculating the pairwise weight matrix from $O(N^2)$ to $O(N\log(N))$ [39] or even to $O(N)$. Next, we devised a data-driven approach for automating the selection of the best hyper-parameter. As for image alignment, we adopted multi-reference-alignment (MRA) [27, 31] with the introduction of a rapid dimension reduction (DR) termed 2SDR [40, 41] to de-noise cryo-EM images prior to alignment. This de-noising approach has greatly improved the accuracy of noisy image alignment, which in turn leads to a significant reduction of the number of iterations required for current 2D classification methods [36]. As MRA can be enabled by GPU iterations required [42], we compared the time on RE2DC with and without GPU. Finally, as the main purpose of 2D classification is to curate

a particle set, it is unnecessary to merge all images of similar views into one class. We then relaxed this constraint imposed by current 2D class methods and introduced a distributed computation strategy. This strategy that divides a large dataset into several smaller datasets for separate classification in parallel further speeded up RE2DC. Notably, to unveil the black box of 2D classification, we sought to track particle clustering in high-dimension during 2D classification [43]. To do so, we here, for the first time, employed a popular embedding tool of t-SNE [44, 45] for visualizing the clusters in two dimensions since t-SNE embedding can preserve the relationships in high-dimension (Figure 1).

To test RE2DC, we used various experimental single-particle cryo-EM sets, including five representative datasets from the public domain and our own dataset of glutamine synthetase (GS) (see the review of [46]). By benchmarking against ISAC, we first demonstrated RE2DC is indeed robust. We then explored whether or not GS could exist in oligomeric configuration other than dodecamer—RE2DC uncovered a peculiar oligomer that was missed by the current methods. We further tested the efficiency of RE2DC using large datasets ($N > 10^5$ particles). For a curated dataset of 80S ribosome ($N \sim 10^5$) [EMPIAR-10028], RE2DC finished the task in approximately 2 hours, whereas RELION took a much longer time. Similar strengths of RE2DC were observed for two more challenging datasets—non-curated TRPV channels ($N \sim 0.8 \times 10^5$) [EMPIAR-10005] and TRPV channels in nanodisc ($N \sim 2 \times 10^5$) [EMPIAR-10059]. Finally, curating sub-millions of COVID-19 spike particles ($N \sim 4 \times 10^5$) we picked from 3,511 movies [EMPIAR-10514] [14] took RE2DC approximately 8 hours, demonstrating its processing capability for a massive dataset that can now be obtained in one day using high-throughput collection [11, 47, 48, 49]. By using SCIPION [19] to couple RE2DC with 3D reconstruction algorithm from other packages, we showed those particles from RE2DC produced a high-quality density map with resolution comparable to the original work [14] (both < 3.5Å). In summary, our findings demonstrate RE2DC is a robust and efficient 2D classifier with particular strength in uncovering subtle structural variations.

Results and Discussion

New γ-SUP variants reduce the issue of computation complexity associated with a large number of images

To build the clustering algorithm of RE2DC, we have introduced variants of γ-SUP. This new γ-SUP uses k-nearest neighbor (knn) to approximate the weight matrix. As shown in Figure 1(b), the computation time of this knn γ-SUP and its advanced derivative called dynamic γ-SUP scale

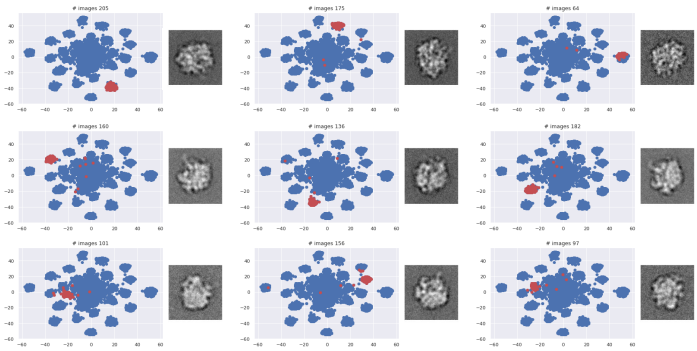
up linearly with the number of images, contrasted to the quadratic dependence of image number in the original γ-SUP.

Test on 70S ribosome particles highlights the robustness of RE2DC

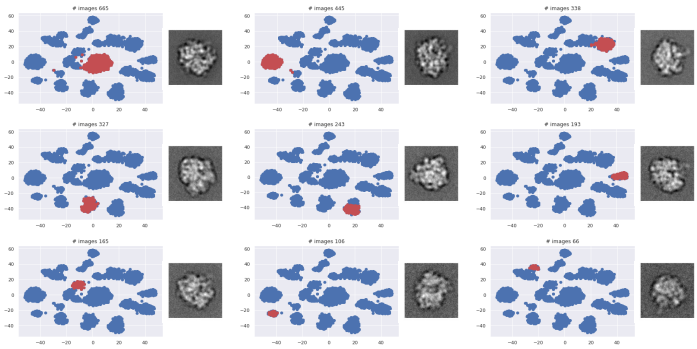
To test whether or not the new γ-SUP algorithm preserves the robustness of the original γ-SUP, we employed the same E. coli 70S ribosome dataset and compared its performance against that of ISAC [34] since it is a 2D classification approach well-known for its robustness. This set of 5,000 particles, windowed from CCD micrographs by Joachim Frank's lab prior to the cryo-EM revolution, was used by the early version of RELION for tutorial purposes. It is the first half of the 10,000 hand-picked particles composed of two half sets, each with a unique conformation. To unveil the clustering process, we employed t-SNE visualization [44, 45] to monitor inter-particle distances resulting from each iteration of 2D classification (Figure 1(c)). This visualization approach allows us to find that the clustering pattern of RE2DC has remained nearly unchanged after three iterations (Figure 1(c)). This notion that stable results are achieved with a minimal number of iterations is consistent with the fact that the number of harvested particles has reached a peak after three iterations. Noticeably, RE2DC generates 27 good classes for this 70S dataset (Figure 2(b) and Supplementary Figure 1), similar to that previously reported by the original γ-SUP [38].

For ISAC, it took many more iterations to converge. As summarized by Supplementary Figure 2, ISAC reports a total of 29 stable classes for this 70S ribosome dataset, similar to that from RE2DC. The t-SNE plot for the ISAC particles (blue dots in Figure 2(a)) reveals an interesting pattern. Unlike the RE2DC particles that segregate into well-dispersed islands (Figure 2(a)), a large fraction of particles is aggregated as a core in the center, surrounded by small satellite clusters with similar size in the peripheral. Nearly 60% (18 out of 29) of the stable ISAC classes (red dots in Figure 2(a)) correspond well to those off-center satellite clusters, whereas the remainder classes to the clusters located in the outer rim of the core. For those core-related ISAC classes, the class averages seem to be more blurred, as exemplified by the one in the lower-left corner in Figure 2(a), indicative of potential unresolved heterogeneity. Indeed, closer inspection reveals that almost every stable ISAC class contains impurity, as evidenced by that small numbers of red dots assigned to the same class by ISAC are scattered afar from those clusters (Figure 2(a) and Supplementary Figure 2), a minor imperfection of ISAC previously unknown until the present diagnosis using t-SNE visualization.

By contrast, the stable RE2DC classes are virtually without impurities (red dots in Figure 2(b) and Supplementary Figure 1). Besides, the size RE2DC classes exhibit a wide range in



(a)



(b)

Figure 2: t-SNE plot of particle images from the final iteration of a 2D classification with representative resulting 2D class-average images. (a) Results from ISAC: each particle image is denoted by a blue point, where the member in a class is denoted in red. Note that the pattern of particle distribution from ISAC is different from RE2DC. In addition, an ISAC class usually contains members scattered far from the main cluster. For RE2DC, it is seldom the case. (b) Results from RE2DC: each particle image is denoted by a blue point, where the member in a class is denoted in red.

distribution as revealed on the t-SNE plot, which we believe to better reflect the experimental reality than the equal size scenario [33]. Those findings have established that RE2DC is indeed a robust classification method where decent results can be obtained with a minimal number of iterations. Of note, RE2DC gives the highest particle yields ($\sim 92\%$) as compared to RELION ($\sim 84\%$) and ISAC ($\sim 76\%$). It is noted that a large fraction of good particles are rejected by ISAC due to being taken at lower defocus as previously found [36].

RE2DC reveals assembly intermediates of glutamine synthetase in solution

Having established that RE2DC is a robust 2D classification method, we employed it to explore a cryo-EM data of E. coli glutamine synthetase (GS) [46]. GS is a ubiquitous enzyme found in all organisms. This enzyme catalyzes the condensation of glutamate and ammonia to form glutamine at the expense of ATP, and it can be organized in different types of quaternary structures: dodecameric (mostly found in prokaryotes); octameric or dodecameric (mostly located in eukaryotes), and hexameric (also found in prokaryotes) [46]. The structure of GS from several organisms except E. coli has been determined to atomic resolution using X-ray crystallography, where the asymmetric unit has been found to be a dodecamer built by two back-to-back hexameric rings. An early negative electron microscopy study revealed that E. coli GS could exist as dodecamer or helical tubes, with the organization sensitive to pH or divalent ions [50]. To pave the way for elucidating subunit-subunit interactions of E. coli GS, we embarked on a cryo-EM approach using conditions that favored dodecamer. We then collected a cryo-EM dataset of E. coli GS using our Titan Krios cryo-EM. This dataset containing 94,154 particles was 2D classified using different approaches, including RE2DC, RELION and ISAC. Remarkably, all three different approaches (Figure 3) generated a number of high-quality top views and a side view corresponding to the back-to-back hexameric rings, indicating the tendency of E.coli GS in adopting preferred orientations in air-water interface. Those top views revealed variations of E coli GS in terms of compositions, including dodecamer, and other kinds of oligomers with fewer subunits, which may reflect assembly intermediates toward the dodecamer. This surprising finding may explain why E. coli GS is refractory to form good crystals for X-ray diffraction.

Remarkably, a comparison of the 2D classes from RE2DC (Figure 3(a)) against those of RELION (Figure 3(b)) or ISAC (Figure 3(c)) shows that RE2DC identified a peculiar form missed by RELION or ISAC. This peculiar form of crescent shape seems to correspond to an octamer. For those RE2DC classes (upper panel in Figure 3(a)), we located the corresponding clusters on the t-SNE plot (Figure 1(a)); it is noted that the peculiar form is located in the center of the t-SNE plot, denoted by No. 3. We hypothesized that RELION might also be able to reveal this form when larger prescribed number of classes was used to mitigate the attractor effect. To maximize the potential of RELION for sorting out heterogeneity in this dataset, we gradually increased the prescribed number from 50 to 200 with a step of 50. As the number of 150 was used (Supplementary Figure 3), RELION was indeed able to reveal this peculiar form of crescent shape, which was hidden when more popular parameters such as 50 or 100 were used (Figure 3(b)). This line of study attests that differentiating subtle structural variation in a heterogeneous dataset generally entails the use of a large number of class. Compared

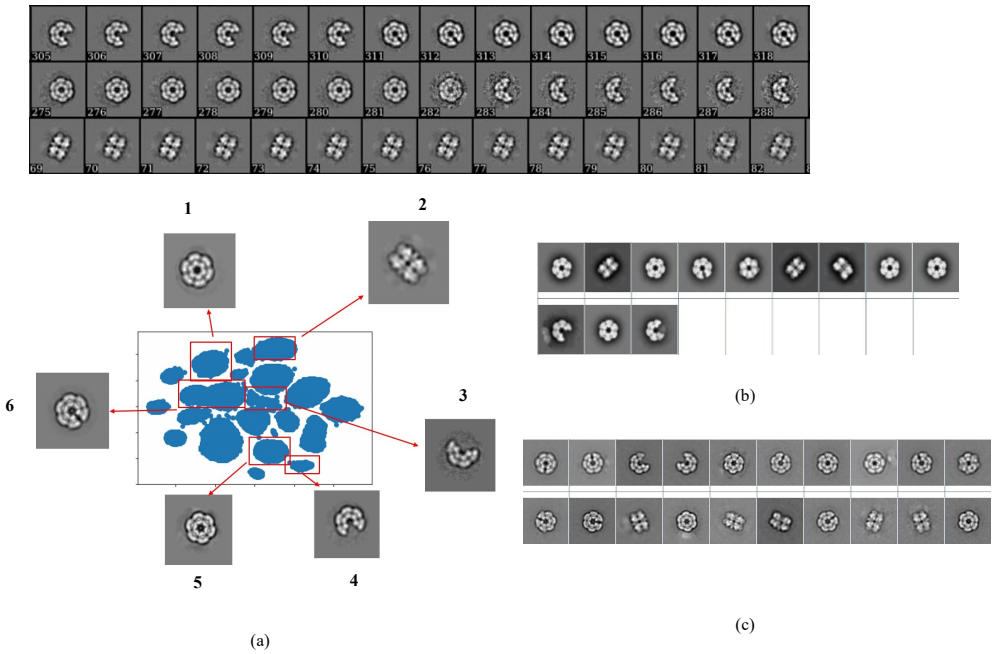


Figure 3: 2D classification results for GS particle images by three different methods. (a) Upper panel: Class averages from RE2DC. Lower panel: a t-SNE plot with six characteristic classes identified, where No. 3 denotes the peculiar form of crescent shape. (b) From RELION, where the number of pre-scribed class was set to 50. (c) From ISAC, where the size limit of each class was set to be 2000.

to these conventional 2D classification approaches, RE2DC does not require a user to specify this parameter as it is automatically determined by a data-driven approach. This unexpected finding suggests RE2DC and like approaches can serve as an ideal tool for discovering structural intermediates in vast classes of data in the absence of prior knowledge.

RE2DC with random partition classifies a large 80S ribosome dataset into a large number of homogenous classes in two hours

To test the efficiency of RE2DC on large datasets, we first tested a highly curated 80S ribosome dataset from Scheres's lab [51], provided as a RELION Benchmark example (accession number EMPIAR-10028). This dataset that contains a total of 105,247 excellent particle images with 360×360 pixels of 1.34 Å, was used for demonstrating RELION [20, 5] in reaching 3 Å at the dawn of resolution revolution, thus suitable for testing whether or not a different package has similar capability [36]. Drugged with an antibiotic molecule [51], this 80S ribosome does not exhibit significant movement between the 60S and 40S subunits. As a result, the major source of data heterogeneity is associated with the diversity in particle orientations. Our previous work of 3D reconstruction from those particles indicated the particle orientations were evenly

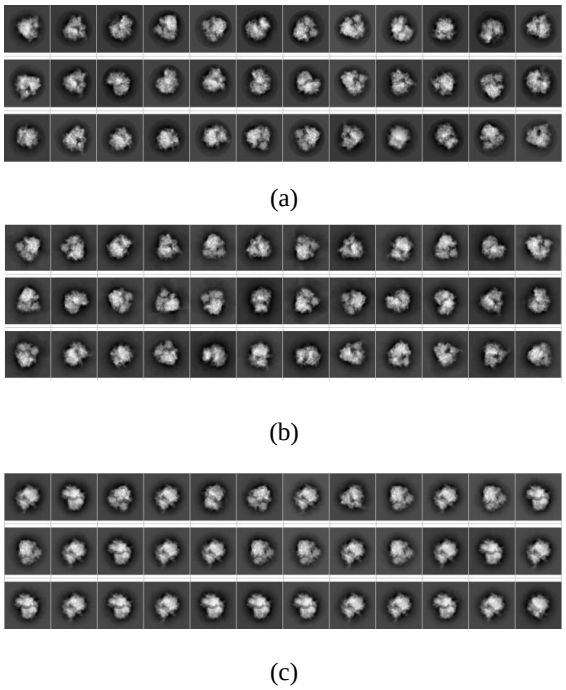


Figure 4: Representative class averages resulted from the 2D classification on a curated 80S particle dataset [EMPIAR-10028] using three different methods. (a) RELION. (b) ISAC. (c) RE2DC. (See Supplementary Figure 4 for all the class averages).

distributed in the angular space [36], which is also reflected by the large number of stable classes (520) generated by ISAC with the size limit of each class set to 200 (see [36] and Table 1 and Table 2). Since this 80S dataset is already highly curated, a good classification algorithm should give a high particle yield close to 100%. Indeed, RELION (one passage) gives 97% whereas ISAC gives 98% (Table 1 and Table 2). It is noted that conventional ISAC that relies on CPU hardware consumes 124 hours for the task of 2D classification. In this work, we also tested the newly released GPU-enabled ISAC and surprisingly found that it only took 14 hours, comparable to RELION [36].

With those figures of merits of current algorithms in mind, we went on to test RE2DC on this 80S ribosome dataset. To this end, we introduced a partition scheme where we set the size per partition to 5,000. Strikingly, RE2DC automatically generated 514 homogeneous classes (Figure 4 and Supplementary Figure 4(c)) with a particle yield of 97% (Table 1 and Table 2). The plethora of class averages resulting from RE2DC reflecting a large number of different views indicated RE2DC results could be used for generating a high-quality ab initio 3D reconstruction [36]. Remarkably, CPU-based RE2DC took only 2.5 hours with whereas the time is further

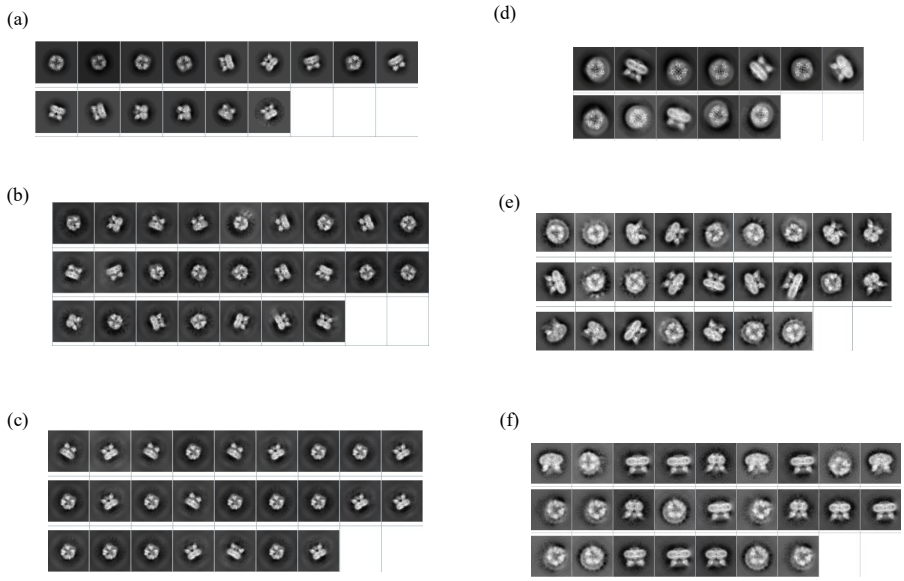
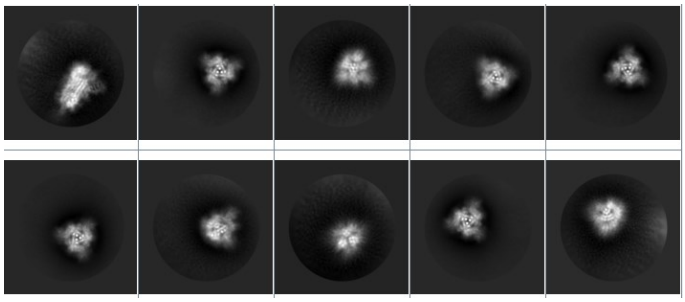


Figure 5: 2D classification average images from two non-curated TRPV channel datasets using three different classification methods. (a) - (c) TRPV channel [EMPIAR-10005] by RELION, ISAC, and RE2DC respectively. (d) (f) TRPV channel in nanodisc [EMPIAR-10059] by RELION, ISAC, and RE2DC respectively.

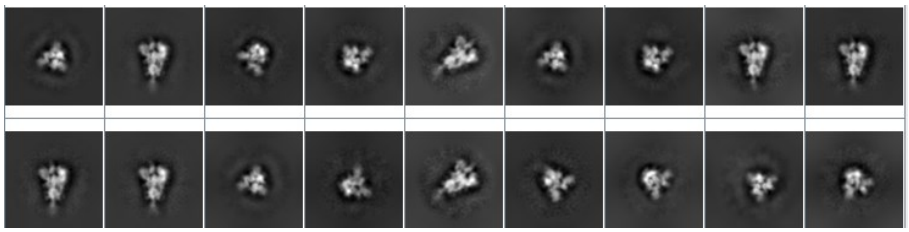
reduced to 2 hours when a GPU-based algorithm was used (Table 1 and Table 2). Those results clearly demonstrate RE2DC champions in its efficiency without compromising its robustness.

RE2DC classifies large non-curated large TRPV datasets less than four hours

Next, we tested RE2DC on non-curated membrane protein datasets, which are considered to be much more challenging for 2D classification. We first tested a famous TRPV dataset [EMPIAR-10005] from Yifan Cheng and David Julius [52]. This dataset contains a total of 88,915 particle images with 256×256 pixels of 1.2 Å, where the subset selected by the authors using extensive 2D and 3D classification was used to derive a first near-atomic resolution TRPV structure (3.4 Å) at the dawn of resolution revolution. While RELION spent 16 hours and harvested 57% particles, and ISAC spent 55 hours and obtained a similar yield, RE2DC only spent less than 2 hours and harvested 67% particles (Table 1 and Table 2). We further tested a dataset of ligand-bound TRPV embedded in nanodisc [EMPIAR-10059], also provided Yifan Cheng and David Julius [53]. This dataset contains a total of 218,787 particle images with 192×192 pixels of 1.2 Å. This dataset is much more heterogeneous due to the nanodisc density surrounding the channel, but was used to extend the resolution of TRPV cryo-EM structure to sub-3Å. Again, RE2DC spent the least amount of time (3.86 hours with CPU, or 3.3 hours with GPU) compared



(a)



(b)

Figure 6: 2D classification results of COVID-19 spike particles pick from 3,511 cryo-EM movie data [EMPIAR-10514] by two classification methods. (a) 2D class averages obtained by RELION. (b) 2D class averages obtained by RE2DC.

to more than 10 hours for RELION or ISAC, and gives the highest particle yield (73%)(Table 1 and Table 2), virtually twice the size of the subset selected by the authors using tedious 2D and 3D classification. Importantly, a 3D structure with 2.9 Å was obtained from the harvested particles. Those results show that for non-curated datasets only one pass of 2D classification is required for RE2DC, and again RE2DC is much faster than the popular method.

RE2DC enables timely structural determination for COVID-19 spike protein

This RE2DC algorithm, together with tests on those famous datasets, was completed prior to the breakout of COVID-19. Due to the fact that single particle cryo-EM is without the need for crystallization, it has been employed to deliver the structure of COVID-19 spike protein almost immediately after its emergence [14]. However, during the three years of this global pandemic, the strain of COVID-19 has kept evolving with the changing of the sequence of the spike protein. Keeping up with the protein structures of a new COVID-19 strain is of extreme importance for

understanding its pathological mechanisms and providing a rationale for the development of neutralizing antibodies. We then wondered whether or not RE2DC can be used for accelerating the structural research of COVID-19 proteins. To this end, we obtained 3,511 cryo-EM movies that contained images of the spike protein stabilized spike in the pre-fusion state, which was recently deposited by the Houston group [14] [EMPIAR-10514]. It is noted that this amount of movies can now be collected within 8 to 10 hours with a cryo-EM instrument operated in a high-throughput mode [47, 48, 49]. We performed motion correction on the movies to obtain 3,511 micrographs and used crYOLO [23] to automatically pick 375,076 particles representing the largest set in this study. For this dataset, we only compared RE2DC with RELION. Seeing that this dataset may contain a high degree of heterogeneity due to subtle conformation states or high levels of glycosylation [15], we set the prescribed class number for RELION to be 500 in order to avoid the necessity of doing multiple rounds of classification as in the case when a small class number is used. RELION and RE2DC sifted a similar number of particles where the particle yields are similar, 63% (235,300 particles) from RELION and 59% (221,916 particles) from RE2DC. However, RELION seems to capture top views more often, whereas the RE2DC results are more balanced between top and side views. Besides, the 2D class averages from RE2DC are better centered and exhibit more details (Figure 6). Remarkably, the computation time is drastically different—for RELION, it was 50.21 hours; whereas for RE2DC, also with GPU acceleration, it was only 7.76 hours. This finding suggests usage of RE2DC with a current workflow (Figure 1) may permit the revelation of the density map of COVID-19 spike protein on the day of cryo-EM data collection. Furthermore, we perceive that, with on-the-fly usage of RE2DC, realizing robust real-time single cryo-EM structural determination is possible.

Conclusions

Here, we have successfully developed a new unsupervised 2D classification method called RE2DC for processing single particle cryo-EM data. Our tests using various datasets have demonstrated RE2DC is a robust and efficient tool that can be widely used, especially useful when the data is massive or highly heterogeneous. Of note, RE2DC has particular strengths in discovering rare structural intermediates. RE2DC can be implemented with modest computation resources and easily integrated with any cryo-EM package.

Table 1: The classification time of the CPU implementation on six datasets.

Algorithm \ Dataset	RELION	ISAC	RE2DC
70S ribosome	0.43 (30 classes)	1.78 (200 members)	0.28
GS Protein	22.91 (150 classes)	50.92 (1000 members)	1.82
80S ribosome	62.32(f) (520 classes)	124.50 (4X binning, 200 members)	2.54 (4X binning)
NC-TRPV1	16.03(f) (200 classes)	55.61 (3X binning, 1000 members)	1.76 (3X binning)
NanoD-TRPV1	12.36(f) (100 classes)	*	3.86 (3X binning)
SARS-CoV-2	120.03(f) (500 classes)	*	8.45 (5X binning)

[a] Fast-subset is used for RELION whenever possible and is labeled with (f). The size of partition used in RE2DC is set to 5,000 or 70S ribosome, 80S ribosome and NC-TRPV1, 20,000 is used for SARS-CoV-2 and 10,000 is used for others and dynamic-SUP is used for the clustering step. The number of Message Passing Interface (MPI) used is 44 and the number of GPU is 1. Finally, the execution time is the average of 5 rounds for 70S ribosome and 3 rounds for all the other datasets. Note we do not count the data format conversion time on the platform.

Methods

To cope with those issues of current 2D classification regarding robustness or efficiency, we introduced a new 2D classification tool, termed RE2DC, which stands for robust and efficient 2D classification. As a 2D classification algorithm consists of sub-algorithms of image alignment and clustering [34], a robust and efficient classification method demands its sub-algorithms to be also robust and efficient. With this in mind, we first developed a clustering algorithm that is both robust and efficient. We employed γ-SUP clustering algorithm as the cornerstone [38] because of its robustness [38], with the promise of eliminating the issues of current classification methods stemming from K-means clustering. However, the computation costs of original γ–SUP prevented it from being useful when a large number of particles are encountered ($N > 10^4$) — its computation complexity for calculating the pairwise weight matrix scales quadratically with N, the number of data points (particle images). In practice, γ-SUP has another shortcoming—it requires human intervention for choosing a good hyper-parameter, which is critical for determining the number of clusters and the route of data points approaching their destiny clusters. In the present work, we provided novel solutions to eliminate these drawbacks, and obtained an efficient and automated version of γ-SUP. First, noticing that the weight matrix of γ-SUP is

Table 2: The classification time of the GPU implementation on six datasets.

Algorithm \ Dataset	RELION	ISAC	RE2DC
70S ribosome	0.28 (30 classes)	0.33 (200 members)	0.18
GS Protein	8.68 (150 classes)	4.08 (1000 members)	1.58
80S ribosome	30.12(f) (520 classes)	14.55 (4X binning, 200 members)	2.02 (4X binning)
NC-TRPV1	13.56(f) (200 classes)	5.38 (3X binning, 1000 members)	1.51 (3X binning)
NanoD-TRPV1	9.49(f) (100 classes)	12.65 (3X binning, 1000 members)	3.33 (3X binning)
SARS-CoV-2	50.21(f) (500 classes)	*	7.76 (5X binning)

[a] Fast-subset is used for RELION whenever possible and is labeled with (f). The size of partition used in RE2DC is set to 5,000 or 70S ribosome, 80S ribosome and NC-TRPV1, 20,000 is used for SARS-CoV-2 and 10,000 is used for others and dynamic-SUP is used for the clustering step. The number of Message Passing Interface (MPI) used is 44 and the number of GPU is 1. Finally, the execution time is the average over 5 rounds for 70S ribosome and 3 rounds for all the other datasets. Note we do not count the data format conversion time on the platform.

highly sparse, we constructed a sparse approximation for it by employing the well-known k-d-tree method. The usage of this approximation appears to have successfully decreased the computation complexity for calculating the pairwise weight matrix from $O(N^2)$ to $O(N)$. Secondly, to automate the selection of an optimal hyper-parameter, we devised a data-driven approach that screened over a wide range to search for an estimate that maximized a target function—e.g., the number of reasonable-sized clusters. The search is expedited by a highly efficient two-stage procedure. As for the sub-algorithm for image alignment, we adopted a straightforward approach based on multi-reference-alignment (MRA) framework [27, 31] with the reference images directly generated from the data and then used for initiating the alignment process. As this framework utilizes cross-correlation scores for image alignment, the outcome will not be reliable when image SNR is too low [29]. To overcome this issue, we introduced dimension reduction MRA (DRMRA), where we used a rapid dimension reduction method (2SDR) [40, 41] to de-noise cryo-EM images. This 2SDR strategy has been demonstrated to improve the accuracy of image alignment, which in turn leads to a significant reduction of the number of iterations required for current 2D classification methods [36]. Since the open-source code for GPU-accelerated MRA is

available [42], we used it to make a GPU version of RE2DC. Finally, we introduced a distributed computation strategy that divided a large dataset into several smaller datasets to classify them separately and in parallel, which further speeded up RE2DC.

Dimension Reduction Multi-reference Alignment (DRMRA)

In this section, we elaborate on the algorithm of dimension reduction MRA (DRMRA). PCA methods or the extension have been popular for dimension reduction. PCA was introduced for processing single particle negative-stained EM images in early days [54], which was pivotal to the development of single particle cryo-EM [8]. Unfortunately, PCA vectorization involved in dimension reduction has been found to render the eigen-space estimate an unstable and time-consuming process when the image size becomes large [55]. This inevitable caveat is inevitable hinders the pursuit of cryo-EM structure to high resolution for large-pixelated images. To tackle this problem, 2SDR [40, 41] has been proposed recently. This approach utilizes a two-stage procedure for dimension reduction, thereby circumventing the need of image vectorization. Crucially, in the first stage, 2SDR preserves the tensor form of the data matrix, which avoids the vectorization process and estimates the two eigenvector sets in a much lower dimensional space from column and row space. The images are then projected onto the space spanned by the Kronecker product of these two eigenvector sets. In the second stage, the highly correlated features of the projected core tensor are then de-correlated by modeling it as a PCA model and further reducing its dimensionality. It has been shown that 2SDR lowers the computation complexity of traditional PCA by several orders of magnitude [40, 41] and can be used to empower current cryo-EM algorithm in the context of experimental images [36]. The hybrid PCA model associated with 2SDR and a consistent rank selection method is discussed in [41]. The hybrid PCA model has been demonstrated to be a better approximation for cryo-EM images than the PCA model. It is noted that the reconstructions of 2SDR can also serve as an effective de-noising tool that can improve the contrast of the particle image. When combining 2SDR with reference-free alignment, a processing tool called PrePro was made [36]. This PrePro has been proven effective in improving the classification results and conferred a faster convergence rate [36] [1]. We therefore added 2SDR to MRA to create DRMRA by directly employing PrePro. As shown in Supplementary Figure 5, by priming MRA with PrePro for initial alignment, we found the class averages became stable in the second iteration of RE2DC. By contrast, conventional MRA required many more iterations to converge. This phenomenon that de-noised particles can lead to improvement in aligning cryo-EM images has also been observed by other groups [57],

[1] The fact that better initial can improve the alignment results or speed up the classification has also been observed in [56]

with a completely different approach that resorted to AI.

Transforming γ-SUP from a robust to robust and efficient clustering

Image clustering is primary for resolving different views and/or other types of heterogeneity in cryo-EM data. In a non-curated cryo-EM dataset, there always exist non-particle artifacts. As non-particles should not be aligned well with particles, a good algorithm with the capability of labeling misaligned images as outliers can be used to identify those non-particles. In addition, when separating different projections or conformations into clusters becomes a primary interest, a large number of classes are desired. Considering those aspects, γ-SUP is an ideal clustering algorithm to suit a dataset that potentially contains a large number of clusters or is contaminated with noise or outliers. Previously, γ-SUP applied to simulated cryo-EM datasets demonstrated this clustering method conferrer higher accuracy in identifying misaligned outliers over existing methods, attesting it is a much more robust clustering algorithm than current algorithms [38]. At a high level, γ-SUP models the data with the q-Gaussian mixture and employs the minimum γ -divergence [58] as the metric for similarity estimation. It then aims to obtain a numerical solution by adopting the self-updating procedure (SUP) [59]. In the updating process, each data point moves to a new position during the iterations. The new position depends on the influences the data point receives from other points. At the end of the process, every point reaches an equilibrium position and points that arrive at the same position are regarded to be in the same cluster. The influence force or weight can be written as: $w_{ij}^{(t)} = exp_{1-s}(\frac{-1}{\tau^2}|x_i^{(t)} - x_j^{(t)}|_2^2)$, where exp_q is the q-exponential function, τ and s are tuning parameters. Notably, γ-SUP eliminates the need to set random cluster centers for initializing the process because it treats each data point as a singleton cluster for initialization. In practice, γ-SUP is quite insensitive to s. Thus, only the scaling parameter τ is left to be tuned, where the tuning is needed for different datasets. Since the parameter τ is proportional to the influence region in the updating process, it ignores the influence of data outside a certain range determined by τ. In other words, τ determines how many points can be accommodated in the same cluster. For instance, as shown in Supplementary Figure 6, when τ is small enough, there exists no influence between any two data points such that each point forms an individual cluster by itself. As τ increases, data points start to attract each other to form a cluster. For the original γ –SUP [38], an appropriate choice of τ can be made by inspecting the transition point in a plot as previously suggested [38]. However, in practice, the behavior of phase transition is not always clear that the accordingly selection of hyperparameters τ becomes difficult. In addition, when sample size n increases, the computation time that scales with the order of $O(N^2)$ has prohibited the application of original γ –SUP to large datasets.

In order to deliver a practical clustering method based on γ –SUP for large datasets, we made several modifications to the original algorithm as follows. First, we observed the weight matrix for each sample point was sparse. This sparsity of the weight matrix is the inherent nature of γ –SUP due to the use of γ -divergence and the q-Gaussian, which sets a rejection bound. We then exploited this sparsity property by employing the k-nearest neighbor graph to approximate the weight matrix. This is achieved by building a space-partitioning tree for sample points at the beginning of each iteration [60]. This approach enables an efficient search for the nearest neighbor. The overall complexity can thus be reduced to $O(N\log(N))$ or even $O(N)$. Secondly, we incorporated dynamic-SUP [61] into the new clustering algorithm so as to let the number of neighbors grow with iteration. This approach, which allowed us to focus on the local structure at the beginning, indeed gave a better performance in practice. Finally, to select the exact τ, we devised a two-stage procedure for the search. In the first stage, binary search is utilized to find the region that gives non-trivial solutions. Next, we utilized grid search to scan this region to find the parameter that maximized a targeted function, for instance, particle yield, number of reasonable-size clusters, or combination of these terms.

Using t-SNE visualization for diagnosis of 2D classification

Visualizing high-dimensional data in low dimensions has been recognized as an important problem in many scientific applications, for which numerous statistical tools have been developed. Those approaches include PCA, Multidimensional Scaling (MDS), and t-Distributed Stochastic Neighbor Embedding (t-SNE) [44]. As the history of methodology development is concerned, t-SNE is the first one that can successfully separate the famous MNIST dataset into 10 clusters. Since such success, t-SNE has been widely applied to exploring the heterogeneity of high-dimensional data with applications in diverse areas, including cancer biology and bioinformatics. 2D classification in cryo-EM is virtually treated as a black box for an end-user and it is hard to have an intuitive insight into the progress or the quality of classification results. Here, to unveil the black box, we introduced t-SNE to display high-dimension image data in low dimensions for visualizing the final 2D classification results and monitoring the progress of the classification. The t-SNE algorithm comprises two main stages. First, it transforms the similarity matrix into a probability distribution (e.g., Gaussian distribution and t-distribution) for both the input data of high dimension and the visualization in low dimension (usually 2 or 3). Secondly, it minimizes the Kullback–Leibler (KL) divergence between the two distributions by a gradient descent algorithm. In practice, under a suitable parameter setting, t-SNE can effectively preserve the high-level qualitative structures of the input data, such as clusters or groups of nearest

neighbor [62]. Our implementation of t-SNE with RE2DC chose FFT-accelerated interpolation-based t-SNE [63] and utilized scaled PCA for its initialization [45]. We surprisingly found the t-SNE plot could be used to evaluate the quality of the classification results, as demonstrated in Supplementary Figure 7, in which red points denote the data points that are "considered" by the classification algorithm to belong to the same group. We can see that in the middle panel of Supplementary Figure 7, some of the red points are scattered across the t-SNE plot, which indicates that the algorithm under our test groups together those data points that belong to different blobs. As a result, the corresponding class average becomes blurred. In contrast, in the right panel of Supplementary Figure 7, the red points assigned by the algorithm are localized within the same blob. In this case, the corresponding class average shows more structural details.

Accelerated alignment by GPU

It is noted that the calculation of similarities in the alignment step is independent of particles, 2D reference or orientation. It is thus possible to calculate them in parallel using GPU. In our algorithm, the reference-based alignment and the reference-free alignment in PrePro can be GPU-accelerated using the routines from Cryo-RALib [42].

Implementing RE2DC sub-algorithms

RE2DC consists of two sub-algorithms, one for image alignment and the other for image clustering as shown in Algorithm 1 and Algorithm 2. The sub-algorithm for alignment follows the general architecture of MRA, but with major advancements as follows. First, to implement DRMRA with 2SDR, we employed PrePro [36] as it uses 2SDR instead of PCA for dimension reduction to achieve fast dimension reduction. The PrePro was found to add insignificant overhead, but speed up the convergence [36]. It was shown that for most datasets pre-processed by PrePro, only three iterations are required by RELION or ISAC to achieve convergence. As for the sub-algorithm of image clustering, we used γ-SUP instead of K-means for clustering to avoid the shortcoming of clustering with popular classification methods that are largely rooted in K-means. In this development, we implemented and tested six versions of γ–SUP. These six versions are divided into two sets for test. The first set includes the original one [38], a new one with k-nearest neighbors (knn), which we term knn γ –SUP, and a variant of knn γ –SUP where k increases in each iteration, which we call 'dynamic knn γ–SUP' (Figure 1(b)). The three in the second set called γ–estimator [38], are the same as those in the first set, except that the sample points do not move in each iteration. The number of k is set to 50 in the knn

Algorithm 1: RE2DC

Input: Set of experiment images X with N particles and feature size p

Number of iteration J

Target dimension for 2SDR p_0, q_0, r

Number of particles per-partition n

Number of reference for global alignment m

The size of largest/smallest cluster g_l/g_s

Output: Set of class averages C

```
Applying PrePro on X.
iteration = 0. repeat
    iteration = iteration +1.
    Reduce the dimension of X from p to r using 2SDR to obtain X_r.
    Random partition X_r into I = N/n partitions to obtain X_{r_i}.
    Make the t-SNE plot of X_{r_0} and label each cluster.
    for each partition do
        Cluster X_{r_i} using modified γ-SUP (Algorithm 2).
        Take averages of each cluster and discard the cluster with a size smaller than g_s.
    if convergence then
        Break.
    else
        if iteration ≠ J then
            Discard the class averages with cluster size larger than g_l. Combine the class
              averages from all partitions into k class averages using γ-SUP.
        X is aligned to the combined m class averages through reference-based alignment to
          obtain orientation parameters.
        Orientation parameters are applied to X.
until iteration = J;
(Optional) Split the cluster with a sizes larger than g_l using k-means.
Take the class averages of each cluster to obtain C.
Return C
```

Algorithm 2: Modified γ-SUP

Input: Set of experiment images X with n particles and feature size r

Tuning parameter s, τ

Number of neighbors k

Output: Set of clusters $\tau\mu$

iteration $= 0$.

Start with $\mu_i = X_i/\tau, i = 1, \cdots, n$.

repeat

  tree = construct-Tree(μ).

  for $i = 1 : n$ **do**

    NN = tree.nearestNeighbors(μ_i).

    $w_{ij} = exp_{1-s}(-|\mu_i - \mu_j|^2), j \in NN$.

    $z_i = \sum_{j \in NN} \frac{w_{ij}}{\sum_{k \in NN} w_{ik}} \mu_j$.

  $\mu_i = z_i, i = 1, \cdots, n$.

  iteration $=$ iteration $+1$.

until *Convergence*;

Return $\tau\mu$.

version and set to $10 + j \times 5$ in the dynamic-knn at iteration j. The clustering performance of the advanced variants of γ-SUP surpasses the original one in greatly reducing the computation time, as shown in Figure 1(b). To combine DRMRA and a new γ–SUP iteratively, we decided not to feed too many references to perform the reference-based alignment. This way, the class assignment can be largely determined in the clustering stage. The class averages generated by the clustering step are further grouped by γ-SUP into a smaller set of class averages as references for the next round of image alignment. Furthermore, the class average that contains more than g_l or smaller than g_s will be screened out in the alignment step. This also has the benefit of speeding up the alignment step. Finally, the algorithm architecture of RE2DC is highly flexible. For instance, one can swap in those MRA used for maximum likelihood approach [5, 33] or stochastic hill-climbing scheme [6].

Random partition

We used random partition to partition a large dataset is used at the start of each iteration to reduce the computation time through parallelization. We found this scheme also improves the classification performance other than saving time. In practice, we suggest to set number

of particles per-partition n to 5000, number of references m to 20, and the threshold of group size to $g_l = 1000$, $g_s = 15$, respectively. Moreover, we set the default iteration number to 3 for the convergence criterion and further validated this through the t-SNE plot. In a typical scenario, one can use a strategy without partitioning the data but instead setting the g_s to a small number. This should allow the removal of apparent non-particles. In other scenarios, one can adjust the τ, g_s and the number of particles per partition. This may allow the user to inspect subtle structural variations within clusters to analyze the potential heterogeneity hidden within the dataset.

Pre-treatment of the data prior to RE2DC

Considering the contrast transfer function (CTF), a user has an option to choose whether or not to correct it prior to the classification. For example, when RE2DC is implemented on SCIPION, one can choose to apply "phase flipping" to particle images before the classification as commonly performed before popular classification algorithms. Like most of MRA-based algorithms, we down-sampled images with large box size. Therefore, noise could be further reduced from the data, but high-resolution information will be removed as well so that fine details may not be evident in the class averages generated by RE2DC. If loss of details has rendered the distinction between particles and invalid particles harder, we provide an option to generate the class averages using the members in the original dataset, which is not down-sampled, while using the alignment parameters estimated based on the down-sampled images. If one is still concerned about the quality of the class averages from RE2DC, we have tested the possibility of enhancement by CTF correction and local 2D refinement using the CTF-correction module available from other packages such as SPHIRE [35]. The class averages can be further enhanced by adjusting their power spectrum using the procedure previously described by Penczek [64]. Those options will further render RE2DC to meet the demand from more difficult datasets when fine details matter.

Acknowledgments

This work was supported by Academia Sinica Grants [AS-GCS-108-08] and [AS-IA-110-M05] to I-P.T.; Academia Sinica SUMMIT Projects [AS-SUMMIT-107], [AS-SUMMIT-108], [AS-SUMMIT-109] to W.-H.C.; and by the Ministry of Science and Technology Taiwan [MOST 106-2118-M-001-001-MY2] to I-P.T.; [MOST 106-2321-B-001-050-MY3] to K.-P.W.; and [MOST 110-2118-M-110-003-MY2] to S.-Z.C.. The authors are thankful for the excellent support from the Academia Sinica Cryo-EM Center [AS-CFII-108-110] for cryo-EM data collection.

Author contributions

I-P.T. and W.-H.C. conceived and designed the project. T.-L.C. contributed to algorithm development. S.-C.C. wrote the codes, conducted the classification experiments. K.-P.W. purified GS and provided GS cryo-EM data. S.-C.C., H.-H.L., and W.-H.C. analyzed the classification results. S.-C.C., I.-P.T., and W.-H.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Code availability

The code will be available upon request once this manuscript is published in a peer-reviewed journal.

References

[1] Liang Jin, Anna-Clare Milazzo, Stuart Kleinfelder, Shengdong Li, Philippe Leblanc, Fred Duttweiler, James C Bouwer, Steven T Peltier, Mark H Ellisman, and Nguyen-Huu Xuong. Applications of direct detection device in transmission electron microscopy. Journal of structural biology, 161(3):352–358, 2008.

[2] G McMullan, S Chen, R Henderson, and AR Faruqi. Detective quantum efficiency of electron area detectors in electron microscopy. Ultramicroscopy, 109(9):1126–1143, 2009.

[3] Melody G Campbell, Anchi Cheng, Axel F Brilot, Arne Moeller, Dmitry Lyumkis, David Veesler, Junhua Pan, Stephen C Harrison, Clinton S Potter, Bridget Carragher, et al. Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure, 20(11):1823–1828, 2012.

[4] X. Li, P. Mooney, S. Zheng, C.R. Booth, M.B. Braunfeld, S. Gubbens, D.A. Agard, and Y. Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em. Nature Methods, 10:584–590, 2013.

[5] Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. Journal of structural biology, 180(3):519–530, 2012.

[6] Cyril F Reboul, Frederic Bonnet, Dominika Elmlund, and Hans Elmlund. A stochastic hill climbing approach for simultaneous 2d alignment and clustering of cryogenic electron microscopy images. Structure, 24(6):988–996, 2016.

[7] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. Nature methods, 14(3): 290, 2017.

[8] Joachim Frank. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press, 2006.

[9] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia MGE Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, et al. Single-particle cryo-em at atomic resolution. Nature, 587(7832):152–156, 2020.

[10] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-em. Nature, 587(7832):157–161, 2020.

[11] Radostin Danev, Haruaki Yanagisawa, and Masahide Kikkawa. Cryo-em performance testing of hardware and data acquisition strategies. Microscopy, 70(6):487–497, 2021.

[12] Wei-Hau Chang, Shih-Hsin Huang, Hsin-Hung Lin, Szu-Chi Chung, I Tu, et al. Cryo-em analyses permit visualization of structural polymorphism of biological macromolecules. Frontiers in Bioinformatics, page 74, 2021.

[13] Shuwen Zhang, Shitao Zou, Deyao Yin, Lihong Zhao, Daniel Finley, Zhaolong Wu, and Youdong Mao. Usp14-regulated allostery of the human proteasome by time-resolved cryo-em. Nature, 605(7910):567–574, 2022.

[14] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. Science, 367(6483):1260–1263, 2020.

[15] Tzu-Jing Yang, Pei-Yu Yu, Yuan-Chih Chang, Kang-Hao Liang, Hsian-Cheng Tso, Meng-Ru Ho, Wan-Yu Chen, Hsiu-Ting Lin, Han-Chung Wu, and Shang-Te Danny Hsu. Effect of sars-cov-2 b. 1.1. 7 mutations on spike protein structure and function. Nature structural & molecular biology, 28(9):731–739, 2021.

[16] Dhiraj Mannar, James W Saville, Xing Zhu, Shanti S Srivastava, Alison M Berezuk, Katharine S Tuttle, Ana Citlali Marquez, Inna Sekirov, and Sriram Subramaniam. Sars-cov-2 omicron variant: Antibody evasion and cryo-em structure of spike protein–ace2 complex. Science, 375(6582):760–764, 2022.

[17] Yunlong Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, Shuo Du, Jing Wang, Qianqian Li, Xiaosu Chen, et al. Ba. 2.12. 1, ba. 4 and ba. 5 escape antibodies elicited by omicron infection. Nature, 608(7923):593–602, 2022.

[18] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. Journal of structural biology, 157(1):38–46, 2007.

[19] JM De la Rosa-Trevín, A Quintana, L Del Cano, A Zaldivar, I Foche, J Gutiérrez, J Gómez-Blanco, J Burguet-Castell, J Cuenca-Alba, V Abrishami, et al. Scipion: a software framework toward integration, reproducibility and validation in 3d electron microscopy. Journal of structural biology, 195(1):93–99, 2016.

[20] Dari Kimanius, Björn O Forsberg, Sjors HW Scheres, and Erik Lindahl. Accelerated cryo-em structure determination with parallelisation using gpus in relion-2. elife, 5:e18722, 2016.

[21] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cistem, user-friendly software for single-particle image processing. Elife, 7:e35383, 2018.

[22] Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim JH Hagen, Erik Lindahl, and Sjors HW Scheres. New tools for automated high-resolution cryo-em structure determination in relion-3. elife, 7:e42166, 2018.

[23] Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, et al. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. Communications biology, 2(1):1–13, 2019.

[24] Dimitry Tegunov and Patrick Cramer. Real-time cryo-electron microscopy data preprocessing with warp. Nature methods, 16(11):1146–1152, 2019.

[25] Joseph Caesar, Cyril F Reboul, Chiara Machello, Simon Kiesewetter, Molly L Tang, Justin C Deme, Steven Johnson, Dominika Elmlund, Susan M Lea, and Hans Elmlund. Simple 3.0. stream single-particle cryo-em analysis in real time. Journal of structural biology: X, 4:100040, 2020.

[26] Richard Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proceedings of the National Academy of Sciences, 110(45):18037–18041, 2013.

[27] Marin van Heel and M Stöffler-Meilicke. Characteristic views of e. coli and b. stearothermophilus 30s ribosomal subunits in the electron microscope. The EMBO journal, 4(9): 2389–2395, 1985.

[28] Fred J Sigworth. A maximum-likelihood approach to single-particle image refinement. Journal of structural biology, 122(3):328–339, 1998.

[29] Fred J Sigworth, Peter C Doerschuk, Jose-Maria Carazo, and Sjors HW Scheres. An introduction to maximum-likelihood methods in cryo-em. In Methods in enzymology, volume 482, pages 263–294. Elsevier, 2010.

[30] Amit Singer and Fred J Sigworth. Computational methods for single-particle electron cryomicroscopy. Annual Review of Biomedical Data Science, 3, 2020.

[31] Peilong Lu, Xiao-chen Bai, Dan Ma, Tian Xie, Chuangye Yan, Linfeng Sun, Guanghui Yang, Yanyu Zhao, Rui Zhou, Sjors HW Scheres, et al. Three-dimensional structure of human γ-secretase. Nature, 512(7513):166–170, 2014.

[32] Philip R Baldwin, Yong Zi Tan, Edward T Eng, William J Rice, Alex J Noble, Carl J Negro, Michael A Cianfrocco, Clinton S Potter, and Bridget Carragher. Big data in cryoem: automated collection, processing and accessibility of em data. Current opinion in microbiology, 43:1–8, 2018.

[33] COS Sorzano, JR Bilbao-Castro, Y Shkolnisky, M Alcorlo, R Melero, G Caffarena-Fernández, M Li, G Xu, R Marabini, and JM Carazo. A clustering approach to multireference alignment of single-particle projections in electron microscopy. Journal of structural biology, 171(2):197–206, 2010.

[34] Zhengfan Yang, Jia Fang, Johnathan Chittuluru, Francisco J Asturias, and Pawel A Penczek. Iterative stable alignment and clustering of 2d transmission electron microscope images. Structure, 20(2):237–247, 2012.

[35] Toshio Moriya, Michael Saur, Markus Stabrin, Felipe Merino, Horatiu Voicu, Zhong Huang, Pawel A Penczek, Stefan Raunser, and Christos Gatsogiannis. High-resolution single particle analysis from electron cryo-microscopy images using sphire. JoVE (Journal of Visualized Experiments), (123):e55448, 2017.

[36] Szu-Chi Chung, Hsin-Hung Lin, Po-Yao Niu, Shih-Hsin Huang, I-Ping Tu, and Wei-Hau Chang. Pre-pro is a fast pre-processor for single-particle cryo-em by enhancing 2d classification. Communications biology, 3(1):1–12, 2020.

[37] Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. Cell, 161(3):438 – 449, 2015. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2015.03.050. URL http://www.sciencedirect.com/science/article/pii/S0092867415003700.

[38] Ting-Li Chen, Dai-Ni Hsieh, Hung Hung, I-Ping Tu, Pei-Shien Wu, Yi-Ming Wu, Wei-Hau Chang, and Su-Yun Huang. γ-sup: A clustering algorthm for cryo-electron microscopy images of asymmetric particles. The Annals of Applied Statistics, 8(1):259–285, 2014.

[39] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research, 15(1):3221–3245, 2014.

[40] Szu-Chi Chung, Po-Yao Niu, Su-Yun Huang, Wei-Hau Chang, and I-Ping Tu. 2sdr: Applying kronecker envelope pca to denoise cryo-em images. CoRR, 2019.

[41] Szu-Chi Chung, Shao-Hsuan Wang, Po-Yao Niu, Su-Yun Huang, Wei-Hau Chang, I Tu, et al. Two-stage dimension reduction for noisy high-dimensional images and application to cryogenic electron microscopy. Annals of Mathematical Sciences and Applications, 5(2): 283–316, 2020.

[42] Szu-Chi Chung, Cheng-Yu Hung, Huei-Lun Siao, Hung-Yi Wu, Wei-Hau Chang, and I-Ping Tu. Cryo-ralib-a modular library for accelerating alignment in cryo-em. In 2021 IEEE International Conference on Image Processing (ICIP), pages 225–229. IEEE, 2021.

[43] Jie Fu, Haixiao Gao, and Joachim Frank. Unsupervised classification of single particles by cluster tracking in multi-dimensional space. Journal of structural biology, 157(1):226–239, 2007.

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[45] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. Nature communications, 10(1):1–14, 2019.

[46] Marco Betti, Margarita García-Calderón, Carmen M Pérez-Delgado, Alfredo Credali, Guillermo Estivill, Francisco Galván, José M Vega, and Antonio J Márquez. Glutamine

synthetase in legumes: recent advances in enzyme structure and functional genomics. International Journal of Molecular Sciences, 13(7):7994–8024, 2012.

[47] Anchi Cheng, Edward T Eng, Lambertus Alink, William J Rice, Kelsey D Jordan, Laura Y Kim, Clinton S Potter, and Bridget Carragher. High resolution single particle cryo-electron microscopy using beam-image shift. Journal of structural biology, 204(2):270–275, 2018.

[48] Chunling Wu, Xiaojun Huang, Jing Cheng, Dongjie Zhu, and Xinzheng Zhang. High-quality, high-throughput cryo-electron microscopy data collection via beam tilt and astigmatism-free beam-image shift. Journal of structural biology, 208(3):107396, 2019.

[49] Rouslan G Efremov and Annelore Stroobants. Coma-corrected rapid single-particle cryo-em data collection on the cryo arm 300. Acta Crystallographica Section D: Structural Biology, 77(5):555–564, 2021.

[50] R Cl Valentine, BM Shapiro, and ER Stadtman. Regulation of glutamine synthetase. xii. electron microscopy of the enzyme from escherichia coli. Biochemistry, 7(6):2143–2152, 1968.

[51] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. Elife, 3: e03080, 2014.

[52] Maofu Liao, Erhu Cao, David Julius, and Yifan Cheng. Structure of the trpv1 ion channel determined by electron cryo-microscopy. Nature, 504(7478):107, 2013.

[53] Yuan Gao, Erhu Cao, David Julius, and Yifan Cheng. Trpv1 structures in nanodiscs reveal mechanisms of ligand and lipid action. Nature, 534(7607):347–351, 2016.

[54] Marin Van Heel and Joachim Frank. Use of multivariates statistics in analysing the images of biological macromolecules. Ultramicroscopy, 6(1):187–194, 1981.

[55] Hung Hung, Pei-Shien Wu, I-Ping Tu, and Su-Yun Huang. On multilinear principal component analysis of order-two tensors. Biometrika, 99:569–583, 2012.

[56] Ayelet Heimowitz, Nir Sharon, and Amit Singer. Centering noisy images with application to cryo-em. SIAM Journal on Imaging Sciences, 14(2):689–716, 2021.

[57] Eugene Palovcak, Daniel Asarnow, Melody G Campbell, Zanlin Yu, and Yifan Cheng. Enhancing the signal-to-noise ratio and generating contrast for cryo-em images with convolutional neural networks. IUCrJ, 7(6), 2020.

[58] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis, 99(9):2053–2081, 2008.

[59] Ting-Li Chen and Shang-Ying Shiu. A new clustering algorithm based on self-updating process. JSM proceedings, statistical computing section, Salt Lake City, Utah, pages 2034–2038, 2007.

[60] Yury A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence, 2018.

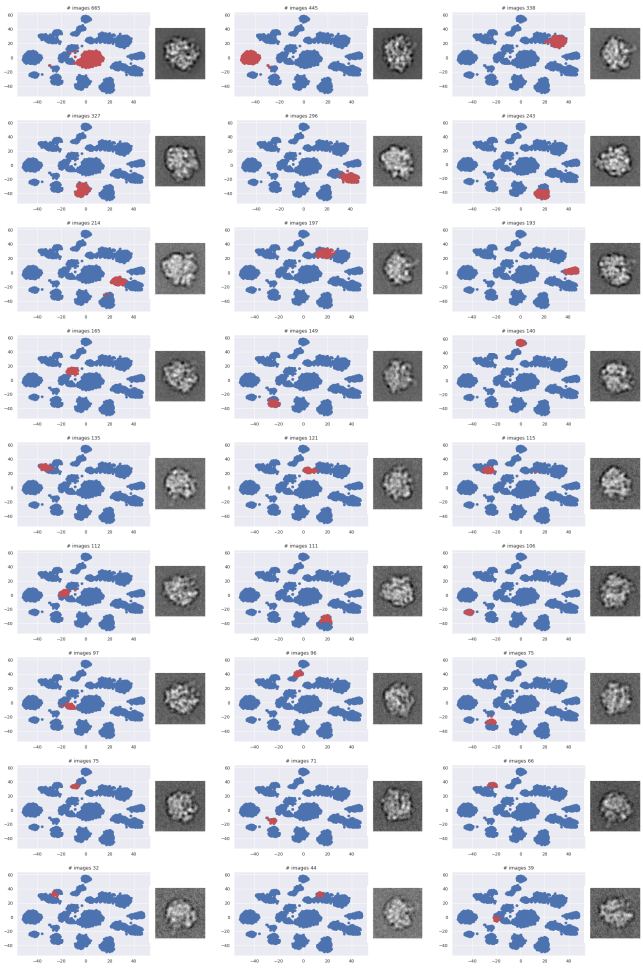
[61] Shang-Ying Shiu and Ting-Li Chen. On the strengths of the self-updating process clustering algorithm. Journal of Statistical Computation and Simulation, 86(5):1010–1031, 2016.

[62] George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. SIAM Journal on Mathematics of Data Science, 1(2):313–332, 2019.

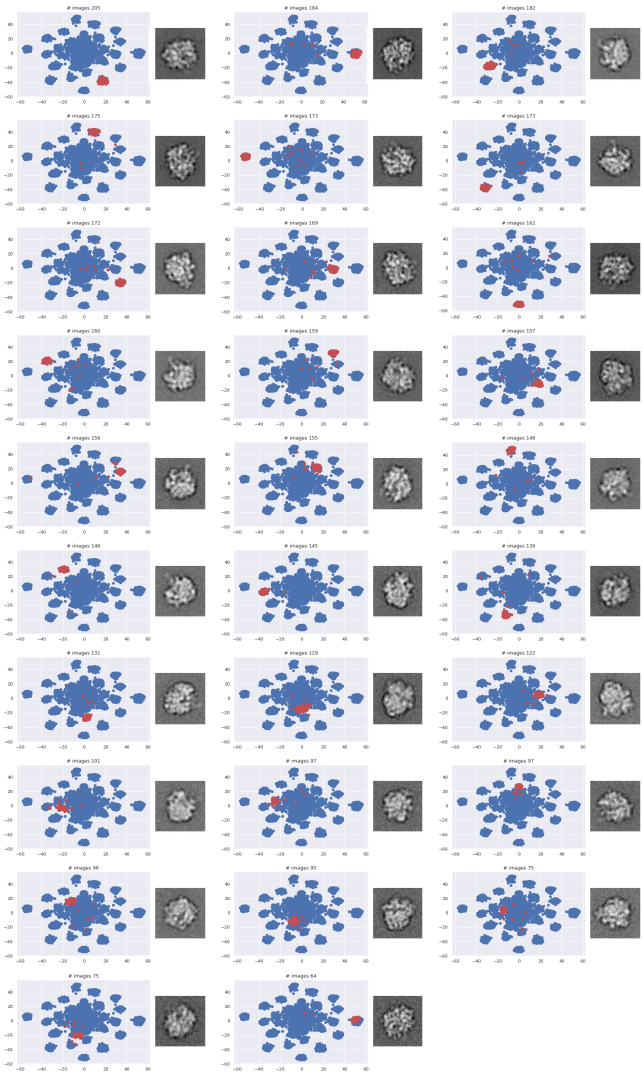
[63] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. Nature methods, 16(3):243–245, 2019.

[64] Pawel A Penczek. Image restoration in cryo-electron microscopy. Methods in enzymology, 482:35–72, 2010.

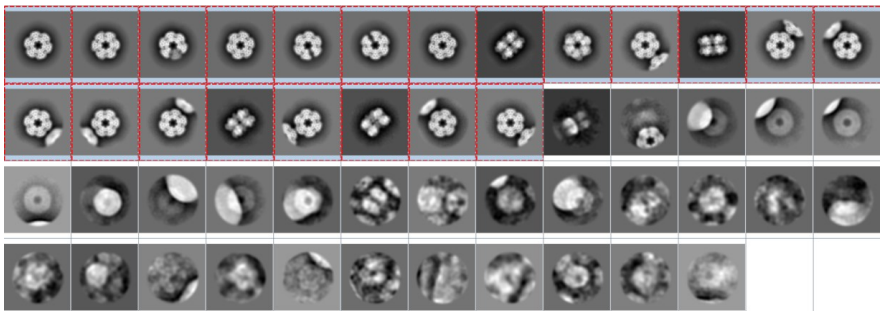
A Supplementary Figures



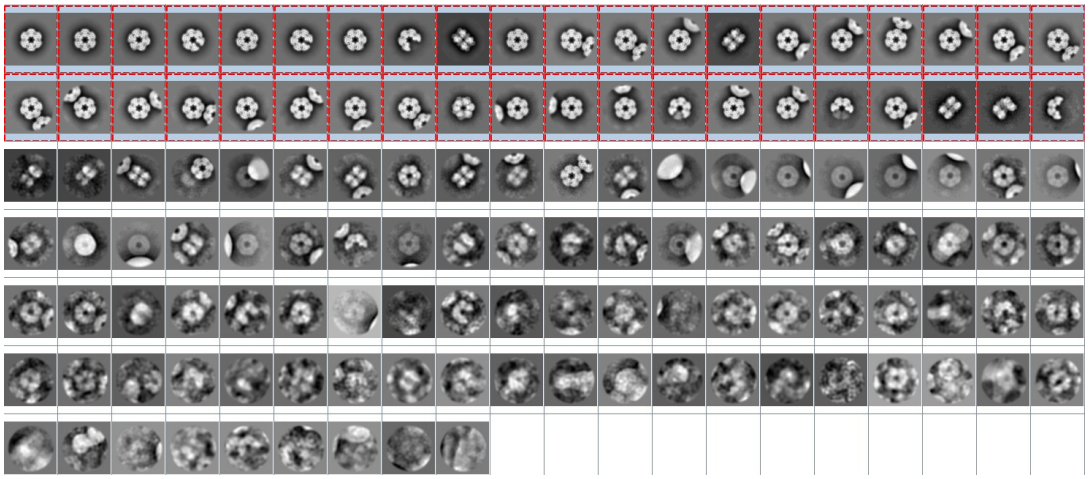
Supplementary Figure 1: The t-SNE plot of classification results of 70S ribosome by RE2DC with cluster size larger than 15. The red points denote the positions of data points that belong to the same group assigned by the classification algorithm.



Supplementary Figure 2: The t-SNE plot of classification results of 70S ribosome by ISAC with the number of members in each class set to 200. The red points denote the positions of data points that belong to the same group assigned by the classification algorithm.

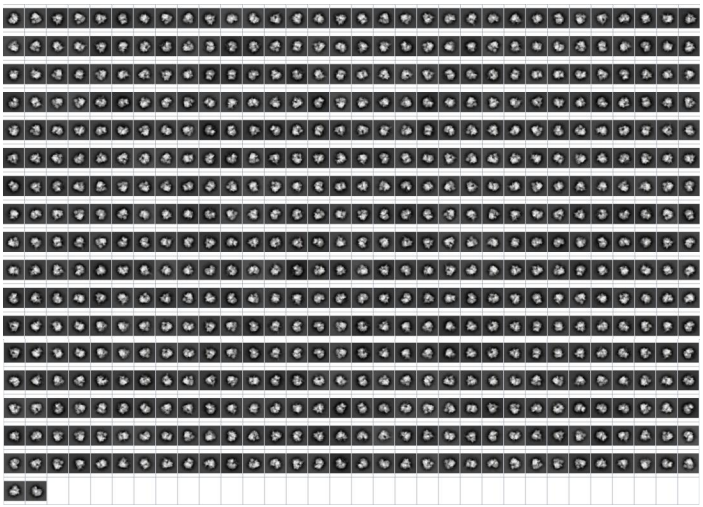


(a)

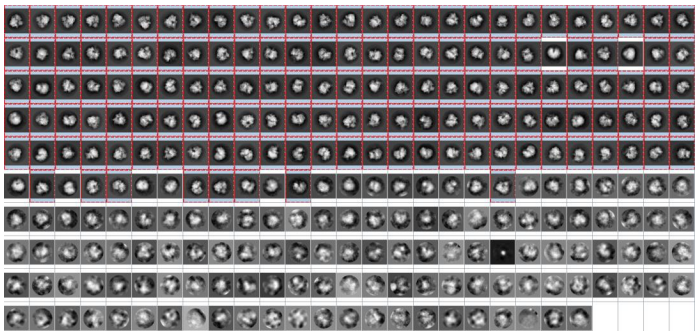


(b)

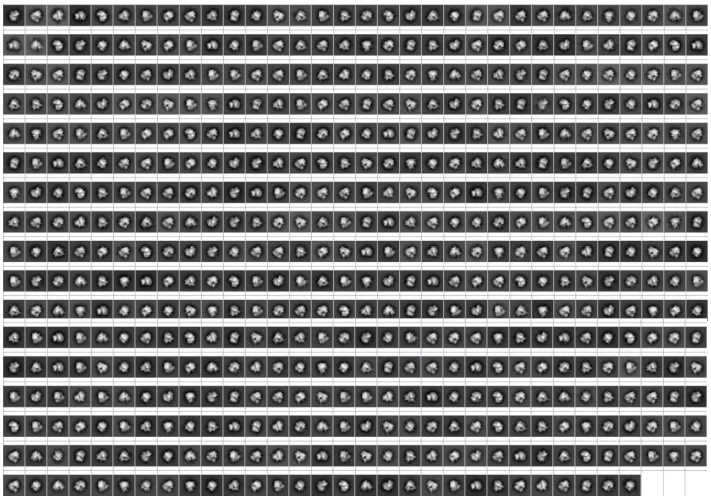
Supplementary Figure 3: Classification results of GS protein. (a) RELION with 50 classes. (b) RELION with 150 classes.



(a)

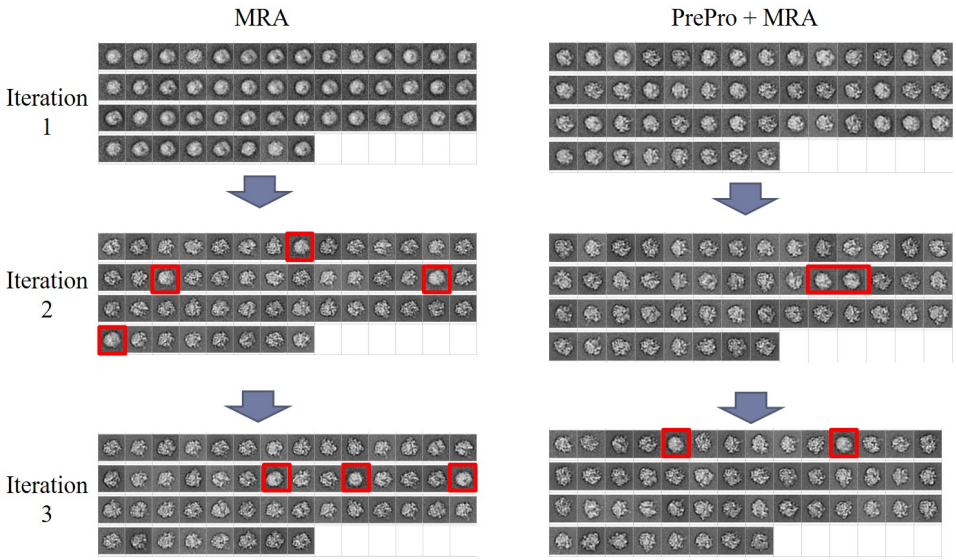


(b)

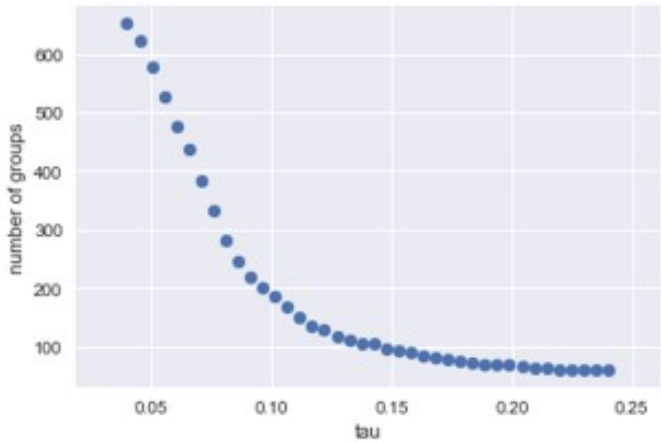


(c)

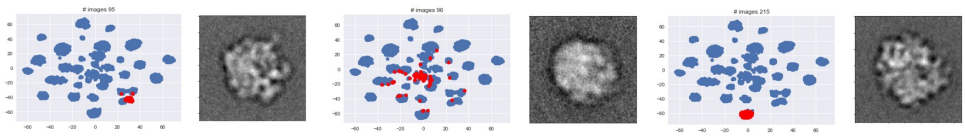
Supplementary Figure 4: Classification Results of 80S ribosome using (a) ISAC with class size set to 200. (b) RELION with the prescribed number of classes set to be 520, where the high-quality class averages are boxed in red. There are a total of 143 high-quality classes. (c) RE2DC with the least number per class set to 15.



Supplementary Figure 5: The classification results for MRA with and without PrePro on 70S ribosome dataset. The class averages surrounded by red blocks in the second and third iterations are considered to be unstable clusters. Note that by adding PrePro before MRA, different views of particles are already revealed in the first iteration.



Supplementary Figure 6: The number of clusters created by γ-SUP under various τ for the 80S ribosome dataset.



Supplementary Figure 7: Three representative t-SNE plots and class averages of the classification result.