The Dissertation Committee for Eric James Verbeke Certifies that this is the approved version of the following Dissertation:

Single Particle Electron Microscopy of Native Cell Extracts

Committee:

David Taylor, Supervisor

Edward Marcotte, Co-Supervisor

Daniel Dickinson

Jason McLellan

Andreas Matouschek

Single Particle Electron Microscopy of Native Cell Extracts

by

Eric James Verbeke

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2021

Acknowledgements

This work would not have been possible without the enormous support of family, friends, colleagues, collaborators and mentors, for which I am very grateful. In particular, I want to thank my advisors David and Edward for their patience, guidance, and willingness to let me pursue my interests.

Abstract

Single Particle Electron Microscopy of Native Cell Extracts

Eric James Verbeke, PhD

The University of Texas at Austin, 2021

Supervisors: David Taylor, Edward Marcotte

After the linking of genetic information to the biochemical composition of proteins in the mid 20th century, the emergent field of structural biology has focused on how the three-dimensional arrangement of atoms in a protein defines its cellular function. One rapidly evolving method for probing this structure-function relationship, and the focus of this work, is single particle transmission electron microscopy. Traditionally, single to few proteins of interest are first purified to near homogeneity from a biological source before structural characterization. However, a key advantage of electron microscopy over other methods is that proteins do not need to be purified, or in the case of electron tomography, even removed from the cell. The ability to study protein structure in as close to native conditions as possible can inform biology broadly. In this dissertation, I will present work towards expanding the use of single particle electron microscopy to native cell extracts. First, we explore a pilot analysis characterizing protein structures from chromatographically separated cell lysate guided by information from mass spectrometry. Extending on our initial studies, we next investigate protein structures from individual *Caenorhabditis elegans* embryos. We then introduce an image processing algorithm developed to assist single particle analysis of samples containing

multiple distinct protein structures. Finally, we demonstrate an application combining methods presented in this dissertation to investigate protein-protein interactions in red blood cells and their structural architectures.

Table of Contents

List of Figures

Chapter 1: Introduction

Part of this introduction is adapted from a joint perspective written by Caitlyn McCafferty and myself and is published as McCafferty, C.L., Verbeke, E.J., Marcotte, E.M., and Taylor, D.W. (2020). Structural Biology in the Multi-Omics Era. J. Chem. Inf. Model. 60, 2424–2429.

1.1. Structural Biology in a multi-omics era

With the sequencing of thousands of genomes, large biological data sets (-omics data) have become pervasive in most fields of biology, including development (Kumar et al., 2017; Wang et al., 2009), the classification of organisms (Joyce and Palsson, 2006; Raupach et al., 2016), and disease (Hasin et al., 2017; Karczewski and Snyder, 2018; Potter, 2018), among many others. Disciplines embracing -omics strategies reach well beyond the central dogma of biology—genomics, transcriptomics, and proteomics—into such areas as metabolomics (Riekeberg and Powers, 2017), epigenomics (Jones and Baylin, 2007), pharmacogenomics (Daly, 2017), and interactomics (Luck et al., 2017). As with these other endeavors, structural biology has also expanded to embrace -omics approaches.

Major historic interactions of structural biology and -omics approaches have included, for example, electron tomography (Lučić et al., 2005) to provide cellular context and spatial information to complement proteomics and interactomics data (Güell et al., 2009; Kühner et al., 2009; Yus et al., 2009), many efforts at proteome-scale modeling of three-dimensional (3D) structures and interactions (Aloy, 2004; Baker, 2001;

Vakser, 2014), and the entire field of structural genomics (Chandonia and Brenner, 2006; Kim, 1998; Skolnick et al., 2000; Stevens, 2001). Structural genomics has employed techniques such as X-ray crystallography, NMR spectroscopy, and electron microscopy (EM) to solve structures of purified macromolecules in a high-throughput manner, targeting new protein folds and entire proteomes, which have been supplemented by molecular modeling and structure prediction to extend structural insights to new molecules.

More recently, advances in single particle cryogenic electron microscopy (cryo-EM) have opened interesting new opportunities to connect -omics approaches and structural biology. In particular, cryo-EM boasts several important features: it requires only small amounts of sample, there is no requirement for crystal screening and optimization, and as a result, it is possible to capture several states of a macromolecular machine of interest. Cryo-EM is also capable of imaging a large field of individual macromolecular complexes in a single image. With the advent of direct electron detectors, ultrastable electron microscopes, automated data collection strategies (Li et al., 2020), and real-time data processing (Tegunov and Cramer, 2019), the "resolution revolution" in cryo-EM provides a definite route forward for increasing the throughput of structural biology (Kühlbrandt, 2014). We can anticipate that structures from these methods, in combination with electron tomography, will produce information-rich cell atlases capturing high-resolution structures of the proteome and its spatial context that will synergize with other -omics approaches. Here we focus specifically on efforts to increase the applicability of single-particle cryo-EM to increasingly complex and heterogeneous samples, approaching cell lysates in complexity (as in shotgun cryo-EM), thus furthering the transformation of cryo-EM into a pipeline for structural-omics.

1.2. INTRODUCTION TO DISSERTATION

This dissertation broadly encompasses approaches, challenges and results from applying single particle analysis to transmission electron microscopy data of native cell extracts. In Chapters 2 and 3, two biological samples are examined with low resolution negative stain EM in combination with other systems biology analyses as a justification for investigating native cell extracts. In Chapter 4, a model-independent approach is introduced for separating particles by structure prior to 3D classification in single particle cryo-EM data. Chapter 5 contains early results combining methods from the previous chapters using red blood cells as a model system. Finally, in Chapter 6, I summarize the arc of this dissertation and provide a brief outlook on single particle cryo-EM for native cell extracts.

1.3. PUBLISHED PAPERS

The following papers were published during my time at the University of Texas at Austin. These publications include many important contributions and collaborations from members of the labs of Edward Marcotte, David Taylor, Daniel Dickinson and Keith Keitz at the University of Texas at Austin. This dissertation focuses on the publications that are marked with an asterisk.

- * Verbeke, E.J., Mallam, A.L., Drew, K., Marcotte, E.M., and Taylor, D.W. (2018). Classification of Single Particles from Human Cell Extract Reveals Distinct Structures. Cell Reports 24, 259-268.e3.

- * Yi, X., Verbeke, E.J., Chang, Y., Dickinson, D.J., and Taylor, D.W. (2019). Electron microscopy snapshots of single particles from single cells. J. Biol. Chem. 294, 1602–1608.

- * Verbeke, E.J., Zhou, Y., Horton, A.P., Mallam, A.L., Taylor, D.W., and Marcotte, E.M. (2020). Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections. Journal of Structural Biology 209, 107416.

- * McCafferty, C.L., Verbeke, E.J., Marcotte, E.M., and Taylor, D.W. (2020). Structural Biology in the Multi-Omics Era. J. Chem. Inf. Model. 60, 2424–2429.

- Lucas, M.J., Pan, H.S., Verbeke, E.J., Webb, L.J., Taylor, D.W., and Keitz, B.K. (2020). Functionalized Mesoporous Silicas Direct Structural Polymorphism of Amyloid-β Fibrils. Langmuir 36, 7345–7355.

Chapter 2: Classification of single particles from human cell extract reveals distinct structures

As an alternative approach to structural proteomics, we demonstrate that single particle electron microscopy is a natural extension of co-fractionation mass spectrometry and can be used for direct visualization of molecular machines from native cell extracts. The work in this chapter was published as Verbeke, E.J., Mallam, A.L., Drew, K., Marcotte, E.M., and Taylor, D.W. (2018). Classification of Single Particles from Human Cell Extract Reveals Distinct Structures. Cell Reports 24, 259-268.e3. Anna Mallam was the lead on all biochemistry aspects of the paper, and I was the lead on processing and interpreting electron microscopy and mass spectrometry data. Kevin Drew assisted with bioinformatics analyses.

2.1. ABSTRACT

Multi-protein complexes are necessary for nearly all cellular processes, and understanding their structure is required for elucidating their function. Current high-resolution strategies in structural biology are effective but lag behind other fields (e.g., genomics and proteomics) due to their reliance on purified samples rather than heterogeneous mixtures. Here, we present a method combining single-particle analysis by electron microscopy with protein identification by mass spectrometry to structurally characterize macromolecular complexes from human cell extract. We identify HSP60 through two-dimensional classification and obtain three-dimensional structures of native proteasomes directly from *ab initio* classification of a heterogeneous mixture of protein complexes. In addition, we reveal an ~1-MDa-size structure of unknown composition and

reference our proteomics data to suggest possible identities. Our study shows the power of using a shotgun approach to electron microscopy (shotgun EM) when coupled with mass spectrometry as a tool to uncover the structures of macromolecular machines.

2.2. INTRODUCTION

Protein complexes play an integral role in all cellular processes. Understanding the structural architecture of these complexes allows direct investigation of how proteins interact within macromolecular machines and perform their function. In an effort to understand which proteins assemble into these machines, proteome-wide studies have been conducted to determine the composition of protein complexes (Drew et al., 2017a; Gavin et al., 2002; Havugimana et al., 2012; Hein et al., 2015; Ho et al., 2002; Huttlin et al., 2015, 2017; Kastritis et al., 2017; Kristensen et al., 2012; Krogan et al., 2006; Wan et al., 2015). Similar studies have identified direct contacts between protein complex subunits computationally (Drew et al., 2017b) or by cross-linking mass spectrometry (Leitner et al., 2016; Liu and Heck, 2015; Rappsilber et al., 2000), and although these studies provide insightful predictions on protein-protein interactions, they lack directly observable structural information that can inform us on function and subunit stoichiometry.

Structural genomics approaches, such as the Protein Structure Initiative, have thus far been the most successful way to systematically solve structures for proteins lacking a model (Chandonia and Brenner, 2006). These approaches have removed several bottleneck steps in traditional structural biology by applying high-throughput technology to sample preparation, data collection, and structure determination. Although many high-

resolution structures have resulted from structural genomics, these approaches typically miss large complexes and perform best on single proteins or low-molecular-weight complexes that can be purified and crystallized for X-ray crystallography or labeled for nuclear magnetic resonance (Montelione, 2012).

Recent advances in electron microscopy (EM) software and hardware have dramatically increased our ability to solve the structures of native protein complexes and allow for increased throughput approaches using EM. Automated microscopy software, such as Leginon (Suloway et al., 2005), SerialEM (Mastronarde, 2005), and EPU (FEI), allow for the collection of large datasets in a high-throughput, semi-supervised manner. RELION, a Bayesian algorithm for 3D classification, allows users to sort conformationally heterogeneous samples to define structurally homogeneous classes (Scheres, 2012). Furthermore, 3D reconstructions can now be done *ab initio* (without an initial model) by a computationally unsupervised approach using cryoSPARC (Punjani et al., 2017). These strategies potentially allow for analysis of heterogeneous mixtures, although this aspect has not been explored extensively.

Advances in hardware, such as direct electron detectors and Volta phase plates, allow visualization of particles at near atomic resolutions and smaller molecular weights, which was previously only possible for larger particles or particles with high symmetry (Danev and Baumeister, 2016; Kühlbrandt, 2014). Despite these revolutionary advances, single-particle EM is still largely used to study homogeneous samples, where the identity of the protein complex is known *a priori*.

Here, we take a different approach to structure determination by exploiting advances in EM software to structurally classify native protein complexes from human cell lysate. By using a shotgun approach to EM (shotgun EM), we chromatographically separate cell lysate into tractable fractions before identification by mass spectrometry (MS) and structural analysis by EM. Using this approach, we characterize compositionally and structurally heterogeneous protein complexes from immortalized (HEK293T) cells separated by macromolecular size using size-exclusion chromatography (SEC).

For this study, we determined the protein composition of two different high-molecular-weight samples from SEC by MS experiments. Identified proteins were then mapped to previously generated protein interaction networks to reveal candidate protein complexes. We then collected negative-stain EM data and performed single-particle analysis of heterogeneous particles simultaneously. Using this approach, we identified structurally distinctive macromolecular machines after unbiased 3D classification and *ab initio* reconstruction of single particles.

2.3. RESULTS

2.3.1. Separation and identification of subunits from high-molecular-weight protein complexes

Native macromolecular assemblies from lysed human cells were first separated by macromolecular size using SEC (see Methods). We selected a high-molecular-weight

fraction (fraction 4) for MS and EM analysis (Figure 1) with molecular weights in the range of 1.5 to 2 MDa based on molecular standards (Figure 2A; see Methods).

MS analysis of our sample (Figure 2A) identified 1,401 unique proteins. Over 93% of the identified proteins had a molecular weight under 200 kDa, indicating that the proteins are likely multi-subunit complexes in order to elute in the high-molecular-weight fraction. We then mapped the proteins identified by MS to a combined set of protein-protein interaction networks to suggest the identity of complexes in our sample (Figure 2B). The previously determined protein-protein interaction networks include hu.MAP (Drew et al., 2017a) and CORUM (Ruepp et al., 2010), which were chosen to provide a list of documented and high-confidence protein complexes. Furthermore, hu.MAP incorporates datasets from previous interactome studies (Havugimana et al., 2012; Hein et al., 2015; Huttlin et al., 2015; Wan et al., 2015) and includes greater than 4,000 complexes. In addition, we incorporated interaction networks that exclusively used size-exclusion chromatography and quantitative proteomics to determine protein-protein interactions (Kristensen et al., 2012; Larance et al., 2016). The combined protein interaction network included 7,021 protein complexes. We identified specific, well-annotated protein complexes within our sample, which contains both structurally defined complexes (e.g., the proteasome; Lander et al., 2012; Schweitzer et al., 2016) and complexes without known structures (e.g., the multi-tRNA synthetase complex; Mirande, 2017; Figure 2B).

Complexes with at least 50% of their subunits identified were kept as candidates for subsequent analysis. Many of the resulting candidate complexes shared a number of individual subunits and are different variants of the same complex. In order to group

related complexes, we created a hierarchical network by performing an all-by-all comparison of proteins between each complex (Figure S1; see Methods). Our hierarchies suggest we have 234 groups of related complexes (i.e., with shared subunits) in addition to the remaining 538 unique complexes for a total of 772 complexes in our sample (Table S1).

The abundance of each complex was then calculated using two different label-free quantification strategies to rank the predicted complexes that might be visible by EM. Both normalized spectral counting (Vaudel et al., 2015) and top 3 extracted ion chromatogram areas (Silva et al., 2006; see Methods) produced similar abundance values for each protein complex (Figure S1). By combining our hierarchical network with the relative abundance for each complex, we identified the specific subunit composition of complexes most likely to be present in our sample. As an example, we can examine the group of related proteasome complexes (Figure S1), showing many related complexes, where the canonical 26S proteasome appears to be the most abundant form. This analysis reveals complexes of interest in our sample, which vary in abundance.

2.3.2. EM of single particles from HEK293T cell extract fraction

Having identified candidate complexes in our sample by MS, we next use negative-stain EM to investigate the structures of the complexes. Negative-stain EM samples are easily prepared and are often used to determine the heterogeneity of a sample because of the higher signal-to-noise ratio compared to cryo-EM. Raw micrographs of our negatively stained sample show monodisperse particles with clear structural features (Figure 3A). Intact, structurally heterogeneous complexes can be directly observed. The

proteasome can be seen in three different structural states, as a core (20S), as a single-capped proteasome (20S core with one 19S regulatory particle), and as a double-capped proteasome (26S, 20S core with two 19S regulatory particles). In addition, many other unidentified particles can be clearly seen, with an average particle diameter of ~200 Å.

Template picking from 1,250 micrographs of our sample resulted in a final set of 31,731 particles after filtering out ~67% of particles as "junk" particles (see Methods). To assess the quality of automated template picking, we also manually selected 35,381 particles for alignment and classification. A comparison of the reference-free 2D class averages of both manually and template-picked datasets yielded similar results (Figure S2), and both datasets were used for independent downstream processing. 2D class averages yielded distinct class averages with various morphologies and features. Remarkably, many well-defined classes emerged from this heterogeneous mixture of complexes (Figure 3B).

Interestingly, we observed two distinct heptameric rings in our reference-free 2D classification (Figures 3C and 3D). One of the rings is wider in diameter with a pinwheel-like architecture (Figure 3C), and the second is rounder and narrower (Figure 3D). To uncover the identity of these rings, we turned to our mass spectrometry data for candidate ring-forming complexes. Two of the identified complexes, heat shock protein 60 (HSP60) and the α and β rings of the proteasome core, are known to form heptameric rings. The X-ray crystal structures of both HSP60 and the proteasome core were used to compare to our candidate structures. HSP60 is 135 Å in diameter (PDB: 4PJ1; Nisemblat et al., 2015), and the ring of the 20S core (PDB: 4R3O; Harshbarger et al., 2015) is 115 Å in diameter, which suggested an identity for each of the rings by a comparison of

diameters. To test this hypothesis, we reprojected the X-ray crystal structure of both protein complexes after low-pass filtering to 30-Å resolution to simulate 2D projections and compared them to our class averages. Finally, we compared reference-free class averages of purified GroEL (Danziger et al., 2003; a well-studied HSP60 homolog) and proteasome core to our fractionation data. All of these comparisons provide strong evidence that the pinwheel-like and narrow ring projections correspond to HSP60 and the proteasome core, respectively.

To further validate our identification of HSP60, we performed negative-stain EM on a second fraction from our SEC, fraction 8, where HSP60 was also identified by mass spectrometry. The approximate molecular weight of native macromolecular assemblies in fraction 8 is 500 kDa (Figure 2A). For particle selection of fraction 8 EM data, we used a difference-of-Gaussian picker (Voss et al., 2009). This method was chosen as an orthogonal, reference-free method to independently confirm whether we could identify HSP60. Reference-free 2D class averages obtained using this particle-picking scheme revealed a class average with a well-defined pinwheel-like architecture (Figure S2), suggesting HSP60 was also identified in fraction 8.

2.3.3. 3D classification of a heterogeneous mixture produces distinct structures

Given the success of 2D classification at separating particles into distinct classes, we then performed 3D classification on the entire set of particles using RELION (Scheres, 2012) to simultaneously generate 30 reconstructions (Figure 4A). Whereas RELION was developed to group 2D projections of the same protein or protein complex with conformational heterogeneity into distinct classes, we asked whether RELION could

also classify projections from many distinct complexes in a heterogeneous mixture into internally consistent (low-error) reconstructions.

To test the internal consistency of the 3D reconstructions, we determined the distribution of calculated error within the models and ranked each reconstruction based on a rotational-translational error score (see Methods). The error score distribution was then compared to the rotational-translational error scores of models built from random particles in the dataset to evaluate our ability to classify related particles belonging to a particular model and demonstrated our 3D reconstructions have substantially less error than random reconstructions (Figure 4B). The 30 3D reconstructions generated all contained various degrees of structural details ranging from distinct barrels to more globular shapes (Figure 4C), suggesting it is possible to classify particles from a heterogeneous mixture into distinct structures.

We then performed cross-correlations between our top 3 models and several complexes with known structure from our MS-determined list of high-abundance complexes to determine whether we could link our structural models with complex identity (Figure S3; see Methods). The 20S proteasome emerges as a clear match when compared to our highest scoring model with a cross-correlation score of 0.87. We were also able to distinguish a single-capped proteasome, which matched to our third highest scoring model with a cross-correlation score of 0.81. Interestingly, our second highest scoring model was not readily recognizable, and none of the known structures emerged as a clear match after cross-correlation. Based on the high-abundance 2D class averages and large volume of the unknown complex, we filtered our proteomics data to search for possible identities. Our search suggests the unknown complex is likely a variant of a

mitochondrial ribosome, spliceosome, or DNA-repair complex, but given the current resolution, the results are inconclusive. A much larger set of particles or projections and deeper classification is likely required for assignment of this structure. However, our results suggest it is possible to solve multiple structures from cell lysate in a parallel manner, even in the absence of matching starting models.

2.3.4. Quantification and *ab initio* reconstruction of the proteasome

To determine our ability to further characterize complexes identified in a complex mixture, we investigated our sample specifically in the context of the proteasome, which allowed us to evaluate the success of reconstructions without an initial model. Our goals were to (1) investigate whether *ab initio* reconstructions would reveal clear proteasome structures, (2) determine the ratio of the 20S core and single-capped proteasomes using our single-particle data, and (3) compare single-particle counting of the proteasome to label-free MS quantification.

Class averages of the 20S core and single-capped proteasomes were clearly identified as barrel-shaped particles and barrels with large rectangular caps, respectively (Figure 5A). Based on identifying the proteasome with notably distinct 2D class averages, as well as RELION-based 3D classification producing two identifiable proteasome models, we asked whether *ab initio* reconstructions were capable of correctly recovering proteasome structures. We therefore attempted a completely unsupervised approach for 3D classification using cryoSPARC (Punjani et al., 2017). cryoSPARC was developed for determining multiple 3D structures of a protein without prior structural knowledge or the assumption that the ensemble of conformations resembled each other,

but in this context, we evaluated its ability to classify 2D particles of distinct complexes in a mixture. Remarkably, a 3D reconstruction of the 20S core was generated using *ab initio* reconstruction in cryoSPARC on the entire dataset of particles with 5, 10, and 15 classes (Figure S4).

From the structures generated with 10 classes, a distinct 3D reconstruction of the 20S core showing a clear barrel with a central channel and some separation of co-axial rings was produced (Figure 5B). This 20S core reconstruction contains 3,150 particles with an estimated resolution of 20.4 Å using the 0.143 Fourier shell correlation (FSC) criterion (Figure S4). Our 3D map is consistent with a recent high-resolution structure of the 20S core (EMD-2981; da Fonseca and Morris, 2015) with a cross-correlation score of 0.94.

We were unable to distinguish a 3D structure of the single-capped proteasome from cryoSPARC. However, going back to our single-capped proteasome from 3D classification using RELION, we were able to dock in a high-resolution structure determined previously (EMD-4002; Schweitzer et al., 2016; Figure 5B). The high-resolution structure can be unambiguously docked into our EM density (cross-correlation score of 0.76) albeit with less agreement given the low number of particles in the model (1,121 particles). Using RELION to refine the structure of our single-capped proteasome, we achieved a nominal resolution of 31 Å (Figure S4).

We then quantified the ratio of 20S core to single-capped proteasome particles by directly counting individual particles from our EM data of fractionated cell lysate. Revisiting our 2D classification, we compared the number of particles aligned in the side

view of the 20S core and single-capped proteasome (Figure 5A). The ratio of 20S core to single-capped proteasome particles in our sample was calculated to be 3:2 or 1 bound 19S regulatory particles for every 2.5 20S core particles in our sample by EM. This is similar to our MS data, which suggest the ratio of 19S regulatory particles to 20S core particles is 1:1 (Figure S5). Collectively, our study suggests it is not only possible to solve structures of protein complexes from cell lysate *ab initio* but also quantify the stoichiometry of biochemical states.

2.4. DISCUSSION

One bottleneck of structural biology is the current limitation of studying only a single protein or protein complex structure in a single experiment. However, recent advances in detectors and software for EM bring about the possibility of high-throughput structural determination using EM. To this end, we have demonstrated shotgun EM as a potential pipeline for high-throughput identification and structural determination of macromolecular machines. By combining MS and EM, we demonstrate it is possible to structurally characterize and identify protein complexes from a cellular sample containing many native complexes. This pipeline was used to successfully identify the proteasome in two biochemical forms and HSP60 from a cellular fraction with minimal user input. HSP60 was then independently verified through another SEC fraction identified as containing HSP60 by MS. Additionally, we construct a self-consistent structural model of an ~1-MDa protein complex of unknown identity.

A recent study showed that higher order assemblies from a eukaryotic thermophile could be separated chromatographically, identified by MS, and visualized

through cryo-EM to obtain a high-resolution structure (Kastritis et al., 2017). The authors performed cryo-EM on particles from a complex mixture to solve a 4.7-Å-resolution structure of fatty acid synthase from cell lysate separated by molecular size after a 50% enrichment for fatty acid synthase. In our study using human cells, which have a canonical proteome approximately 3 times larger than C. thermophilum, we are able to obtain structural information from a complex mixture without enrichment, suggesting that sample heterogeneity is a surmountable problem. A combined approach using shotgun EM and the cryo-EM protocol presented by Kastritis et al. (2017) provides a potential strategy for recovering multiple high-resolution structures from fractionated cellular extracts.

Several key barriers to structurally classifying heterogeneous mixtures remain, with the main challenge being to correctly assign different orientations of the same complex in large datasets of heterogeneous mixtures. Additionally, assigning the correct subunit composition to the unidentified molecular models (UMMs) uncovered using shotgun EM, particularly for complexes lacking structural information, will present a unique challenge to structural biology. Whereas currently we cannot identify each class average or 3D structure obtained in this study, we are able to distinguish different structural states of the proteasome using current *ab initio* methods, suggesting that shotgun EM is a promising tool to characterize the heterogeneity of protein complex forms. Our top-scoring UMM was not readily recognizable and had no apparent match from model fitting. It is possible our model has been structurally annotated previously but was not covered in our search. Alternatively, it is possible our model remains unidentified because it is structurally novel. In future experiments, a comprehensive list

of solved structures coupled with optimal volume alignment and cross-correlation can be used to identify likely matches to models generated using shotgun EM.

One challenge when dealing with protein complexes is defining their precise subunits. MS does not indicate which complex a protein belonging to multiple complexes was identified from. Many of these related complexes and sub-complexes have yet to be structurally or biochemically characterized. Our hierarchical network strategy allows us to make an initial estimate on which form of a complex might be in our EM data. Using shotgun EM, we aim to validate these uncharacterized and other less-characterized forms of complexes that may be more amenable to our separation scheme.

A key proof of concept in this study was the proteasome, which is a structurally distinct complex and serves a crucial role in protein degradation in eukaryotic cells (Finley, 2009). The native stoichiometry of the proteasome has been studied in different ways by multiple groups (Asano et al., 2015; Havugimana et al., 2012). Our template-picked counting of single proteasome particles has an advantage over MS approaches by identifying which form of a complex an identified protein belongs to. Although our MS and EM quantification were similar, showing an approximate ratio of 20S core to 19S regulatory particles ranging from 1:1 to 2:1, a separate study using corrected spectral counts suggests the ratio is closer to 4:1 (Havugimana et al., 2012). To reconcile these two observations, more chromatographic fractions containing the proteasome would need to be quantified by EM and MS to see whether there is agreement. As more protein complexes become structurally annotated, shotgun EM can be used as an auxiliary method for quantifying the abundances of native complexes, as well as their stoichiometry.

After *ab initio* 3D classification, we obtained a reasonable reconstruction of the 20S core in cryoSPARC from 3,150 particles. Although only half of these particles are accounted for from 2D class averaging of all particles, it is likely that the discrepancy results from proteasome particles that are misclassified or exist in different, less-populated orientations in our 2D class averages. Alternatively, because the number of models we could reconstruct in 3D was limited by the small populations of each complex we had in our micrographs, it is possible that non-proteasome particles were grouped into our 3D class of the proteasome. These misclassified particles would have a small contribution to the overall likelihood of the 3D map as it is reconstructed (Punjani et al., 2017). One method to separate misclassified particles would be to do iterative rounds of 3D classification.

In this study, we used a 60S ribosome class average as a template for auto-picking due to its large molecular weight and round shape. Interestingly, none of the resulting averages resembled the 60S, providing evidence that we were not biasing the results from template picking and subsequent data analysis. A similar concern for model bias exists when using RELION to generate 3D models. Despite this, none of the 3D classes are visually identical to the reference 3D model, with most EMD structures selected from our MS data outscoring the reference model by cross-correlation score when compared to our top 3 RELION models. In future experiments, more sophisticated template matching, deep learning algorithms, or *ab initio* methods can be introduced to improve particle identification and model building (Punjani et al., 2017; Rickgauer et al., 2017; Wang et al., 2016).

This study represents an advance into structural proteomics using EM, suggesting that parallel structural determination of protein complexes shows promise for alleviating bottlenecks in structural biology. In the interim before high-resolution data are collected, it is possible to search for structurally uncharacterized complexes through the addition of protein tags (Flemming et al., 2010) to identify complexes in a heterogeneous mix without the need to purify the sample. One could also utilize integrative structural biology approaches to have a predicted model with which to search for structures in cell extract. We envision using cryo-EM for this pipeline to solve sub-nanometer-resolution structures, where homology models and known structures can be more clearly compared. Moving this pipeline to cryo-EM will likely aid in our identification of candidate complexes; however, several obstacles will need to be overcome, including (1) lower signal-to-noise ratio, (2) complex instability (i.e., protein complexes being degraded into non-native compositions), and (3) the increased amount of data required for reconstructions. Future studies will be required to determine whether we can overcome these potential pitfalls when transitioning the pipeline into cryo-EM.

Shotgun EM will accelerate the pace at which structural information is generated and allow us to better understand the structure-function relationship of proteins. Optimization of this technique has the potential to address questions about many macromolecular machines across different cell types, disease states, and species. We propose that investigating the collective protein complexes in a cell, or the "complexome," using shotgun cryo-EM will help inform us broadly on systems biology, cell biology, and changes in complexes that contribute to human diseases.

2.5. METHODS

2.5.1. Cell culture and extract preparation

HEK293T cells were harvested at 80%–100% confluence without trypsin by washing in ice cold phosphate buffered saline (PBS) pH 7.2 (0.75 mL; GIBCO) and placed on ice. Cells (approximately 10 mg) were lysed on ice (5 min) by resuspension in Pierce IP Lysis Buffer (0.8 mL; 25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol; Thermo Fisher) containing 1x protease inhibitor cocktail III (Calbiochem). The resulting lysate was clarified (17,000 g, 10 min, 4°C) and filtered (Ultrafree-MC filter unit (Millipore); 12,000 g, 2 min, 4°C).

2.5.2. Biochemical fractionation using native size-exclusion chromatography

Size-exclusion chromatography (SEC) was performed at 4°C on an AKTA FPLC (GE Healthcare). Approximately 6 mg of soluble protein was applied to a Superdex 200 10/300 GL analytical gel filtration column (GE Healthcare) equilibrated in PBS, pH 7.2 at a flow rate of 0.5 mL min-1. Fractions were collected every 0.5 mL. The elution volumes of molecular weight standards (Thyroglobulin, 670,000 Da; γ-globulin, 158,000 Da; Ovalbumin, 44,000 Da; Myoglobin, 17,000 Da; Vitamin B12, 1,350 Da; Biorad) were additionally measured to calibrate the column (Figure 2A). Fraction 4 (concentration ~1 mg/mL) was deemed most likely to contain a high number of large complexes, as determined by A_{280}, and was subjected to further proteomic and structural analysis.

2.5.3. Mass Spectrometry

50 µL of Fraction 4 (Figure 2A) was denatured and reduced in 50% 2,2,2-trifluoroethanol (TFE) and 5 mM tris(2-carboxyethyl)phosphine (TCEP) at 55°C for 45 minutes, followed by alkylation in the dark with iodoacetamide (55 mM, 30 min, RT). Samples were diluted to 5% TFE in 50 mM Tris-HCl, pH 8.0, 2 mM CaCl2, and digested with trypsin (1:50; proteomic grade; 5 hours: 37°C). Digestion was quenched (1% formic acid), and the sample volume reduced to ~100 µL by speed vacuum centrifugation. The sample was washed on a HyperSep C18 SpinTip (Thermo Fisher), eluted, reduced to near dryness by speed vacuum centrifugation, and resuspended in 5% acetonitrile/ 0.1% formic acid for analysis by liquid chromatography tandem mass spectrometry (LC-MS/MS). Peptides were separated on a 75 µM x 25 cm Acclaim PepMap100 C-18 column (Thermo) using a 3%–45% acetonitrile gradient over 60 min and analyzed on line by nanoelectrospray-ionization tandem mass spectrometry on an Orbitrap Fusion (Thermo Scientific). Data-dependent acquisition was activated, with parent ion (MS1) scans collected at high-resolution (120,000). Ions with charge 1 were selected for collision-induced dissociation fragmentation spectrum acquisition (MS2) in the ion trap, using a Top Speed acquisition time of 3 s. Dynamic exclusion was activated, with a 60 s exclusion time for ions selected more than once.

2.5.4. Proteomic and bioinformatics analyses

The mass spectrometry data were processed independently using searchGUI and PeptideShaker (Vaudel et al., 2011, 2015) and Proteome Discoverer (ThermoFisher Scientific). Data were searched against a target-decoy human database downloaded from

Universal Protein Resources Database (UniProtKB/Swiss-Prot comprising human proteins supplemented with common contaminants). Fixed modifications of carboxyamidomethylated cysteine and variable modifications of oxidized methionine and acetylation of protein N terminus were permitted to allow for detection of modified peptides. Peptide spectral matches, peptides and proteins were considered positively identified if detected within a 1% false discovery rate cut off (based on empirical target-decoy database search results). Additionally, proteins were only considered for further processing if at least one unique peptide was identified. This screening procedure resulted in 1,402 distinct human proteins. To facilitate mapping to a protein ID, we used UniProtKB accession numbers as a common identifier and the UniProt ID mapping tool to interconvert different gene and protein identifiers.

Relative abundance for each complex was determined using two different methods of label-free quantification, one calculated using peptide spectral matches and the other calculated using extracted ion chromatogram area (XIC). Protein length was used for normalizing the number of peptide spectral matches observed for each protein using the Normalized Spectral Abundance Factor (NSAF) as calculated by PeptideShaker (Vaudel et al., 2015). Proteins expected to participate in a complex as predicted by our combined protein interaction network, which were not identified by MS, were assigned a NSAF value of zero. The NSAF values for all proteins in a complex were then averaged to estimate the relative abundance of each complex.

To calculate relative abundance based on XIC, each protein was assigned an abundance by taking the average of the top-3 peptide areas identified for that protein using Proteome Discoverer (ThermoFisher Scientific). Proteins expected to participate in

a complex as predicted by our combined protein interaction network, which were not identified by MS, were assigned an abundance of zero. The average area values for all proteins in a complex were then averaged to estimate the relative abundance of each complex.

The hierarchical network of protein complexes in Figure S1 was created by determining the percent of shared subunits between all complexes. For a predicted protein complex A with subunits $\{a_1, a_2, \dots, a_n\}$ and B with subunits $\{b_1, b_2, \dots, b_n\}$, the similarity score (S) of A to B was calculated by finding the intersection of A and B divided by the size of set A as follows.

$$S = \frac{|A \cap B|}{|A|}$$

If the similarity score between complexes was 90% or greater, it was considered a related complex. The resulting network shows related groups of complexes where at least 90% of subunits in higher-order complexes are shared between sub-complexes. 837 of the 1375 complexes identified by MS belong to a group of shared complexes. Furthermore, the 837 shared complexes in our sample can be organized into 234 distinct hierarchies. The network of related complexes was then visualized using Cytoscape with edges corresponding to the similarity score (Shannon et al., 2003).

2.5.5. Negative stain electron microscopy sample preparation

4 μL of fractionated human cell lysate was applied to a glow-discharged 400-mesh continuous carbon grid. After a 1 min adsorption, the sample was negatively stained with five consecutive droplets of 2% (w/v) uranyl acetate solution, blotted to remove residual stain, and air-dried in a fume hood.

2.5.6. Electron Microscopy

Data was acquired using a JEOL 2010F transmission electron microscope operated at 200 keV with a nominal magnification of x60,000 (3.6 Å at the specimen level). Each image was acquired using a 1 s exposure time with a total dose of ~30-35 e^{-} $Å^{-2}$ and a defocus between –1 and –2 μm. A total of 1,250 micrographs were manually recorded on a Gatan OneView.

2.5.7. 3D reconstruction and analysis

Two independent particle stacks were generated from the same 1,250 micrographs using either template or manual particle picking. The contrast transfer function (CTF) of each micrograph was estimated using CTFFIND4 (Rohou and Grigorieff, 2015). FindEM (Roseman, 2004) was used for template-based particle picking using a reference-free 2D class average of our negatively stained 60S Ribosome from *Saccharomyces cerevisiae* (a gift from A. Johnson). We chose this template for particle picking as it picked virtually all particles in each micrograph. It would also be easily recognizable in class averages if there were a template bias. Importantly, none of the resulting class averages matched this

ribosome. ~97,000 and ~37,000 particles were selected by template picking and manually selecting particle images, respectively. All image pre-processing was done in Appion (Lander et al., 2009). After removing junk particles, 31,731 particles were left from template picking and 35,381 particles from manual picking, respectively. The majority of junk classes from template picking can be attributed to the picking of particles within aggregates and two particles as one. Particle box size was set to 576 Å x 576 Å. For our second fraction analyzed by EM (fraction 8), particles were selected in an automated manner using a Difference of Gaussian (DoG) particle picker (Voss et al., 2009). ~75,000 particles were picked from 300 micrographs. Junk particles were filtered from the dataset resulting in a final set of 28,553 particles. Particle box size was set to 518.4 Å x 518.4 Å.

Reference-free 2D class averages were generated with 300 classes for both fraction 4 and fraction 8 datasets using RELION (Scheres, 2012). Next, 3D classification was performed on fraction 4 data using RELION to create 30 classes of both datasets. The structure of DNA-dependent protein kinase catalytic subunit was chosen as an initial model using a negative stain structure low-pass filtered to 60 Å as a starting model (Sibanda et al., 2017) (Figure S3). Autorefine in RELION was used to refine the putative single-capped 26S proteasome structure from the manually-picked dataset using the corresponding class reconstruction low-pass filtered to 60 Å as a starting model. The manual picked dataset was used for subsequent analysis using cryoSPARC (Punjani et al., 2017). cryoSPARC was used to *ab initio* reconstruct 5, 10 and 15 3D models. The class corresponding to the 20S proteasome from the 10-model run, containing 3,150 particles, was then subjected to homogeneous refinement using cryoSPARC.

Random particle models were generated using RELION with the template picked particle dataset. Each model was reconstructed using the mean number of particles from the 30 models in Figure 4, ~1000 particles. Particles were sampled without replacement. Model error (E) was calculated for each RELION generated model by taking the harmonic mean of their respective rotational accuracy (R) and translational accuracy (T) as determined using RELION. Model error values were normalized between 1 and 2.

$$E = \frac{2}{\left(\frac{1}{R}\right) + \left(\frac{1}{T}\right)}$$

We then performed a two-sided Kolmogorov-Smirnov test between the distribution of model error from our models and the distribution of model error from the random particle models.

Several high-abundance complexes from our MS data with identifiable, previously solved structures were used to compare with our top 3 models generated using RELION. All models were first low-pass filtered to 30 Å before being aligned using Chimera's Fit in Map function (Pettersen et al., 2004). The cross-correlation score was then calculated by using the model with a larger volume as the region of computation, essentially sliding the larger complex across the smaller complex.

Purified proteasomes (a gift from A. Matouschek and C. Davis) were prepared as described above. 80 micrographs were manually recorded and processed using reference-free 2D alignment and classification in RELION.

2.5.8. Quantification and statistical analysis

The statistical tests and associated p values are reported in the figures and/or figure legends for the specific analysis. Distributions of the rotational-translational error for the reconstructed 3D models were compared using a two-sided Kolmogorov-Smirnov test (Figure 4B). For the comparison of the two label-free quantification strategies, each point represents the relative abundance of a given protein complex determined using the two methods (Figure S1B). The Pearson correlation coefficient was then calculated for the resulting data.

2.5.9. Data and software availability

The EM reconstruction for both the 20S and 26S (presented in Figure 5B) were deposited in the EM Data Bank (EMDB) under accession codes EMD-7946, EMD-7947, respectively. The accession number for the MS data reported in this paper is PRIDE: PXD010026.

2.6. FIGURES

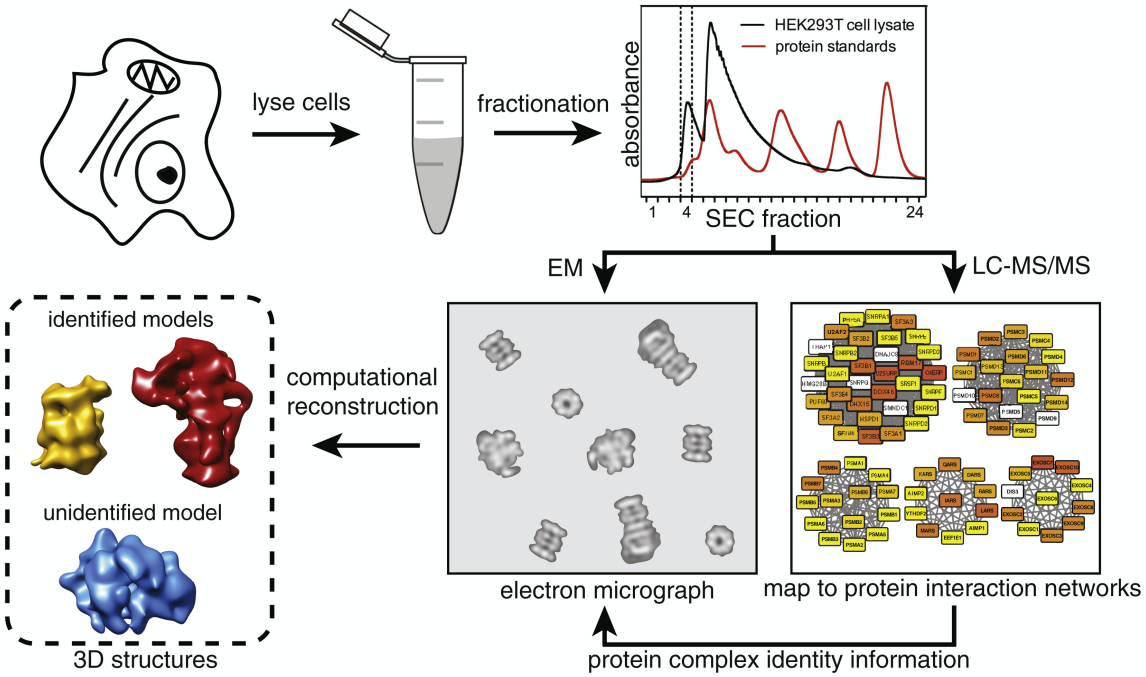

Figure 2.1. Shotgun EM pipeline used for structural determination of multiple macromolecular complexes.

HEK293T cells are subjected to lysis and separation using SEC. The resulting fractions are characterized separately by electron microscopy and mass spectrometry. Proteins identified from mass spectrometry are mapped to known and predicted protein complexes to identify which complexes are present in a given fraction. Electron microscopy data are then used to generate structures of multiple protein complexes.

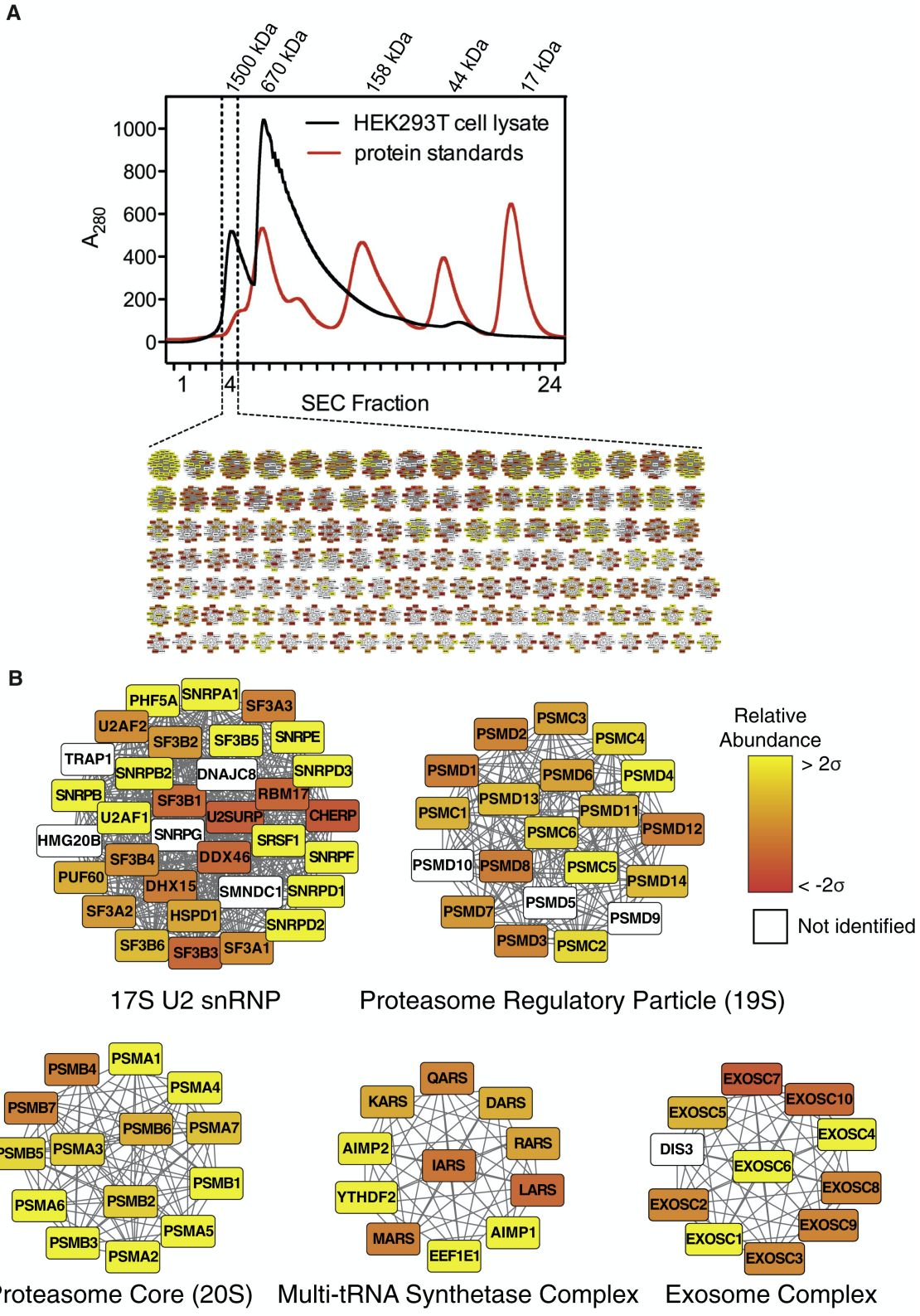
A
1500 kDa
670 kDa
158 kDa
44 kDa
17 kDa
HEK293T cell lysate
protein standards
A280
SEC Fraction
B
PHF5A
SNRPA1
SF3A3
U2AF2
SF3B2
SF3B5
SNRPE
TRAP1
SNRPB2
DNAJC8
SNRPD3
SNRPB
SF3B1
RBM17
U2AF1
U2SURP
CHERP
HMG20B
SNRPG
SRSF1
SF3B4
DDX46
SNRPF
PUF60
DHX15
SMNDC1
SNRPD1
SF3A2
HSPD1
SNRPD2
SF3B6
SF3B3
SF3A1
17S U2 snRNP
PSMC3
PSMD2
PSMC4
PSMD1
PSMD6
PSMD4
PSMC1
PSMD13
PSMD11
PSMC6
PSMD12
PSMD10
PSMD8
PSMC5
PSMD14
PSMD7
PSMD5
PSMD9
PSMD3
PSMC2
Proteasome Regulatory Particle (19S)
Relative Abundance
> 2σ
< -2σ
Not identified
PSMB4
PSMA1
PSMA4
PSMB7
PSMB6
PSMA7
PSMB5
PSMA3
PSMB1
PSMA6
PSMB2
PSMA5
PSMB3
PSMA2
Proteasome Core (20S)
QARS
KARS
DARS
AIMP2
RARS
IARS
YTHDF2
LARS
MARS
AIMP1
EEF1E1
Multi-tRNA Synthetase Complex
EXOSC7
EXOSC10
EXOSC5
EXOSC4
DIS3
EXOSC6
EXOSC8
EXOSC2
EXOSC9
EXOSC1
EXOSC3
Exosome Complex

Figure 2.2. Identification of protein complexes in a cellular fraction.

(A) Elution profile from SEC. Elution profiles of protein standards are overlaid to estimate the molecular weight range of protein complexes in fraction 4. Inset: a network map displaying a portion of the 1,375 candidate complexes determined by mapping mass spectrometry data to combined protein interaction networks is shown. (B) Enlarged view of a subset of candidate complexes. A filled node indicates a protein was identified by mass spectrometry; a white node indicates the protein was not identified. Color gradation of filled nodes indicates the relative abundance (determined by label-free quantification) ranging from ±2 SDs. See also Figure S1 and Table S1.

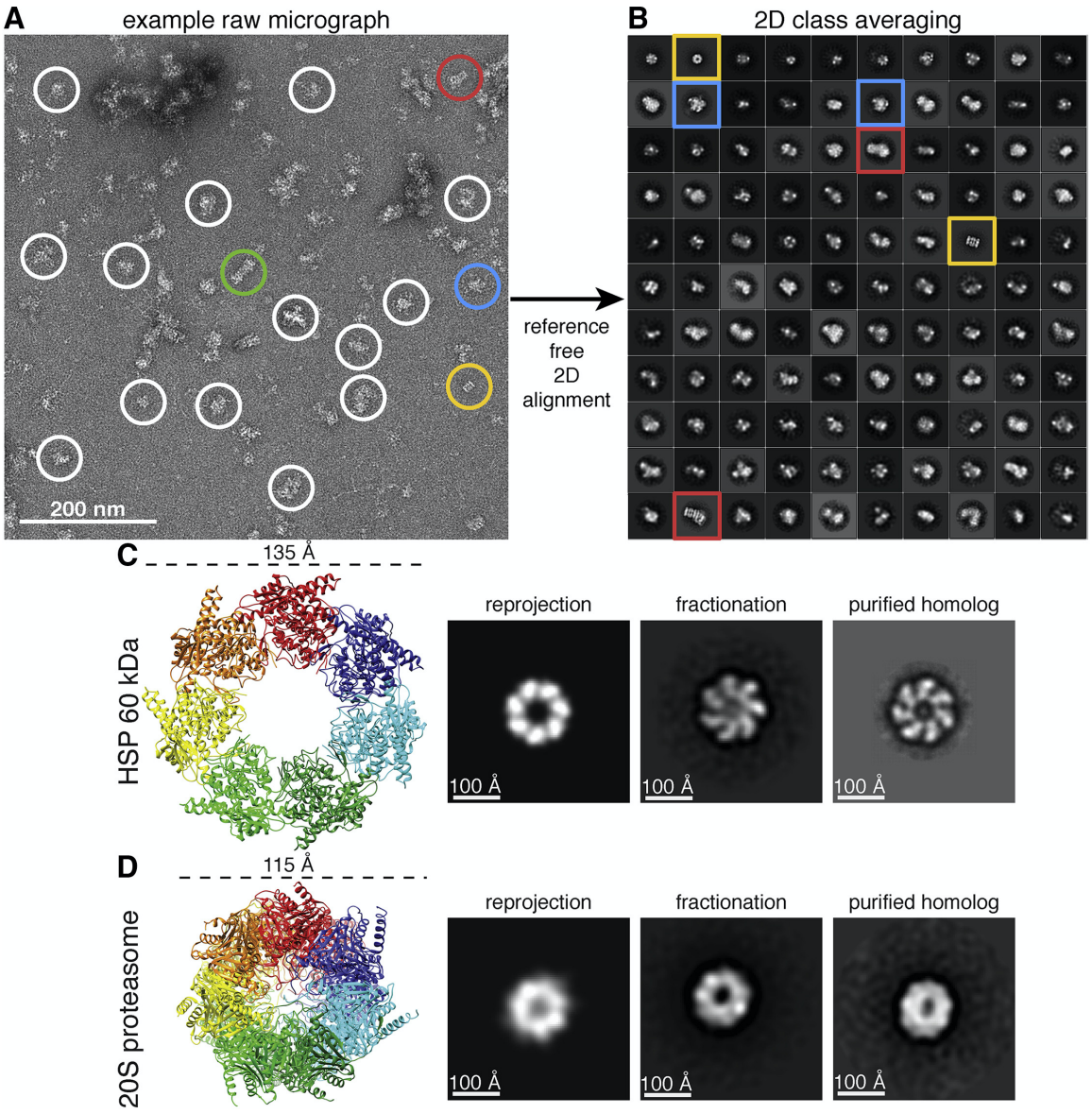

Figure 2.3. Structural characterization of protein complexes from cell extract.

(A) Raw micrograph of negatively stained sample from SEC. Proteasome particles in three different biochemical forms, 20S core, single-capped 26S (20S core with one 19S regulatory particle), and double-capped 26S (20S core with two 19S regulatory particle), are circled in gold, red, and green, respectively. Representative unidentified particles are circled in white. Class averages with well-resolved structural features are circled in blue. (B) Reference-free 2D class averages of 31,731 template-picked particles generated using RELION. The size of each box is 576 × 576 Å. The 2D class averages are sorted in decreasing order based on the number of particles belonging to a class, with 110 out of 300 2D classes shown. (C) Crystal structure of HSP60 (PDB: 4PJ1) identified by MS and

its corresponding reprojection after being low-pass filtered to 30 Å. The 2D class average from our fractionation (fraction 8) matching both the reprojection and a class average of a negatively stained purified homolog (GroEL), adapted from Danziger et al. (2003), suggests the identity of our 2D class average as HSP60. Image box sizes are scaled for consistency. (D) Crystal structure of the 20S proteasome (PDB: 4R30) and its corresponding reprojection after being low-pass filtered to 30 Å. The 2D class average from our fractionation (fraction 4) matching both the reprojection and a class average of a negatively stained, purified *S. cerevisiae* proteasome suggests the identity of our 2D class average as the 20S proteasome. Image box sizes are scaled for consistency. See also Figure S2.

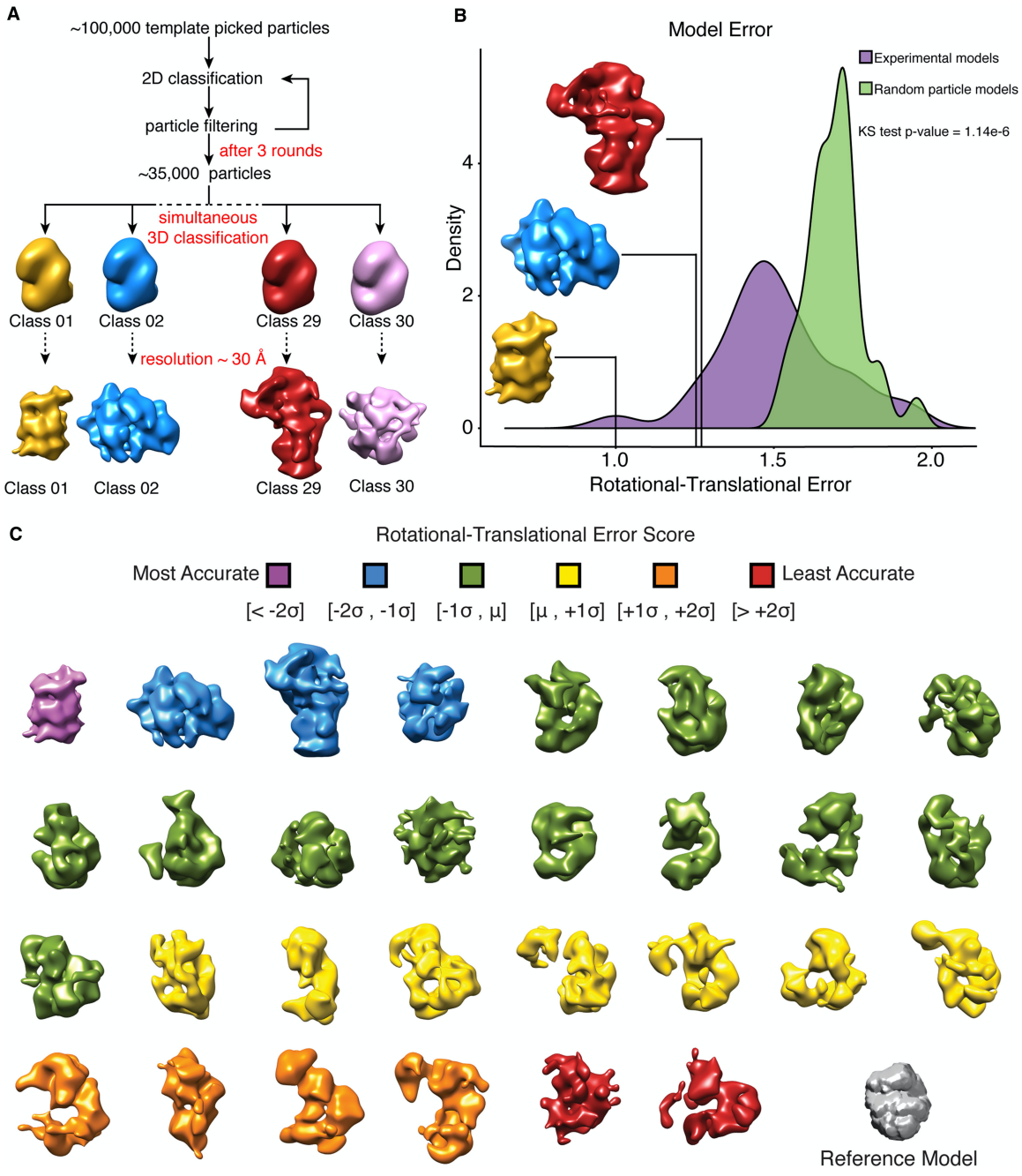

Figure 2.4. Classification of distinct protein complex architectures.

(A) Classification workflow for the simultaneous generation of 30 3D models from the complete dataset of particles using RELION. Models were built using DNA-dependent protein kinase catalytic subunit low-pass filtered to 60 Å as an arbitrary reference model. (B) Top 3 models generated using RELION. Models were scored based on their rotational-translational error (a measure of the internal consistency of the model; see

Methods). The distribution of model error scores was compared to models generated using random particles from our template-picked data. (C) 30 classes generated using RELION from the complete template-picked dataset of particles with the reference model shown in gray. Models are colored by their rotational-translational error and are unrelated to colors in (A) and (B). See also Figure S3.

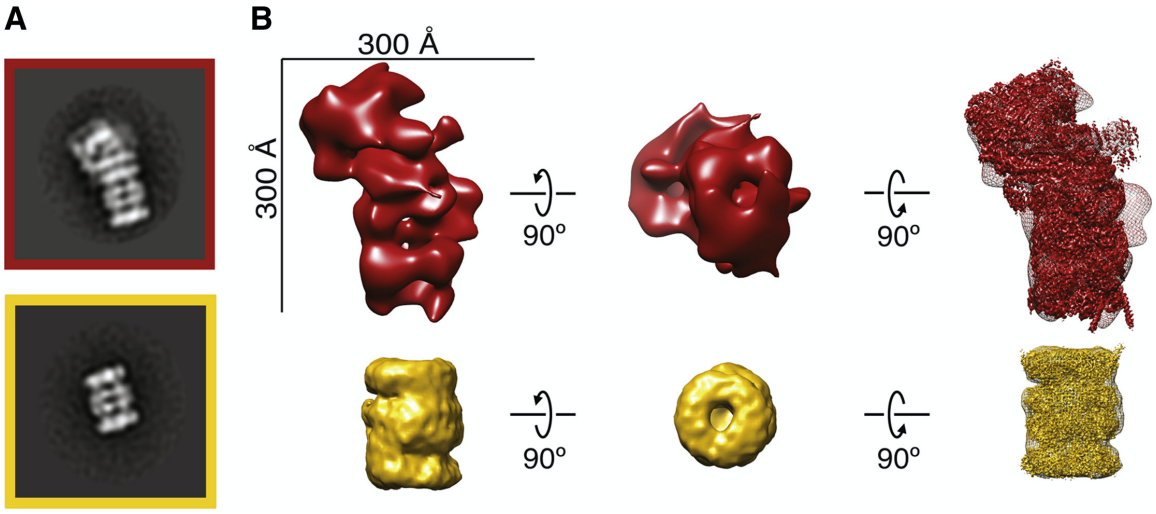

Figure 2.5. *Ab initio* structures from a cellular fraction unambiguously reveal the proteasome.

(A) Reference-free 2D class averages of the proteasome from Figure 3B. (B) Top: structure of single-capped proteasome generated using RELION from manually picked particles. Bottom: *ab initio* structure of the 20S core proteasome generated using cryoSPARC is shown. High-resolution structures EMD-4002 (Schweitzer et al., 2016) and EMD-2981 (da Fonseca and Morris, 2015) are fit into the structures, respectively. See also Figures S4 and S5.

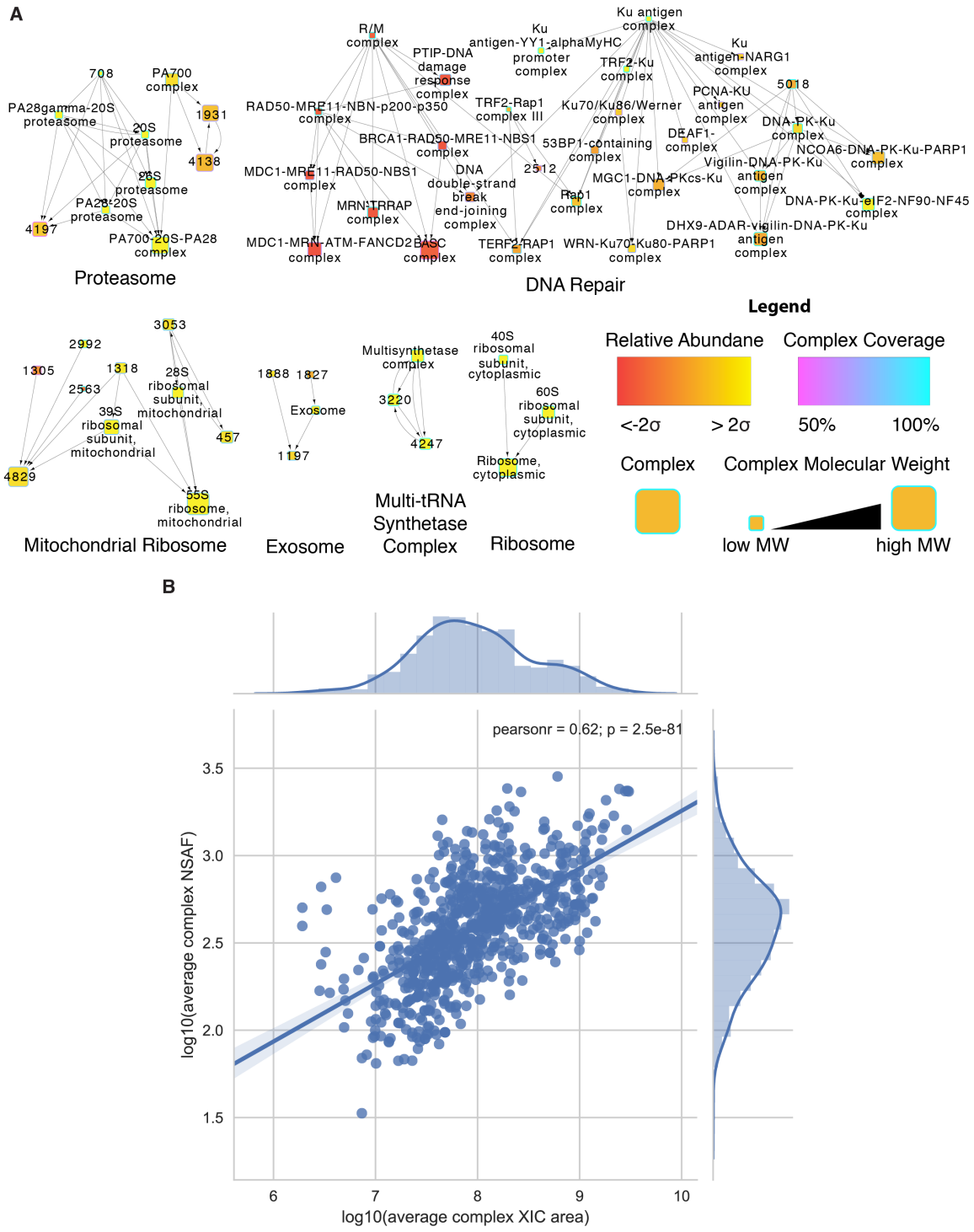

Figure 2.S1. Hierarchical network of related protein complexes. Related to Figure 2 and Table S1.

(A) Subset of the hierarchical network showing related complexes identified by MS in our sample. Each node represents a protein complex and is identified by name or by cluster number from NSAF quantified data (Table S1). The size of each node depicts the molecular weight of the complete complex. Node fill color gradient represents the relative abundance of the complex determined by label-free quantification (see Methods). Node border color gradient represents the percent of subunits in a complex identified by MS. Arrows between nodes indicate at least 90% similarity in subunit composition between source and target node. (B) Comparison of protein complex relative abundance as calculated using two different label-free quantification strategies.

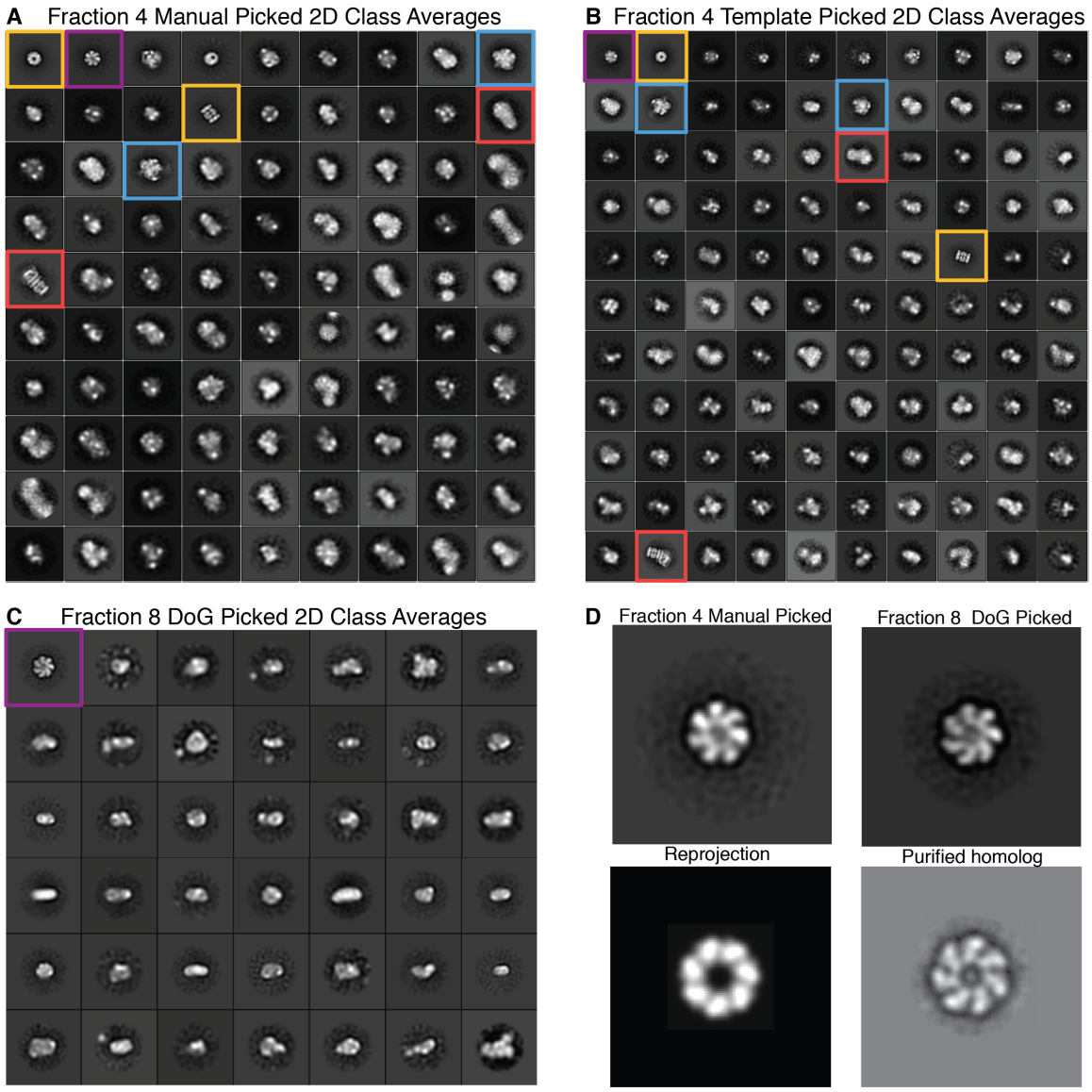

Figure 2.S2. Classification of particles using RELION. Related to Figure 3.

(A) Reference-free 2D class averages of 31,731 template picked particles generated using RELION. The size of each box is 576 Å x 576 Å. The 2D class averages are sorted by the number of particles belonging to each class. Highlighted boxes show examples of similar 2D classes from both particle selection methods of fraction 4 data. (B) Reference-free 2D class averages of 35,381 manual picked particles generated using RELION. The size of each box is 518.4 Å x 518.4 Å. The 2D class averages are sorted by the number of particles belonging to each class. (C) Reference-free 2D class averages of 28,553

Difference of Gaussian picked particles generated using RELION. The size of each box is 518.4 Å x 518.4 Å. (D) Reference-free 2D class averages of HSP60 identified in both fraction 4 and fraction 8. Reprojection of the HSP60 X-ray crystal structure (PDB 4PJ1) low-pass filtered to 30 Å and a 2D class average of a negatively stained purified protein homolog adapted from (Danziger et al., 2003) shown as comparison. Image box sizes are scaled for consistency.

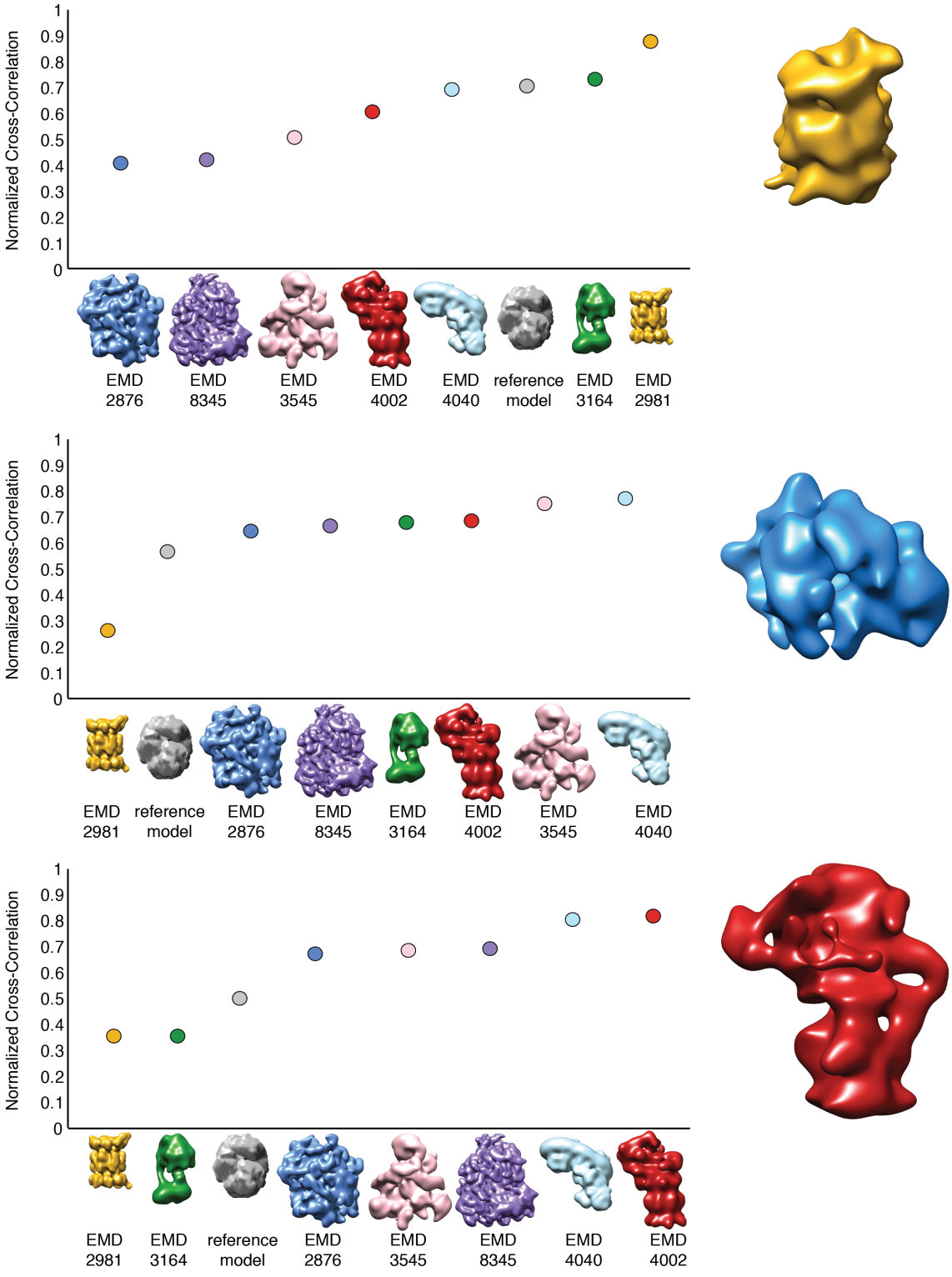
Normalized Cross-Correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
EMD 2876
EMD 8345
EMD 3545
EMD 4002
EMD 4040
reference model
EMD 3164
EMD 2981
Normalized Cross-Correlation
EMD 2981
reference model
EMD 2876
EMD 8345
EMD 3164
EMD 4002
EMD 3545
EMD 4040
Normalized Cross-Correlation
EMD 2981
EMD 3164
reference model
EMD 2876
EMD 3545
EMD 8345
EMD 4040
EMD 4002

Figure 2.S3. Cross-correlation comparison of top 3 RELION models to complexes identified by MS. Related to Figure 4.

Normalized pairwise cross-correlation scores for our top 3 RELION reconstructions to each of the following previously solved cryo-EM structures: EMD-2876 – mitochondrial ribosome, EMD-2981 – 20S proteasome core, EMD-3164 – bovine mitochondrial ATP synthase, EMD-3545 – c* spliceosome, EMD-4002 – 26S proteasome, EMD-4040 – respiratory complex I, EMD-8345 – 80S ribosome.

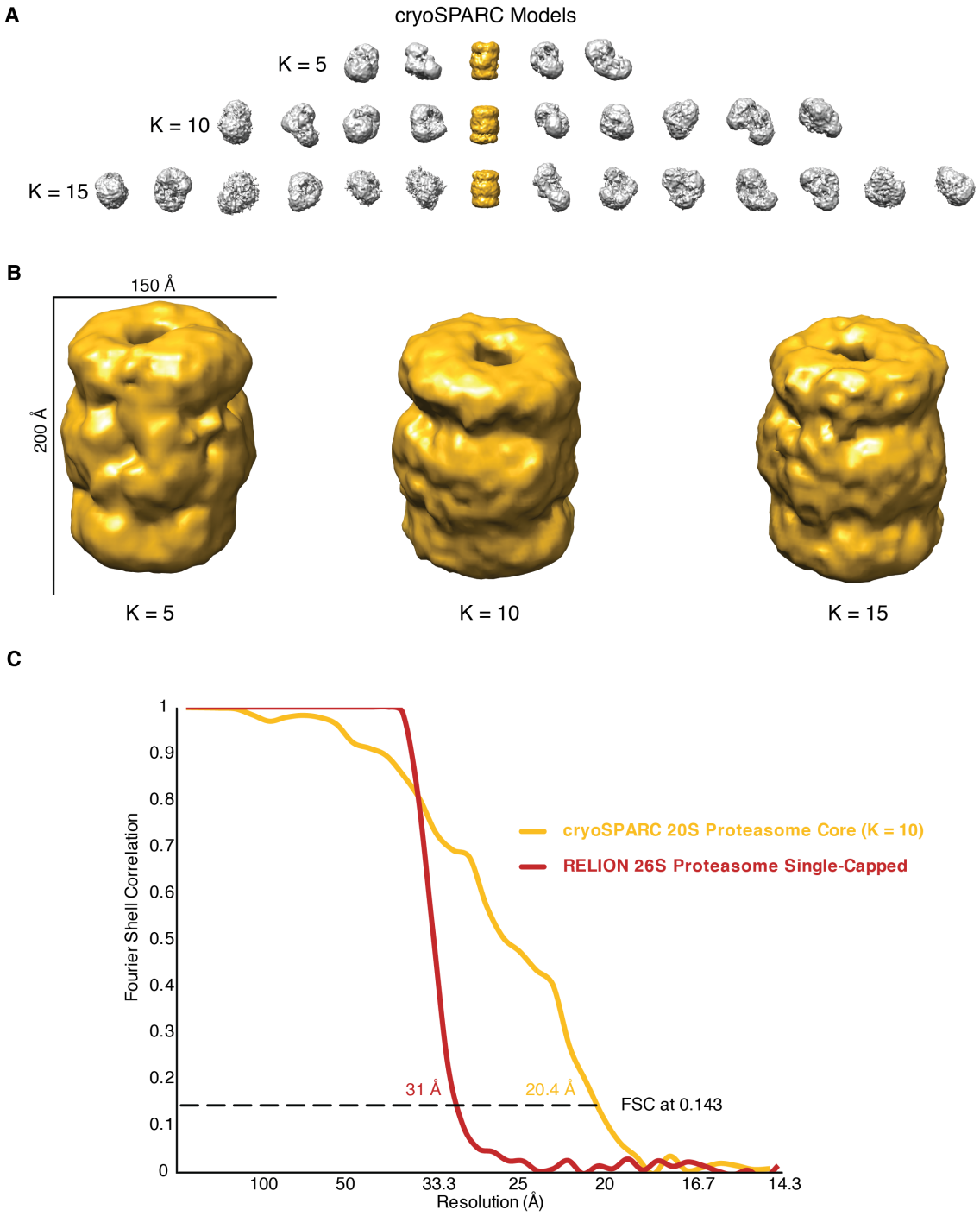

Figure 2.S4. 3D models using cryoSPARC with k = 5,10,15 and related Fourier shell correlations curves. Related to Figure 5.

(A) Reconstructed 3D models from 35,381 manually picked particles when sorted into 5, 10 and 15 *ab initio* classes by cryoSPARC. The 20S proteasome core is highlighted in gold. (B) Comparison of 20S proteasome core models from 5, 10 and 15 classes. (C) FSC curves for the single-capped 26S proteasome (red) and 20S core proteasome (gold) shown in Figure 5B. Nominal resolutions were estimated to be 31 Å and 20.4 Å using the 0.143 gold-standard FSC criterion for the single-capped 26S and 20S core proteasome, respectively.

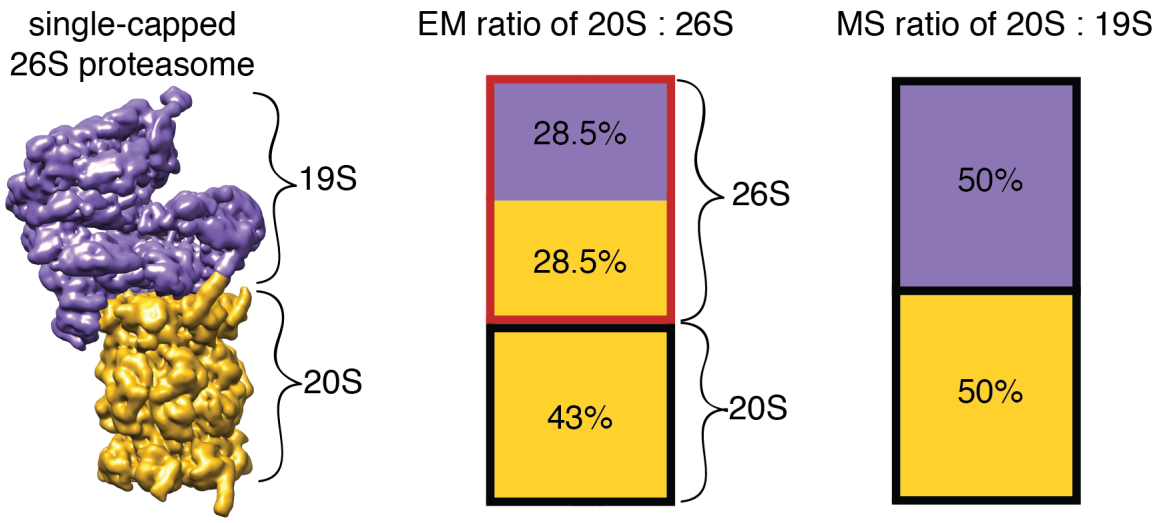

Figure 2.S5. Comparative quantification of the proteasome by MS and EM. Related to Figure 5 and Table S1.

Quantification of proteasome particles by single particle counting of EM data and extracted ion chromatogram areas.

Chapter 3: Electron microscopy snapshots of single particles from single cells

Large biological data sets (-omics data) have become pervasive in most fields of biology, even extending to data collected from single cells. In the field of structural biology, -omics data from single cells is typically acquired by tomography of whole cells or cell sections. While tomographic data uniquely provides spatial context, current limitations often restrict data to the largest protein assemblies and ultrastructure of the cell. The work presented in this chapter demonstrates how simple microfluidic devices can be combined with single particle electron microscopy to investigate protein structures from single cells. This chapter is published as Yi, X.*, Verbeke, E.J.*, Chang, Y.*, Dickinson, D.J., and Taylor, D.W. (2019). Electron microscopy snapshots of single particles from single cells. J. Biol. Chem. 294, 1602–1608. Xiunan Yi and Yiran Chang built the microfluidic devices and prepared the *C. elegans* embryos for electron microscopy. I led the bioinformatics and computational analysis with contributions from the other two co-first authors.

3.1. ABSTRACT

Cryo-electron microscopy (cryo-EM) has become an indispensable tool for structural studies of biological macromolecules. Two additional predominant methods are available for studying the architectures of multiprotein complexes: 1) single-particle analysis of purified samples and 2) tomography of whole cells or cell sections. The former can produce high-resolution structures but is limited to highly purified samples, whereas the latter can capture proteins in their native state but has a low signal-to-noise

ratio and yields lower-resolution structures. Here, we present a simple, adaptable method combining microfluidic single-cell extraction with single-particle analysis by EM to characterize protein complexes from individual *Caenorhabditis elegans* embryos. Using this approach, we uncover 3D structures of ribosomes directly from single embryo extracts. Moreover, we investigated structural dynamics during development by counting the number of ribosomes per polysome in early and late embryos. This approach has significant potential applications for counting protein complexes and studying protein architectures from single cells in developmental, evolutionary, and disease contexts.

3.2. INTRODUCTION

Cell behavior is fundamentally dependent on the activities of macromolecular machines. These machines, comprised of protein (and sometimes RNA) subunits, are responsible for catalytic, structural, and regulatory activities that allow cells to function. Structural biology, by revealing the physical architecture of macromolecules and their assemblies, plays a critical role in efforts to understand how molecular mechanisms contribute to cell behavior *in vivo*.

A crucial feature of most living cells is their ability to adjust their behavior in response to their environment. In a developmental context, cells respond to chemical and mechanical cues from neighboring cells and tissues to coordinate their behavior with their neighbors and to assemble functional tissues. A major goal of developmental biology studies is to understand the molecular mechanisms of these interactions—that is, how dynamic behaviors of macromolecular machines give rise to cell behaviors that support proper organismal development.

Recently, single-cell nucleic acid sequencing approaches have revolutionized developmental studies by allowing gene expression to be interrogated with unprecedented spatiotemporal resolution (Kumar et al., 2017). Such experiments are powerful because they reveal which genes are expressed in which cells at a particular point in development and can thus provide insights into signaling dynamics, mechanisms of cell state changes (e.g. cellular differentiation), and levels of heterogeneity between individual cells. However, sequencing approaches do not shed light on the molecular states or interactions of cellular proteins. A few studies have begun to extend a single-cell approach to biochemical studies of proteins and protein complexes. For example, Huang and Zare (Huang et al., 2007) described a sophisticated microfluidic device for counting protein molecules in single-cell lysates. Allbritton and co-workers (Dickinson et al., 2013; Kovarik and Allbritton, 2011) have developed capillary electrophoreses methods for measuring enzyme activities in whole-cell lysates. Most recently, Dickinson et al. (Dickinson et al., 2017) used microfluidic lysis followed by single-molecule pulldown and TIRF microscopy to measure the abundance of protein complexes in single cells. This single-cell, single-molecule pulldown (sc-SiMPull) approach was sufficiently sensitive to reveal regulated changes in protein– interactions that occurred over ~5 min during development of the *Caenorhabditis elegans* zygote. Thus, single-cell biochemical approaches have the potential to uncover dynamics of macromolecular machines in cell or tissue samples obtained directly from developing embryos.

Although still in their infancy, the initial success of these single-cell biochemical methods raises the question of whether a single-cell approach could be extended to macromolecular structure determination. Such an approach could overcome a classical

limitation of structural biology: its need for highly purified, homogenous proteins (or protein complexes) that represent only a single snapshot from the ensemble of structures that are likely present in cells. Moreover, the ability to determine structures of proteins obtained directly from cells engaged in development would represent a significant step toward the goal of linking the structural dynamics of molecular machines to their cellular and developmental consequences.

One approach to single-cell structural studies is electron tomography. This allows for the study of cell morphologies (Beck and Baumeister, 2016), and in some cases, can be used to reconstruct 3D models directly from native cells (Galaz-Montoya and Ludtke, 2017). However, because of the sensitivity of biological specimens to electron dose, tomographic approaches routinely lead to low- or intermediate-resolution structures of complexes in single cells using subtomogram averaging techniques. The advent of phase plates for EM has revolutionized the information contact extractable from tomograms, but high-resolution structures of less-abundant complexes remain elusive.

Alternatively, single-particle cryo-electron microscopy (cryo-EM) is now capable of routinely achieving high-resolution structures of highly purified samples because of advances in hardware (Kühlbrandt, 2014) and software (Punjani et al., 2017; Scheres, 2012). We and others have recently extended single-particle EM techniques to study heterogeneous mixtures from biochemically fractionated cell lysate (Kastritis et al., 2017; Verbeke et al., 2018). Although these shotgun-EM approaches are able to sort through the heterogeneity of macromolecules, they still rely on a large quantity of cells and mass spectrometry (MS) to characterize the contents of the sample. Furthermore, direct

investigation of proteins at the single-cell level has remained a challenging problem for proteomic studies. This poses a unique challenge to structural studies of single cells.

Some efforts toward applying single-particle EM methods to single cells have been made (Arnold et al., 2017; Kemmerling et al., 2013) that importantly demonstrated the feasibility of extracting material from single cells for EM analysis. However, this earlier work required a complicated apparatus and has not yet yielded any 3D structures of macromolecular complexes. Here, we propose an alternative approach for combining single-cell lysis with EM to investigate macromolecular structures. Our method is technically simpler than previous approaches but is able to directly visualize the contents of a single cell. After computationally classifying the particles from cell lysate, we uncover the 3D structures of 40S and 60S ribosomes from disperse particles and the structure of an 80S ribosome from polysomes. Because we chose to apply our approach to a developmental model system (*C. elegans* zygotes and embryos), we are able to obtain structural information from embryos at specific developmental stages. In one application, we find that the number of ribosomes per polysome remains consistent between early- and late-stage embryos. These results demonstrate the potential of EM for structural characterization of unpurified macromolecular machines obtained from samples as small as a single cell.

3.3. RESULTS

3.3.1. Extracting macromolecules from single embryos

Our primary goal in this study was to determine whether imaging of single-cell lysates with EM could yield sufficient, high-quality particles for 3D structure determination. To obtain intact, native particles from single cells, *C. elegans* zygotes (i.e. 1-cell embryos) were trapped and lysed using microfluidic chambers (Figure 1; see Methods). We then transferred the cell lysates (a volume of ~50 nL) from the microfluidic channels to EM grids using a glass needle (Video S1; see Methods). Because of the small volume, which was insufficient to coat an entire grid, we used reference grids containing alphanumeric markers to locate the placement of our samples under both the dissecting scope and the electron microscope. Each reference grid was then conventionally stained using 2% (w/v) uranyl acetate. We chose to use negative stain EM for its high signal-to-noise ratio to more accurately assess our ability to identify single particles from individual cell lysates. Each grid was then examined by transmission EM to identify grid squares that contained cellular protein particles embedded in stain.

To demonstrate our ability to capture small volumes of samples on EM grids, we first transferred samples of a purified protein kinase (Zhan et al., 2015) from our microfluidic device to an EM grid for visualization, resulting in successful detection of the kinase (Figure S1). We then performed our transfer technique on lysates from seven independent single embryos sampled at different developmental stages, three from zygotes and four from later-stage, multicell embryos. Micrographs of single-cell extract across different embryos show a reproducible mixture of heterogeneous particles that span an order of magnitude in size (Figure S1). These data allowed us to investigate the

dynamics of protein complexes at the single-cell/embryo level between different developmental stages of *C. elegans* zygotes.

3.3.2. EM of extract from a single *C. elegans* embryo

Raw micrographs collected at the locations where embryo lysate had been applied to the EM grid showed distinct, monodisperse particles with varying sizes and distinct shapes. The results unambiguously show that we were able to retrieve cellular contents from our microfluidic lysis chips for subsequent imaging by EM, although we cannot exclude the possibility that some particles may fail to adhere to the grid and be lost during sample preparation (Figure 2A). We collected ~1,400 micrographs between the seven samples. Although small particles were abundant in our micrographs (Figure S1), we first chose to analyze large particles (~150–300 Å in diameter), which were easily recognizable and appeared relatively homogeneous. After manually selecting ~10,000 large particles from a subset of micrographs, we generated reference-free 2D class averages that were subsequently used as templates for automated picking of particles from all micrographs. Using this template picking scheme, ~80,000 large particles were selected from ~1,400 micrographs and used for reference-free 2D alignment and classification. 2D class averages with distinct structural features were generated from ~50,000 particles after removing junk particles (e.g. detergent micelles, irregular small particles, two nearby particles, or particles in aggregates) from the data (Figure 2B, top panel).

To obtain insight into the possible identities of these particles, we used publicly available RNA-seq data from *C. elegans* 1-cell embryos (Gerstein et al., 2010; Hillier et

al., 2009) to inform us about which proteins are likely to be highly expressed. 34 ribosomal protein transcripts and 4 proteasomal protein transcripts were among the 200 most abundant transcripts, suggesting these protein complexes were likely to appear in our micrographs (Figure S2) (Gerstein et al., 2010; Hillier et al., 2009). None of the other 200 most highly expressed proteins are known subunits of large (megadalton) macromolecular complexes, suggesting that ribosomes and proteasomes should be the most abundant large particles in our data set. We therefore performed pairwise cross-correlations of our 2D class averages with 2D class averages of purified 40S ribosome, 60S ribosome (Malyutin et al., 2017), and 26S proteasome (Sone et al., 2004) from *Saccharomyces cerevisiae* to look for structural similarities. The alignment revealed several classes with similar features between our single-cell lysate and the known, purified structures, suggesting the identity of several projections in our sample were in fact the 40S ribosome, 60S ribosome, and 26S proteasome (Figure 2B, bottom panel). This initial 2D classification proved it is possible to obtain structural information from intact protein complexes extracted from lysates of single cells.

3.3.3. Capturing ribosome dynamics in polysomes

Intriguingly, our raw micrographs revealed densely packed clusters of ribosome-like particles (Figure 3A). These ribosome-like particles appeared in organized arrays with a similar appearance to polysomes from *Escherichia coli* (Brandt et al., 2009) and wheat germ (Afonina et al., 2013). Polysomes consist of a pool of actively translating ribosomes on an mRNA transcript. This suggested that our single-cell EM method is capable of capturing protein–mRNA interactions in the cell.

In *C. elegans*, zygotic transcription begins at the four-cell stage (Edgar et al., 1994; Seydoux and Fire, 1994); prior to this, development is driven by maternal RNA and proteins. We were curious whether polysome architecture would change as a consequence of new zygotic transcription. Because we observed polysomes in both early- and late-stage embryos (before and after the onset of zygotic transcription), we addressed this question by counting the number of ribosomes per polysome in our samples from each developmental stage. Each micrograph was manually annotated to determine the number of ribosomes per polysome cluster for our early-stage and late-stage embryos, respectively. Using this approach, we determined that there are, on average, eight ribosomes per polysome for early-stage embryos and seven ribosomes per polysome for late-stage embryos (Figure 3B). These numbers are consistent with previous studies in which the number of ribosome per mRNA is estimated by isolating polysomes using velocity sedimentation in sucrose gradients (Slavov et al., 2015), but our results add an additional dimension by observing polysomes at defined developmental stages that either have or lack zygotic transcription. Although our data suggested no significant change in the number of ribosomes per polysome between early and late stage embryos, this analysis provides evidence that single-particle counting from single cells could potentially be applied for investigating the dynamics of macromolecules in different cell states.

3.3.4. 3D classification of ribosome particles from single embryo data

We then performed 3D classification of our large particles to determine whether any distinct structures could be obtained from lysates of single cells. Specifically, we were looking for structures of ribosomes because they appeared as clear and abundant 2D

class averages in our data. We first combined two data sets from early-stage embryo samples for 3D classification using RELION (Figure S3) (Scheres, 2012). After removal of junk particles, ~14,000 particles were used for classification. Initially, we used an unbiased approach for 3D classification by using an initial model of a featureless 3D shape with uniform electron density. Using a model reconstructed from this initial classification, which resembled a previously determined 60S ribosome structure (Shen et al., 2015) (EMDB-2811) as a reference, we then performed another round of 3D classification (see Methods). The models from each classification were then compared by docking a high-resolution *S. cerevisiae* 60S ribosome structure (EMDB-2811) into our maps to determine which class, if any, was most similar to the known structure. Our top scoring 60S ribosome reconstruction, containing ~3,400 particles, displayed striking similarity to the *S. cerevisiae* 60S ribosome (Figure 4, top row) with a cross-correlation score of 0.8143 and a nominal resolution of 34 Å calculated using the 0.5 Fourier shell correlation criterion (Figure S4; see Methods).

Performing 3D classification on particles from all data sets combined, a final set of ~17,000 particles after stringent removal of junk particles resulted in an additional class that resembled the *S. cerevisiae* 40S ribosome (Figure 4, middle row; see Methods). Our 40S ribosome reconstruction, containing ~1,450 particles, had a cross-correlation score of 0.8352 with the *S. cerevisiae* model (Scaiola et al., 2018) (EMDB-4214) and a nominal resolution of 48 Å (Figure S4). These results suggest that 3D structures of multiple protein complexes can be obtained from lysates of single embryos using single-particle EM analysis. To explore our ribosome reconstructions, we built a hybrid 80S ribosome model by aligning our 40S and 60S ribosome reconstructions to their respective domains in the 80S ribosome from a previously determined 80S ribosome structure

(Figure 4, bottom row) (Cianfrocco and Leschziner, 2015) (EMDB-2858). As expected, this hybrid model is consistent with the high-resolution 80S ribosome structure. Because of the limited number of proteasome particles in our sample, we did not attempt to obtain a 3D structure of the proteasome.

With clear structures resembling a 40S and 60S ribosome, we next attempted to determine the molecular architecture of the 80S ribosome directly from polysome clusters. Our data contained ~9,000 particles within polysomes that were manually picked for single-particle analysis. We then performed 2D and 3D classification of the selected particles (see Methods). The 2D class averages were ~250 Å in diameter, which is consistent with the size of an *S. cerevisiae* 80S ribosome. Our 3D model, containing ~2,000 particles, had a cross-correlation score of 0.7572 when compared with an *S. cerevisiae* 80S ribosome (Cianfrocco and Leschziner, 2015) (EMDB-2858) and a resolution of 45 Å (Figure S4). Although our 80S ribosome model lacked some areas of density present in the high-resolution structure, the overall size could accommodate both the 40S and 60S ribosome. Collectively, our data show that we are able to distinguish structures of the 40S, 60S, and 80S ribosomes directly from particles isolated from single cells.

3.4. DISCUSSION

A major goal of basic biological research is to connect structural dynamics of macromolecules to their effects on cell behavior. Here, we present an approach for structural characterization of protein complexes isolated from single cells engaged in development. We demonstrate that a single cell contains a sufficient number of protein

particles to enable structural characterization by EM. We think that this approach has significant potential to reveal structural changes in protein complexes across developmental and disease contexts. Our method is also promising for future single-cell cryo-EM, because the lysate transferring procedure can remain the same during cryo-EM sample preparation. The increased resolution from cryo-EM will help in the identification of other macromolecular complexes from lysate. However, we expect several obstacles moving our approach to cryo-EM including freezing small lysate volumes and capturing low-concentration proteins on grids.

Moving forward, a significant challenge will be to extend this approach beyond ribosomes and proteasomes to other macromolecular complexes. We focused here on ribosomes because they are large, highly abundant, and relatively easy to recognize. For complexes that are less abundant and/or less distinctive in shape, we will need to develop methods to identify a complex of interest in a heterogeneous mixture. Correlative light and EM holds promise in this regard (Schorb and Briggs, 2014). We also plan to explore whether particles isolated and characterized via our earlier sc-SiMPull approach (Dickinson et al., 2017) can be eluted and transferred to EM grids for structural analysis. An added advantage of this strategy would be the ability to use multicolor TIRF to characterize the composition of complexes whose structures could then be determined. Taken together, we are optimistic that these strategies will allow us to gain structural information about protein complexes beyond ribosomes and proteasomes using single-cell lysates.

A related question is whether there are enough particles in a single cell to allow high-resolution structure determination. This will of course depend on the protein or

protein complex being studied; it may be more difficult to determine high-resolution structures of low-abundance complexes. However, we note that the time required to prepare and collect data from a single-cell sample is short enough that analyzing 10–20 such samples is realistic. Particles from multiple samples could be pooled to increase resolution, without sacrificing information about which cell each particle in the data set came from. This might represent an ideal compromise between the need for increased numbers of particles for structure determination and the desire for single-cell resolution for detailed developmental studies.

3.5. METHODS

3.5.1. Microfluidic device fabrication

Microfluidic devices were fabricated using a standard soft lithography procedure. A photomask corresponding to the desired channel shape was designed using CAD software and produced by Cad-Art Services (Bandon, OR). An ~30-μm-thick layer of SU8–2025 photoresist was deposited on a plasma-treated silicon wafer by spin coating for 10 s at 400 rpm followed by 30 s at 2800 rpm and 30 s of deceleration. After soft baking at 65 °C for 3 min and 95 °C for 10 min, the films were exposed to 1000 mJ UV light through the photomask. Following a post-exposure bake of 5 min each at 95 °C and 120 °C, the molds were developed in SU8 developer (propylene glycol monomethyl ether acetate, PGMEA) and rinsed with isopropanol. The molds were hard baked at 95 °C for 30 min and then at 120 °C overnight.

PDMS (Sylgard 184 silicone elastomer kit, Dow Corning, Midland, MI) was mixed using a 10:1 ratio of base to curing agent and deposited onto the molds by spin coating at 400 rpm for 30s. The PDMS was cured for 20 min at 95 °C, then peeled off from the molds, and inlet and outlet holes were punched with a 2 mm biopsy punch. Each PDMS device contained 8 channels, and each channel was used for one single-embryo experiment.

24 × 60 mm glass coverslips were cleaned with ethanol and dried under nitrogen flow. Each cleaned coverslip was bonded to a PDMS device by 2 min of treatment with air plasma, then baked at 120 °C for 30 min to form a permanent bond.

The PDMS device was first activated by flowing 1 m KOH through the channels for 20 min, washed three times with water, and then dried. After activation, 2-[methoxy(polyethylenxy)9–12Propyl]-trimethoxysilane was applied to the channels for 30 min to prevent nonspecific protein binding. The channels were then washed three times with water and dried. The dry devices were cured overnight at room temperature and stored with the open holes facing downward, in a closed box, until use.

3.5.2. Sample preparation from staged embryos

WT *C. elegans* embryos (strain N2) were dissected from gravid adults in egg buffer (5 mM HEPES, pH 7.4, 118 mM NaCl, 40 mM KCl, 3.4 mM MgCl2, 3.4 mM CaCl2). Developmental stage was determined by visual inspection of morphology (cell shape and nuclear position) on a dissecting microscope. The embryo with desired stage was transferred to a 3-μL drop of lysis buffer (10 mM Tris, pH 8, 50 mM NaCl, 0.1%

Triton X-100, 10% glycerol) and placed in the inlet well of a prepared microfluidic device using a mouth pipette. A clean 26-gauge needle was used to push the embryo into the microfluidic channel.

Once the embryo was trapped in the center of the chamber, the channel output was sealed with crystallography-grade clear tape (Crystal Clear, Hampton Research, Aliso Viejo, CA) to stop flow. The device was temporarily fixed under the dissecting microscope with the tape. The embryo was then immediately crushed while watching in the stereoscope, by pushing down on the surface of the PDMS with the melted tip of a glass Pasteur pipette. A clean glass needle connected to a 10-mL syringe through a short flexible tubing was used to puncture the top layer of the PDMS channel once the embryo lysed. The lysate (an approximate volume of 50 nL) was sucked into the needle and transferred onto a marked area of a glow discharged reference grid covered with carbon. Two to three different lysates were transferred onto different squares of the same grid, with no overlap. After the last embryo lysate was transferred, the grid was immediately negatively stained with five consecutive droplets of 2% (w/v) uranyl acetate solution, blotted to remove residual stain, and air-dried in a fume hood. Purified JNK2 (a gift from K. Dalby and N. Sun) was used in control experiments (Shaw et al., 2008).

3.5.3. EM and data collection

Data were acquired using a JEOL 2010F transmission electron microscope operated at 200 keV with a nominal magnification of ×60,000 (3.6 Å at the specimen level). Each image was acquired using a 1-s exposure time with a total dose of ~30–35 e^{-} $Å^{-2}$ and a defocus between −1 and −2 μm. A total of 1,402 micrographs from seven

samples (three early embryos and four late embryos) were manually recorded on a Gatan OneView camera.

Seven independent particle stacks were generated from the micrographs of each sample: 341 micrographs of an early-staged embryo sample (E1), 350 micrographs of an early-staged embryo sample (E2), 250 micrographs of an early-staged embryo (E3), 100 micrographs of a late-staged embryo (L1), 111 micrographs of a late-staged embryo (L2), 147 micrographs of a late-staged embryo (L3), and 103 micrographs of a late-staged embryo (L4). FindEM (Roseman, 2004) was used for template-based particle picking with a template selected from reference-free 2D class averages generated from ~10,000 large particles which were manually picked from the E1 data set. In total, ~81,600 particles were selected from template picking of all data sets. All image pre-processing was done in Appion (Lander et al., 2009). After removing junk particles, 17,070 particles remained for further processing. Particle box size was set to 576 × 576 Å. Reference-free 2D class averages were generated with 100 classes using RELION (Scheres, 2012). The 2D class averages of large particles in the embryo lysate were compared with those of purified 40S ribosomes, 60S ribosomes, and 26S proteasomes from *S. cerevisiae* (a gift from A. Johnson, S. Musalgaonkar, A. Matouschek, and C. Davis) using EMAN. The micrographs of the yeast ribosomes and proteasomes were taken using the TEM procedures above.

For our 40S ribosome reconstruction, 3D classification was performed using RELION to create 10 classes. We used the structure of a purified DNA-dependent protein kinase catalytic subunit as an arbitrary initial model after being low-pass filtered to 60 Å.

The top scoring model when compared with the *S. cerevisiae* 40S ribosome structure (EMDB 4214) contained 1,466 particles.

For our 60S ribosome reconstruction, a similar strategy was followed. Two independent particle stacks from E1 and E2 were used. The contrast transfer function of each micrograph was estimated using CTFFIND4 (Rohou and Grigorieff, 2015). Approximately 37,200 particles were selected by template picking. After removing junk particles, 13,916 particles were left. Particle box size was set to 432 × 432 Å. Reference-free 2D class averages were generated with 100 classes. 3D classification was performed to create eight classes. The structure of a featureless 3D shape with uniform electron density was chosen as an initial model after low-pass filtering to 60 Å. A subsequent round of 3D classification was performed on the same data using a reconstructed 3D class that was most similar to the 60S ribosome as the new initial model. From this classification, the best of three classes was determined by comparison to a *S. cerevisiae* 60S ribosome structure (EMDB 2811) and contained 3,431 particles.

For our 80S ribosome reconstruction, an initial stack of ~9,000 particles in polysome-like structures were manually selected from all data sets combined. After removing junk particles, 5,638 particles remained for subsequent 2D and 3D classification. Particle box size was set to 576 × 576 Å. Reference-free 2D class averages were generated with 200 classes. 3D classification was performed to create two classes. The top scoring model when compared with a *S. cerevisiae* 80S ribosome structure (EMDB 2858) contained 1,971 particles.

We additionally performed an initial characterization of small particles found in our micrographs. Using a template-free difference of Gaussian particle picker (Voss et al., 2009), ~165,00 particles were selected from data sets E1 and E2. Particle box size was set to 216 × 216 Å. After removing junk particles, 126,095 particles were classified using reference-free 2D classification to generate 150 classes.

3.6. FIGURES

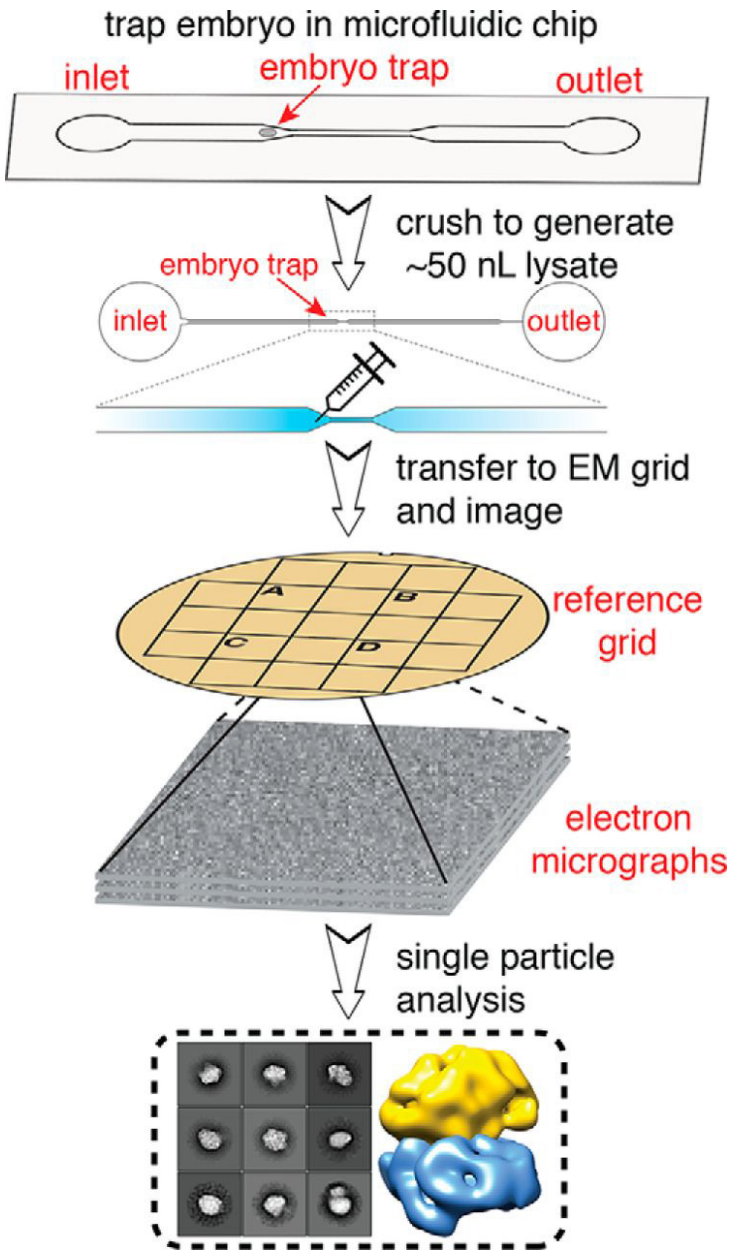

Figure 3.1. Schematic of single-cell structural biology approach.

Single *C. elegans* embryos are trapped in a microfluidic device. After the embryo is crushed, the lysate is extracted using a fine needle and applied to a specific area of an EM grid using a stereoscope. The same area is then visualized using EM, and single-particle analysis is applied for structure determination.

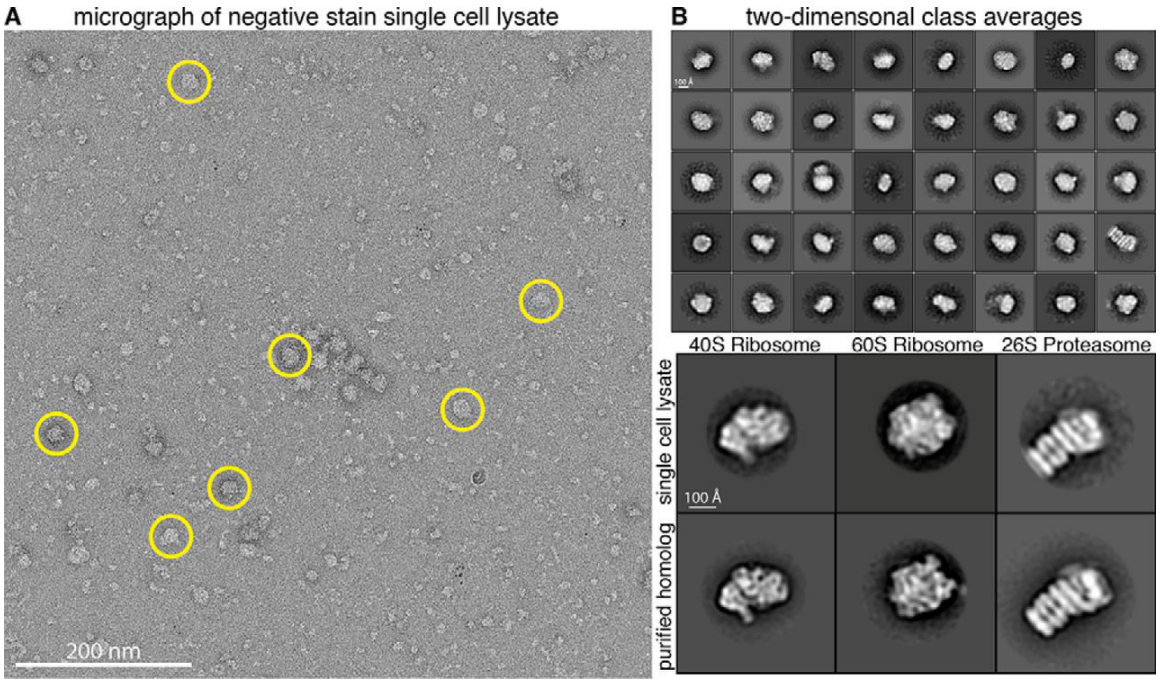

Figure 3.2. Single-particle analysis of extracts from single cells.

(A) Representative raw electron micrograph of negatively stained single-cell lysates. Micrographs show monodisperse particles of varying size. Circled particles are representative of the larger particles (~150–300 Å in diameter) used for subsequent 2D and 3D classification. (B) Top panel, reference-free 2D alignment and classification of a subset of the ~50,000 particles picked from single-cell extract. Classes are sorted in order of decreasing abundance. Box size is 576 × 576 Å. Bottom panel, alignment of 2D class averages from single-cell extract to purified homologs.

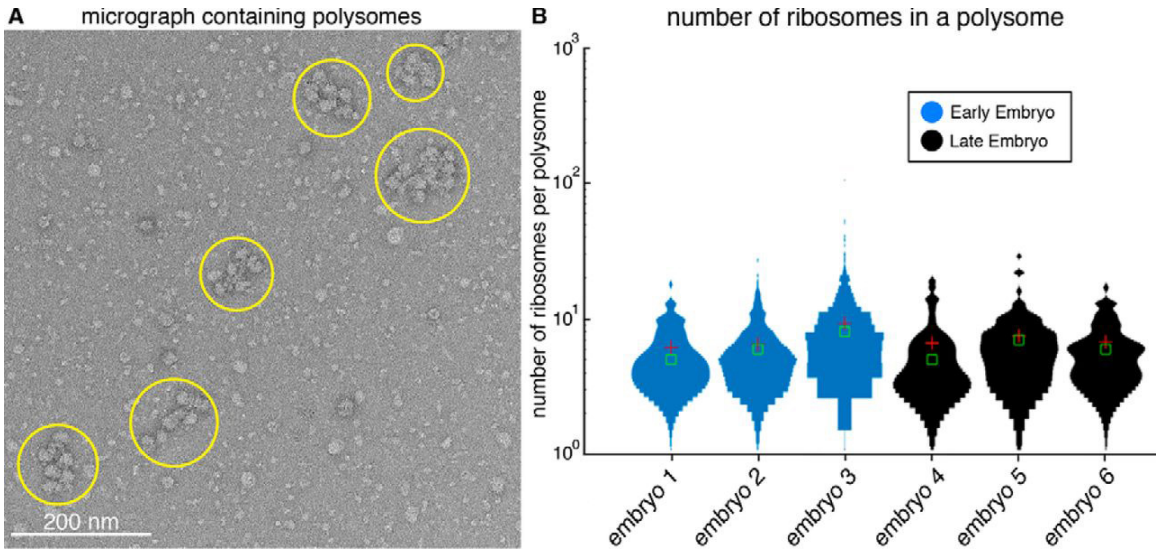

Figure 3.3. Counting ribosomes in polysomes from early- and late-stage *C. elegans* embryos.

(A) Representative raw electron micrograph of negatively stained single-cell lysate showing several distinct polysome clusters of varying size (yellow circles). (B) Distribution of the number of ribosomes in a polysome across three early- and three late-stage embryos. The average numbers of ribosomes for early- and late-stage embryos are eight and seven, respectively. The red cross-hair is the mean value, and the green box is the median (n = 81, 513, 319, 31, 71, and 52 for embryos 1–6, respectively).

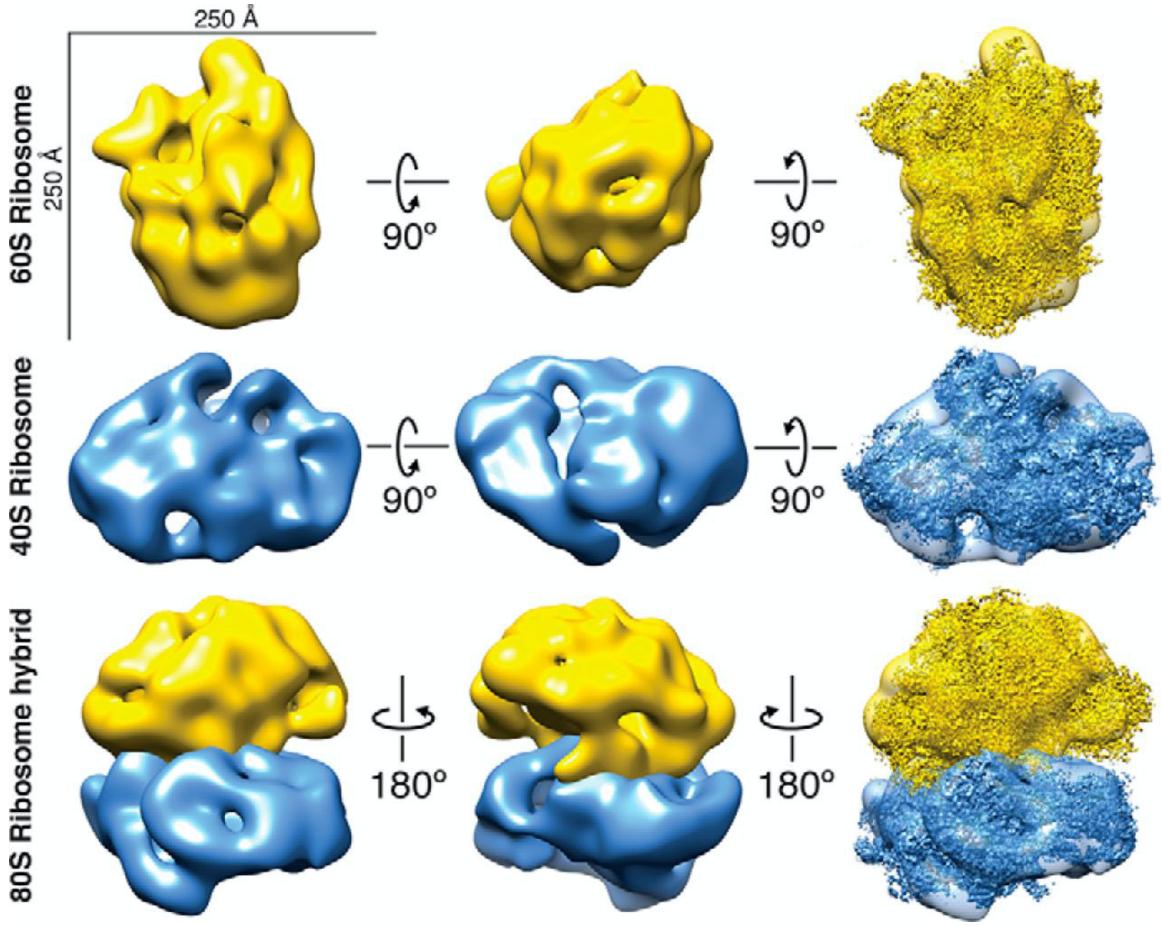

Figure 3.4. 40S and 60S ribosome reconstructions from particles from single cells.

Top row, 60S ribosome reconstruction. High-resolution structure EMDB-2811 (Shen et al., 2015) docked into our 60S map with a cross-correlation score of 0.8142. Middle row, 40S ribosome reconstruction. High-resolution structure EMDB-4214 (Scaiola et al., 2018) docked in to our 40S map with a cross-correlation score of 0.8352. Bottom row, 80S ribosome hybrid model built using our 40S and 60S ribosome aligned to a high-resolution structure of the 80S ribosome EMDB-2858 (Cianfrocco and Leschziner, 2015).

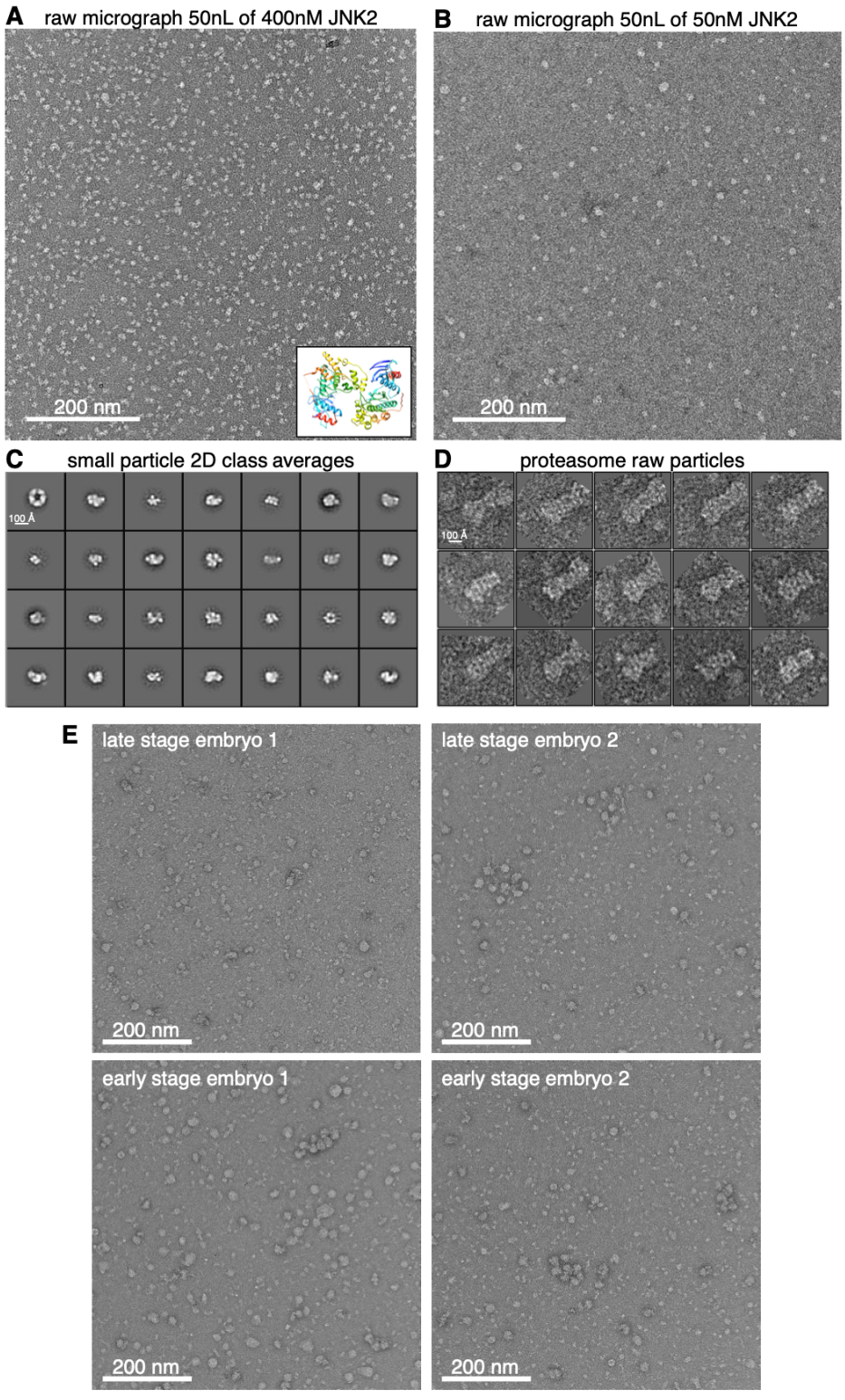
A raw micrograph 50nL of 400nM JNK2
200 nm
B raw micrograph 50nL of 50nM JNK2
200 nm
C small particle 2D class averages
100 Å
D proteasome raw particles
100 Å
E
late stage embryo 1
200 nm
late stage embryo 2
200 nm
early stage embryo 1
200 nm
early stage embryo 2
200 nm

Figure 3.S1. Lysate transfer control experiments and small particle classes.

(A) Micrograph of 50nL of 400nM JNK2 (a kinase protein with known structure). To demonstrate that we can visualize particles using our method, we first flowed 400nM JNK2 solution through the PDMS channel and then ~50nL of the solution was transferred to the reference grid using a glass needle. Inset, a view of the JNK X-ray crystal structure (PDB 3E7O) (Shaw et al., 2008). (B) Micrograph of 50nL of 50nM JNK2. (C) Reference-free 2D class averages of small particles picked from two early-staged embryo datasets containing ~126,000 particles. Classes show unique structural features such as a pentameric ring in the top left corner. (D) Individual raw particles of the 26S proteasome show distinct features directly from micrographs. Particles were visualized using the 'Display' and 'Show particles in selected class' graphical user interface within RELION. Particles have been rotated and translated based on 2D classification in RELION. Box size is 576 Å x 576 Å. (E) Representative micrographs from multiple individual single cell experiments shows similar dispersion and size range of single particles.

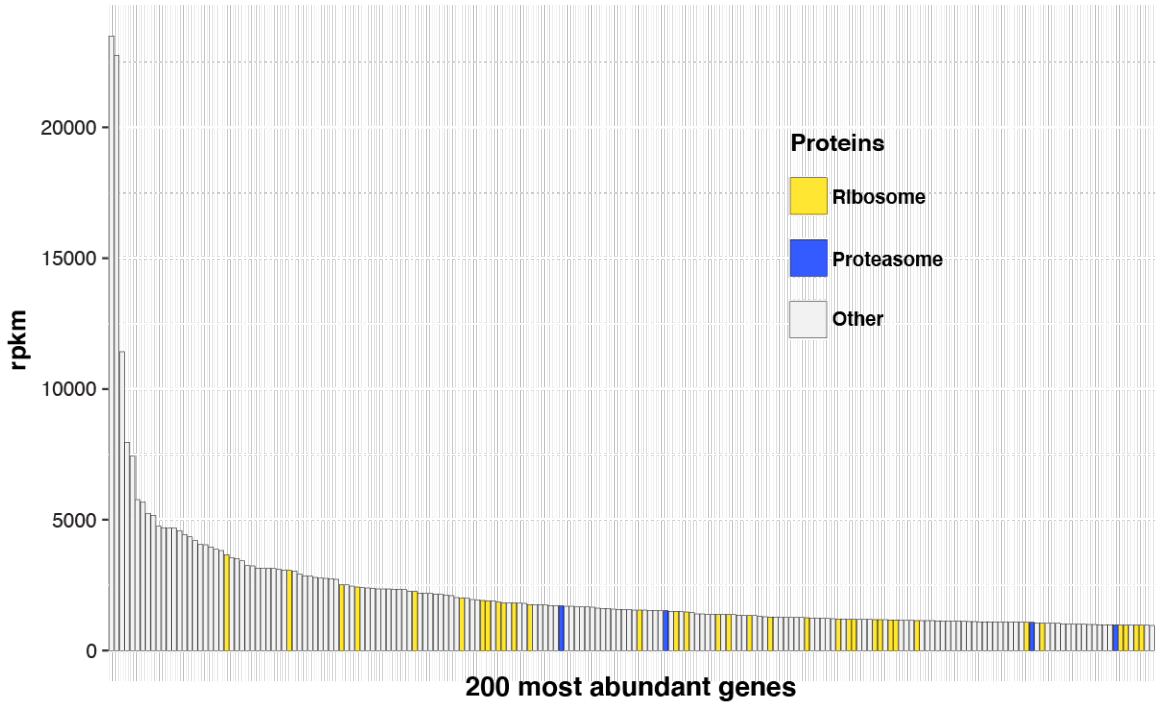

Figure 3.S2. RNA-seq data of *C. elegans* embryos.

The 200 most abundant genes in *C. elegans* zygotes (determined by RNAseq (Gerstein et al., 2010; Hillier et al., 2009)) sorted in order of decreasing abundance.

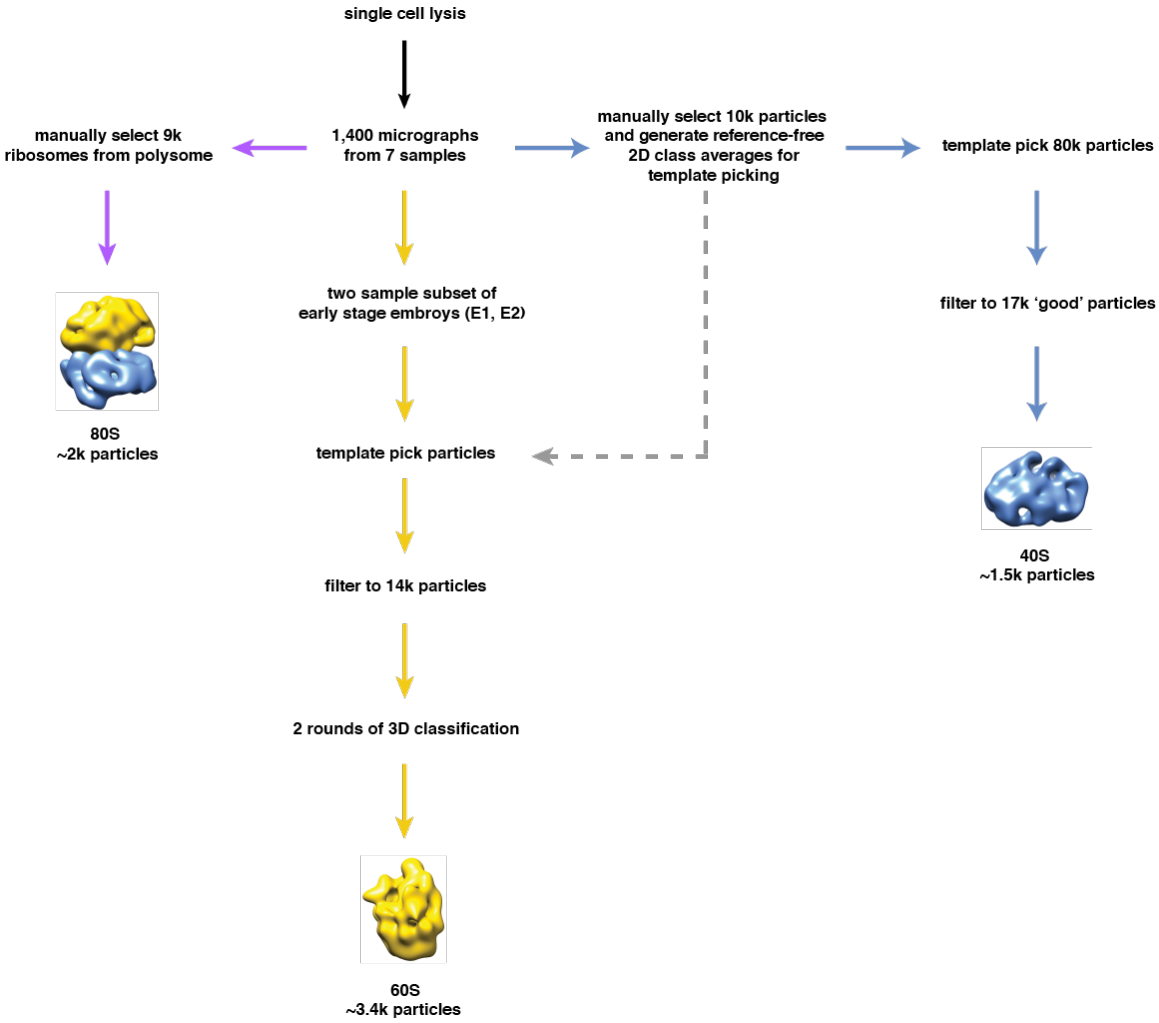

Figure 3.S3. Classification of ribosomes from single cells.

Workflow for classification of particles from single cells into 40S, 60S, and 80S ribosomes structures.

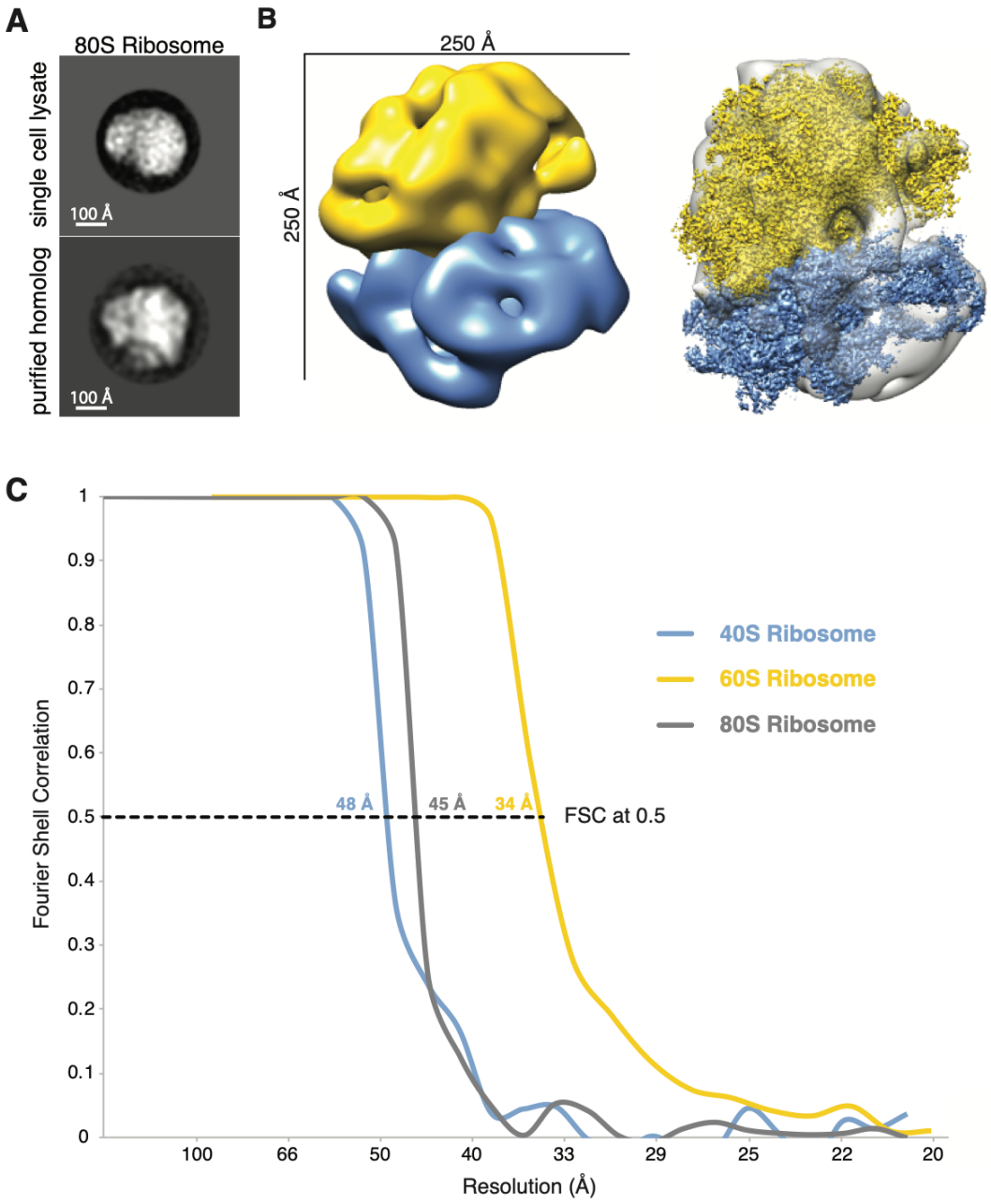

Figure 3.S4. Reconstruction of an 80S ribosome and Fourier shell correlations.

(A) Reference-free 2D class average from ribosomes in polysomes aligned to a purified homolog from *S. cerevisiae*. (B) Left: Our 80S ribosome hybrid model. Right: Our 80S ribosome model reconstructed from ribosomes in polysomes with high-resolution 60S (EMDB-2811) (Shen et al., 2015) and 40S (EMDB-4214) (Scaiola et al., 2018) shown in yellow and blue, respectively. (C) Fourier shell correlations of our 40S, 60S and 80S ribosome models. Nominal resolution values are reported at a correlation score of 0.5.

Chapter 4: Separating distinct structures of multiple macromolecular assemblies from single particle cryo-EM projections

Classically, the goal of single particle cryo-EM is to solve one or more related structures from a highly purified sample. However, a key advantage of single particle cryo-EM over other structural methods is that particles do not need to be homogenous. In this chapter, I demonstrate an algorithm for sorting particle images from multiple distinct structures prior to conventional 3D classification, essentially performing an *in silico* purification. This work is published as Verbeke, E.J., Zhou, Y., Horton, A.P., Mallam, A.L., Taylor, D.W., and Marcotte, E.M. (2020). Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections. Journal of Structural Biology 209, 107416. Anna Mallam helped with protein purification, Yi Zhou prepared samples and collected the cryo-EM data, and I developed the computational pipeline, processed the cryo-EM data and interpreted results. Andrew Horton provided valuable input on early analysis of the pipeline.

4.1. ABSTRACT

Single particle analysis for structure determination in cryo-electron microscopy is traditionally applied to samples purified to near homogeneity as current reconstruction algorithms are not designed to handle heterogeneous mixtures of structures from many distinct macromolecular complexes. We extend on long established methods and demonstrate that relating two-dimensional projection images by their common lines in a graphical framework is sufficient for partitioning distinct protein and multiprotein complexes within the same data set. The feasibility of this approach is first demonstrated

on a large set of synthetic reprojections from 35 unique macromolecular structures spanning a mass range of hundreds to thousands of kilodaltons. We then apply our algorithm on cryo-EM data collected from a mixture of five protein complexes and use existing methods to solve multiple three-dimensional structures *ab initio*. Incorporating methods to sort single particle cryo-EM data from extremely heterogeneous mixtures will alleviate the need for stringent purification and pave the way toward investigation of samples containing many unique structures.

4.2. INTRODUCTION

Cryo-electron microscopy (cryo-EM) has undergone a revolutionary shift in the past few years. Increased signal in electron micrographs, as a result of direct electron detectors, has allowed for the near-atomic resolution structure determination of many macromolecules of various shapes and sizes (Kühlbrandt, 2014). These new detectors combined with automated data collection software and improvements in image processing suggest that cryo-EM could be utilized as a high-throughput approach to structural biology. One emerging field in single particle cryo-EM that seeks to take advantage of these advances is the direct investigation of macromolecules from cellular extracts (Doerr, 2018; Kyrilis et al., 2019). Such an approach is motivated by many observations that fractions from chromatographically separated cell extracts combined with mass spectrometry can be mined for a wealth of information including the organization of macromolecules into larger assemblies (Wan et al., 2015). A natural complement to this information would be direct structural analysis of the macromolecular assemblies from the same fractions of cell extract. Single particle cryo-EM is a promising tool for this goal. Although spatial context is lost when compared to tomography, single

particle approaches are more successful at producing high-resolution structures. However, one major obstacle remains: sorting through the immense heterogeneity that is present in a mixture of tens to hundreds of macromolecular assemblies.

We and others have shown that cellular extracts contain rich structural information which can be used for the identification of multiple structures using conventional single particle analysis (Kastritis et al., 2017; Verbeke et al., 2018). More recently, we extended this approach to reconstruct macromolecular machines from the lysate of a single *C. elegans* embryo (Yi et al., 2019). These studies were limited to the identification of only the most abundant and easily identifiable protein and protein–nucleic acid complexes due to a lack of methods to efficiently categorize which two-dimensional (2D) projection images derive from which three-dimensional (3D) assemblies on the basis of their structural features. While a number of 3D classification schemes exist, all failed to produce reliable reconstructions for the majority of particles in these complicated mixtures. This obstacle emphasizes the long-standing need to sort mixtures of structures in addition to their conformational and compositional heterogeneity.

Several methods have been successfully implemented for sorting heterogeneity in cryo-EM data when there are conformational landscapes or variations in the subunit stoichiometry. These approaches generally fall into three categories. Currently, the most popular approach for sorting heterogeneity in cryo-EM data utilizes a maximum likelihood estimation to optimize the correct classification of particles into multiple structures (Scheres, 2012; Sigworth, 1998; Sigworth et al., 2010). Another approach is to estimate the covariance in cryo-EM data to search for regions of variability between the

models and the data (Katsevich et al., 2015; Liao et al., 2015; Penczek et al., 2006). The last approach, and most relevant to this paper, involves computing similarities between projection images in the data before applying clustering methods to separate the data into homogenous subsets (Aizenbud and Shkolnisky, 2019; Herman and Kalinowski, 2008; Shatsky et al., 2010). All of these approaches have been demonstrated on samples containing a primary structure with multiple conformations or variable subunits. However, little work has been done for sorting heterogeneous samples containing multiple distinct structures.

In particular, this work uses the principle of common lines to score the similarity between many otherwise disparate 2D projection images. The central section theorem states that the Fourier transform of any 2D projection of a 3D object is a 2D section through the center of the 3D Fourier transform of the 3D object. Additionally, the 2D central section is perpendicular to the direction of the projection. It follows a dimension lower that a 1D projection (line projection) of a 2D object is a 1D central section through the 2D Fourier transform of the 2D object. Stated in real space: any two 2D projections of the same 3D object must share a 1D line projection in common (i.e. common lines) (Van Heel, 1987). The central section theorem was initially used for *ab initio* 3D reconstructions but has largely been abandoned in favor of projection matching strategies due to a poor sensitivity to noise (Penczek et al., 1994). For our purposes of investigating structures from lysates, projection matching is largely ineffective because we do not have initial 3D structures or even know how many structures might be present in the data and therefore cannot bootstrap from the models. However, common lines still contain significant information that can be exploited to discriminate 2D projections from a heterogeneous mixture prior to 3D reconstruction by conventional methods.

Here, we develop a pipeline for building 3D reconstructions from rich mixtures of distinct particles by first grouping aligned and averaged 2D projections into discrete, particle-specific classes using the principles of common lines and a novel graphical clustering framework. We demonstrate our method by partitioning reprojections from 35 previously solved structures into their correct groups. Furthermore, we applied this pipeline to an experimental set of cryo-EM micrographs containing a mixture of several macromolecular complexes. We were able to reconstruct multiple 3D structures after our clustering, improving on 3D classification of all particles simultaneously using current 3D reconstruction software. This work adds a new layer to the conventional classification schemes and is a necessary step for moving cryo-EM towards single particle structural biology from samples containing mixtures of many structures.

4.3. RESULTS

4.3.1. Classifying projection images from multiple structures

A major challenge facing "shotgun"-style cryo-EM is to reconstruct models from projection images arising from multiple distinct structures present in a mixture. To overcome this obstacle, we sought a method to computationally group heterogeneous projection images into discrete clusters that each derive from the same structure. In order to partition 2D projections into homogenous subsets, we developed an algorithm for detecting **S**hared **L**ines **I**n **C**ommon **E**lectron **M**aps (SLICEM). Using this algorithm, we score the similarity of 1D line projections between sets of aligned, classified and averaged 2D projection images (referred to as 2D class averages) without knowledge of

the number of underlying 3D objects, or what they look like. Subsequently, these similarity scores can be put into a graphical framework and clustering algorithms can be applied to group related 2D projection images for subsequent 3D reconstructions (Figure 1).

4.3.2. Synthetic data

To test our approach using SLICEM, we generated synthetic reprojections from 35 previously solved structures deposited in the PDB (see Methods). The structures ranged in molecular weight from ~30 to 3000 kDa (Figure 2A). Each PDB structure was low-pass filtered to 9 Å and uniformly reprojected to create 12 2D projection images, forming an initial set of 420 reprojections simulating 2D class averages from a mixture of structures (Ludtke et al., 1999) (Figure S1). Although these reprojections do not perfectly reflect experimentally determined 2D class averages, failure of this test would indicate little power for real data. Each 2D projection is in turn projected down to 1D in 5 degree increments over 360 degrees.

The similarity between all 1D line projections from every 2D reprojection was then scored using different metrics to evaluate their performance for identifying common line projections. The metrics evaluated were Euclidean distance (Eq. (1)), sum of the absolute difference (Eq. (2)), cross-correlation (Eq. (3)) and cosine similarity (Eq. (4)) (see Methods). We additionally tested the performance of the Euclidean distance and cross-correlation after a Z-score normalization of each 1D line projection. Scoring common lines depends heavily on the centering of 2D class averages. We address this in two ways in our algorithm. As an additional layer of image processing, the particle in

each class average is centered by encompassing it in a minimal bounding box. Next, as part of the scoring, if there is a difference in length between a given pair of 1D projections, the smaller of the two vectors is translated pixel-wise relative to the other vector and scored at each position to account for class averages that might be offset relative to other similar class averages. The optimum score during translations is then used as the similarity between the two 1D line projections.

The precision and recall of correctly pairing 2D class averages from the same 3D structures was then computed in order to determine the performance of each metric, and cosine similarity was determined to be the top performing metric (Figure 2B). Euclidean distance and normalized Euclidean distance had identical performance and are overlaid on the plot. Not surprisingly, cross-correlation was the worst performing metric as the dot product between two vectors scales with their magnitude. Thus, 1D projections from larger protein assemblies are more likely to score higher even if there is no true similarity between the 1D projections.

In order to identify sets of 2D projection images from the same 3D particles, we constructed a network from the comparisons between 2D reprojections, or class averages, as follows: Each 2D class average was represented as a node in a directed graph, with each node connected by edges to the nodes corresponding to the 5 most closely-related 2D class averages based on the similarity of their 1D line projections. While the top-scoring metric in our precision/recall analysis was cosine similarity, the network generated from the Euclidean distance similarity most clearly showed communities (clusters of 2D class averages) correctly partitioned by 3D structure (Figure S2). This result is reflected by the well separated distributions of scores for reprojections belonging

to the same structure and scores for reprojections belonging to different structures (Figure 2C). We additionally applied a traditional hierarchical clustering scheme and show the block structure present in the similarity scores between reprojections (Figure S2). These results show that partitioning 2D projection images by scoring the similarity of their 1D line projections is a powerful, unsupervised approach for sorting cryo-EM data from distinct 3D structures within a heterogeneous mixture.

We additionally tested the following cases that are often present in cryo-EM datasets: (1) uneven angular distribution and number of projections (i.e. non-uniform sampling of the structure), (2) molecular symmetry in the structure, and (3) conformational and subunit heterogeneity. In the first test, performance of the algorithm was only slightly diminished over the case of uniform projections (Figure S3). Preferential orientation negatively impacts 3D reconstruction, but has significantly less effect when simply searching for common lines. Our algorithm was also able to effectively distinguish synthetic 2D reprojections for the latter two cases (Figure S4). In the competitive graphical framework, similar but lower scoring projections (e.g. due to a change in conformation) are outcompeted by higher scoring projections in the same conformation. Molecular symmetries may also be beneficial as they increase the chance of finding a common line between structures. Thus, scoring by common lines provides a powerful approach for ranking the similarity of 2D projections in a mixture.

4.3.3. Cryo-EM on a mixture of protein complexes

After validating our SLICEM algorithm on a synthetic dataset, we performed cryo-EM on an experimental mixture of structures and tested our approach as a proof-of-

principle. Our experimental mixture consisted of 40S, 60S and 80S ribosomes at 75 nM, 150 nM and 50 nM, respectively, and apoferritin and β-galactosidase each at 125 nM. We collected ~ 2,400 images and used a template-based particle picking scheme to select ~ 523,000 particles from the entire data set (Roseman, 2004). Raw micrographs showed a mixture of disperse particles with varying size and shape (Figure S5). We then performed 2D classification on the entire set of particles using RELION (Scheres, 2012). After 1 round of filtering junk particles, the remaining ~203,000 particles were sorted into 100 classes using RELION. The class averages contained many characteristic ribosome projections and had distinct structural features (Figure S5). We were unable to identify any β-galactosidase particles in our collected images.

We then applied our SLICEM algorithm to the 100 2D class averages. The identity of each 2D class average was manually annotated, where it was easily recognizable, to assess whether our algorithm was correctly separating the 2D projection images from our heterogeneous mixture (Figure 3). Based on these manual annotations, we again tested the 6-different metrics in a precision-recall framework to determine which metric performed better on experimental data (Figure S6). The Euclidean distance and sum of the absolute difference scoring metrics significantly outperformed the cosine similarity. Using the sum of the absolute difference scoring metric, the network naturally partitioned into 3 distinct communities, one for each ribosome, prior to employing any community detection algorithms (Figure 3).

As part of our algorithm, we evaluated two community detection methods, edge betweenness and walktrap, to determine if the network should be further subdivided (Newman and Girvan, 2004; Pons and Latapy, 2005). We chose to use community

detection algorithms to prevent biasing the data by choosing a specific number of output clusters we expected. Briefly, the algorithms work as follows: For edge betweenness, edges with the highest "betweenness" score in a network are iteratively removed and the betweenness recalculated. At some iteration, the network is separated into separate components (i.e. communities). For walktrap, random walks on a graph tend to stay in the same community if they are densely packed. A similarity score between nodes can then be calculated and used for partitioning of the graph. Both approaches have advantages and disadvantages for our purpose here and the best choice for clustering is largely empirical.

As part of our processing pipeline, we note that the initial choice for the number of 2D class averages, computed here using RELION, can have an effect on the performance of our algorithm. We tested K = 80, 100, 120 and 200 classes to assess the effect on the performance of our algorithm (Figure S7). Despite varying the number of classes, the resulting networks still show correct grouping of 2D class averages from the same 3D structure. At all K values, performance measured by precision and recall is substantially better than random assignment of class averages. However, these results also suggest that moving forward, a more quantitative approach should be taken for selecting the number of 2D class averages. Using our SLICEM algorithm, we demonstrate that it is possible to correctly separate 2D projection images from 3 large, asymmetric macromolecular complexes in the same mixture.

4.3.4. Summed pixel intensity as an additional filtering step

Apart from partitioning 2D projection images into homogenous subsets for 3D reconstruction, one additional goal of shotgun-EM is to determine the identity of each projection image. In previous studies, we and others have leveraged mass spectrometry data to help identify electron microscopy reconstructions from a heterogeneous mixture, such as cell lysate, where the architecture of every protein or protein complex is not known (Kastritis et al., 2017; Verbeke et al., 2018). However, this combined MS-EM approach was only useful for identifying highly abundant and easily recognizable structures.

To provide evidence of macromolecular identity from the electron maps, we calculated the sum of pixel intensities for each manually annotated 2D class average as a proxy for molecular weight (Figure 4). The summed pixel intensities of each annotated 2D class average is plotted as a point on the violin plot to show the distribution of summed pixel intensities between projections of the same structure and between projections of different structures. We found that each of the three ribosomes and apoferritin had unique summed pixel intensities that could be used to distinguish their class averages. Although these values do not directly correspond to molecular weight, and the values will depend on microscope settings or specimen variation, such as ice thickness, class averages belonging to the same structure should have similar values that can be ranked relative to external data (e.g. mass spectrometry data). A least-squares fit to the mean of the summed pixel intensities showed a linear relationship between summed pixel intensity and molecular weight.

The summed pixel intensities were therefore used as an additional filtering step by removing nodes in communities whose summed pixel intensities were outliers in that community. Using this filtering step, the apoferritin class average was removed from the community containing predominantly 40S ribosome reprojections. Our data suggest that, given an appropriate set of standards, summed pixel intensity can be correlated to molecular weight. Thus, summed pixel intensity could be useful in narrowing down the possible identities for a set of electron density maps, when combined with sequence information from mass spectrometry.

4.3.5. 3D classification of a mixture of protein complexes

The ultimate goal of our pipeline is to reconstruct multiple 3D models from our output of clustered 2D projection images. We chose to use cryoSPARC for 3D reconstructions because it can perform heterogeneous reconstruction without *a priori* information on structure or identity (Punjani et al., 2017). We used the particles from each of our 3 distinct communities in addition to the isolated apoferritin node for *ab initio* reconstruction in cryoSPARC (Figure 5). The cluster containing primarily 40S ribosome particles was split into two classes to filter the additional junk particles present in the community. Comparison of our models reconstructed after clustering to the models produced using the entire data set as input for *ab initio* reconstruction in cryoSPARC with 4 classes (one for each protein complex in the mixture) showed our pre-sorting procedure improved the resulting structures (Figure 5). In particular, we were able to build an apoferritin model that was missed in the 3D classification of all particles from cryoSPARC. Our 80S model also shows a more complete density for the small subunit than its counterpart in the model created without clustering. We also observe that

changing the number of classes using *ab initio* reconstruction in cryoSPARC had a substantial impact on the quality of classification (Figure S8).

Each model was refined and evaluated using the gold-standard 0.143 Fourier shell correlation criterion (Figure S9). We obtained easily identifiable 40S, 60S, and 80S ribosome structures at 12, 4, and 5.4 Å resolution, respectively. We were also able to reconstruct the smaller, more compact apoferritin at 19 Å resolution. The ratio of particle numbers for each model was also compared to the input concentrations and shows a bias towards 60S particles (Figure S9). Notably, the 40S and 80S models contain streaks in one dimension, indicating that we are missing several orientations of the particles. We attribute this to preferential orientation of the particles in ice, rather than an inability of our algorithm to properly sort particles into correct communities. Together, these results demonstrate a functioning pipeline for sorting 2D projection images from a heterogeneous mixture of 3D structures, allowing for single particle EM to be applied to samples containing multiple proteins or protein complexes. Importantly, aside from choosing the most appropriate similarity measure, our approach is fully unsupervised, requiring no user defined estimate of the number of existing 3D classes.

4.4. DISCUSSION

As cryo-EM continues to rapidly advance, one potential application would be to perform high-throughput single particle structural biology of the cell. In particular, our goal is to survey macromolecular structures directly from cell lysates. The ability to correctly sort and classify heterogeneous mixtures will become a necessary feature. One advantage of this approach would be to study closer-to-native proteins directly from cells

without the need to purify or alter the sample. Currently, handling compositional and conformational heterogeneity is a major challenge for the EM field, usually requiring expert, time-consuming steps. For our purposes of samples containing many structures, the more sophisticated projection matching algorithms currently used are not effective by themselves as they require an estimate for the number of 3D models expected. Additionally, chromatographic separation of cell lysate is often done on the basis of size, ruling out using the size of 2D projections as a means for separating them.

In this study, we present an unsupervised algorithm, SLICEM, which extends on previous methods and demonstrates that scoring the similarity between 2D class averages based on their 1D line projections contains sufficient information to correctly cluster 2D class averages of the same 3D structure from a mixture of protein and protein-nucleic acid complexes. Using the principal of common lines in a competitive graphical framework provides auxiliary information which can enhance traditional classification. Additionally, as we are not using the common lines to define a relative angle about a tilt axis between 2D projections, many of the pitfalls previously observed with using common lines for 3D reconstruction do not apply. We first demonstrate that the algorithm successfully sorts a synthetic dataset of reprojections created from 35 unique macromolecular structures. Next, we show the same algorithm can successfully partition 2D class averages from an experimental data set containing multiple macromolecular complexes. Pre-sorting 2D projection images prior to 3D classification can allow for current reconstruction algorithms to be employed on datasets containing many unique structures.

Although we demonstrated the feasibility of our approach on synthetic and experimental data, we acknowledge that there are several limitations. In particular, our algorithm relies on the quality of upstream 2D alignment, classification and averaging. One possible approach to better quantify the 2D class averages input to our algorithm would be to sweep multiple values of 2D classes and compare their Fourier ring correlations to see which number of classes has the most similar, high-resolution classes. There will likely be a tradeoff between picking enough classes to cover the heterogeneity present in the data and still having enough signal for accurate common line detection. However, our intent with this algorithm is simply to pre-sort 2D projections belonging to the same structure allowing for more robust 3D classification schemes. As we observed during 2D classification of our cryo-EM data, all apoferritin particles were grouped into a single class average. However, during our network generation step, each class average is given multiple edges to the most similar classes, forcing the single apoferritin class average to have multiple spurious edges. This error will occur any time the number of class averages of a given structure is less than the number of edges used in the graph. Future modifications to the algorithm could include searching for symmetric class averages, where this error is more likely to occur, and removing them prior to community detection.

As we move cryo-EM towards structural determination from complicated mixtures, several other technical challenges will emerge, such as universal freezing conditions. In our mixture of 5 macromolecular complexes, we were unable to easily find freezing conditions that accommodated all proteins. The result was a mixture missing β-galactosidase and containing orientation preferences for the 40S and 80S ribosome. However, previous work has produced e.g. high-resolution structures of fatty-acid

synthase from fractionated cell lysate, suggesting it is possible to find suitable cryo-conditions for solutions containing many macromolecular species (Kastritis et al., 2017). An additional challenge will be developing particle picking algorithms specifically for mixtures, where the particle shape may be unknown and, perhaps more importantly, non-uniform. While in this study we used a template picking scheme, future studies with mixtures of unknown composition will require more sophisticated approaches.

An expert might be able to manually sort the class averages from our cryo-EM data set; however, as mixtures grow in complexity, manual sorting will certainly become infeasible. Introducing algorithms such as SLICEM will provide an unbiased way to group 2D projection images and can be easily implemented in conjunction with a variety of image processing and 3D reconstruction packages. One additional utility of this algorithm could be to remove junk class averages from data in a semi-supervised manner by removal of communities of projection images that do not appear to have structural features. Our approach for sorting mixtures of structures combined with previous approaches for sorting conformational heterogeneity could be a powerful tool for deep classification. Development of methods to sort mixtures of structures in single particle cryo-EM will allow us to solve more structures in parallel and alleviate time-consuming protein purification and sample preparation.

4.5. METHODS

4.5.1. Synthetic data generation

The following list of PDB entries were used to create the dataset of synthetic reprojections (1A0I, 1HHO, 1NW9, 1WA5, 3JCK, 5A63, 1A36, 1HNW, 1PJR, 2FFL, 3JCR, 5GJQ, 1AON, 1I6H, 1RYP, 2MYS, 3VKH, 5VOX, 1FA0, 1JLB, 1S5L, 2NN6, 4F3T, 6B3R, 1FPY, 1MUH, 1SXJ, 2SRC, 4V6C, 6D6V, 1GFL, 1NJI, 1TAU, 3JB9, 5A1A). Each PDB entry was low-pass filtered to 9 Å and converted to a 3D EM density using 'pdb2mrc' in EMAN (Ludtke et al., 1999). These densities were then uniformly reprojected using 'project3d' in EMAN to create 12 2D reprojections for each structure (Ludtke et al., 1999). Reprojections were centered in 350 Å boxes.

4.5.2. Purification of apoferritin and β-galactosidase

Size-exclusion chromatography was performed at 4 °C on an AKTA FPLC (GE Healthcare). Approximately 10 mg of apoferritin (Sigma A3660-1VL) and 5 mg of β-galactosidase G5635-5KU were independently applied to a Superdex 200 10/300 GL analytical gel filtration column (GE Healthcare) equilibrated in 20 mM HEPES KOH, 100 mM potassium acetate, 2.5 mM magnesium acetate, pH 7.5 at a flow rate of 0.5 mL min^{-1}. Fractions were collected every 0.5 mL.

4.5.3. SLICEM algorithm

Our algorithm consists of five main steps: (1) Extracting 2D class average signal from background, (2) Generating 1D line projections from the extracted 2D projection

images, (3) Scoring the similarity of all pairs of 1D line projections, (4) Building a nearest-neighbors graph of the 2D class averages and (5) Partitioning communities within the graph.

4.5.3.1. Extracting 2D class averages from background

The input to our algorithm is a set of centered and normalized 2D class averages. The images are normalized according to the RELION conventions of setting particles to a mean value of zero and a standard deviation of one for all pixels in the background area. We then extract the centered region of positive pixels values from the zero-mean normalized images to remove background signal and extra densities that might be present in a class average. This step also serves to re-center the class average by surrounding it with a minimal bounding box.

4.5.3.2. Generating 1D line projections from extracted 2D projection images

Each newly extracted class average is then projected into 1D over 360 degrees in 5 degree intervals by summing the pixel values along the projection axis. The 1D line projections are then ready to be scored or are independently zero-mean normalized if the normalized cross-correlation or normalized Euclidean distance scoring metric are selected.

4.5.3.3. Scoring the similarity of all pairs of 1D line projections

To score the similarity of the 1D line projections we consider 6 different scoring metrics. The metrics evaluated were Euclidean distance (Eq. (1)), sum of the absolute difference (Eq. (2)), cross-correlation (Eq. (3)) and cosine similarity (Eq. (4)). We additionally consider Euclidean distance and cross-correlation after a Z-score normalization of each 1D line projection. For two 1D line projection vectors p and q, the difference d between the vectors can be calculated as follows:

1) $d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$

2) $d(p,q) = \sum_{i=1}^{n}|p_i - q_i|$

3) $d(p,q) = p_i q_i$

4) $d(p,q) = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$

The similarity of the 1D line projections is calculated for all pixel-wise translations of the smaller 1D projection across the larger 1D projection if there is a difference in projection size, analogous to the 'sliding' feature of standard cross-correlations. The optimum score during the translations is kept for each pair of 1D projections. After pairwise scoring of all 1D line projections from all 2D class averages, the similarity between each pair of 2D class averages is defined by their respective best scoring 1D line projections.

4.5.3.4. Building a nearest-neighbors graph of the 2D class averages

SLICEM then constructs a directed graph using the similarity scores calculated for each pair of 2D class averages. Each node (2D class average) is connected to the 5 most similar (top scoring) 2D class averages. Each edge is assigned a weight computed as a Z-score relative to all scores for a given 2D class average.

4.5.3.5. Partitioning communities within the graph

The resulting graph is then subdivided using a community detection algorithm. Specifically, we evaluated the edge-betweenness and walktrap algorithms to define clusters in the graph. The default parameters for each clustering method implemented in iGraph were used in our algorithm, however we note that different similarity metrics and 'clustering strengths' can be applied. For edge-betweenness, the dendrogram is cut at the level which maximizes the modularity and for walktrap, the length of the random walks is set to 4. Then, the median absolute deviation of summed pixel intensities for each node is calculated to remove outliers from clusters. Finally, for each community, the individual raw 2D particles corresponding to the now-grouped 2D class averages are then used as input for 3D reconstruction in cryoSPARC.

4.5.4. Cryo-EM grid preparation and data collection

C-flat holey carbon grids (CF-1.2/1.3, Protochips Inc.) were pre-coated with a thin layer of freshly prepared carbon film and glow-discharged for 30 s using a Gatan Solarus plasma cleaner before addition of sample. 2.5 µL of a mixture of 75 nM 40S

ribosome, 150 nM 60S ribosome, 50 nM 80S ribosome, 125 nM apoferritin and 125 nM β-galactosidase were placed onto grids, blotted for 3 s with a blotting force of 5 and rapidly plunged into liquid ethane using a FEI Vitrobot MarkIV operated at 4 °C and 100% humidity. Data were acquired using an FEI Titan Krios transmission electron microscope (Sauer Structural Biology Laboratory, University of Texas at Austin) operating at 300 keV at a nominal magnification of ×22,500 (1.1 Å pixel size) with defocus ranging from −2.0 to −3.5 μm. The data were collected using a total exposure of 6 s fractionated into 20 frames (300 ms per frame) with a dose rate of ~8 electrons per pixel per second and a total exposure dose of ~40 e^{-} $Å^{-2}$. A total of 2423 micrographs were automatically recorded on a Gatan K2 Summit direct electron detector operated in counting mode using the MSI Template application within the automated macromolecular microscopy software LEGINON (Suloway et al., 2005).

4.5.5. Cryo-EM data processing

All image pre-processing was performed in Appion (Lander et al., 2009). Individual movie frames were aligned and averaged using 'MotionCor2' drift-correction software (Zheng et al., 2017). These drift-corrected micrographs were binned by 8, and bad micrographs and/or regions of micrographs were removed using the 'manual masking' command within Appion. A total of 522,653 particles were picked with a template-based particle picker using a reference-free 2D class average from a small subset of manually picked particles as templates. The contrast transfer function (CTF) of each micrograph was estimated using CTFFIND4 (Rohou and Grigorieff, 2015). Selected particles were extracted from micrographs using particle extraction within RELION (Scheres, 2012) and the EMAN2 coordinates exported from Appion. Two rounds of

reference free 2D classification with 100 classes for each sample were performed in RELION to remove junk particles, resulting in a clean stack of 202,611 particle images.

4.6. FIGURES

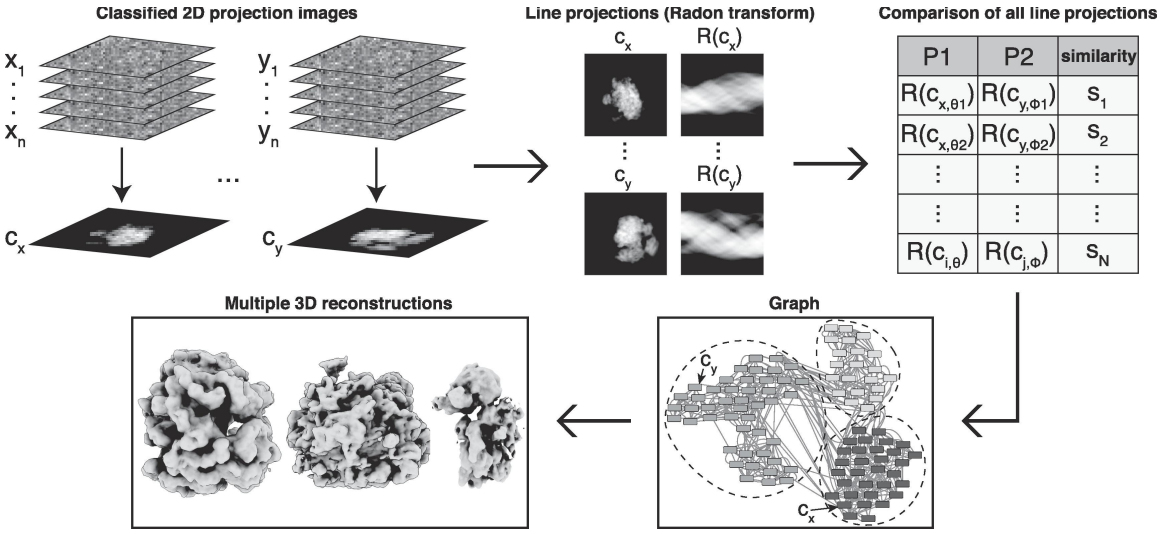

Figure 4.1. Computational pipeline for SLICEM.

Individual particle images are averaged after reference-free 2D alignment and classification. Using a Radon transform, 1D line projections are created from the 2D class averages (also referred to as 2D projections). Each 1D line projection from every 2D projection is then scored for similarity. The top scores between 2D projections are then used to create edges connecting 2D projections that have a similar 1D line projection, forming a graph. 2D projection images are then partitioned into groups belonging to the same putative structure using a community detection algorithm. Individual particle images belonging to each 2D projection within a community are subjected to *ab initio* 3D reconstruction.

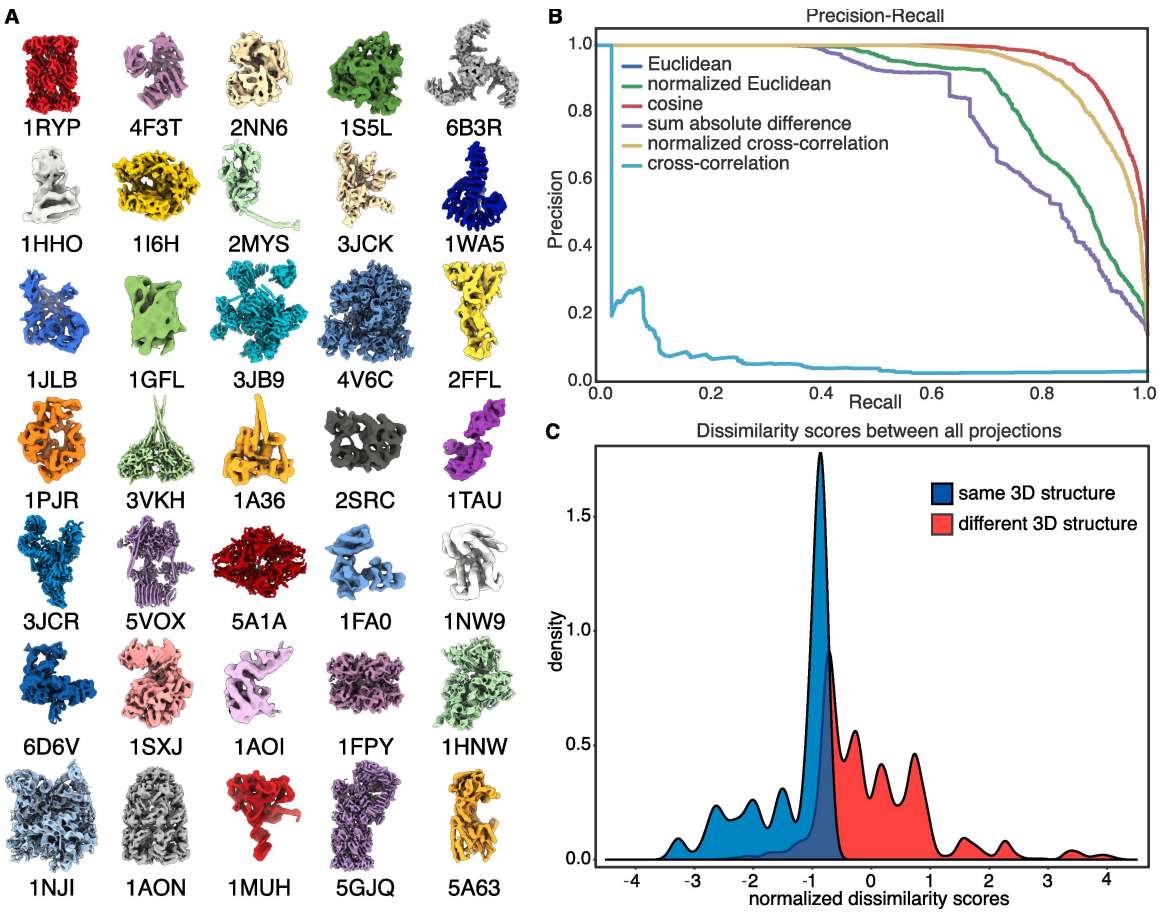

Figure 4.2. Separating mixtures of synthetic 2D reprojections.

Synthetic reprojections were generated from 35 distinct PDB structures low-pass filtered to 9 Å from protein and protein assemblies ranging in molecular weight from ~30 to 3000 kDa, prior to separation using SLICEM. (A) Low-pass filtered models of each PDB structure. (B) Precision-recall plot ranking 6 different metrics at scoring the similarity between 1D line projections from each 2D reprojection. (C) Distribution of scores calculated using Euclidean distance for reprojections belonging to the same structure and reprojections belonging to different structures.

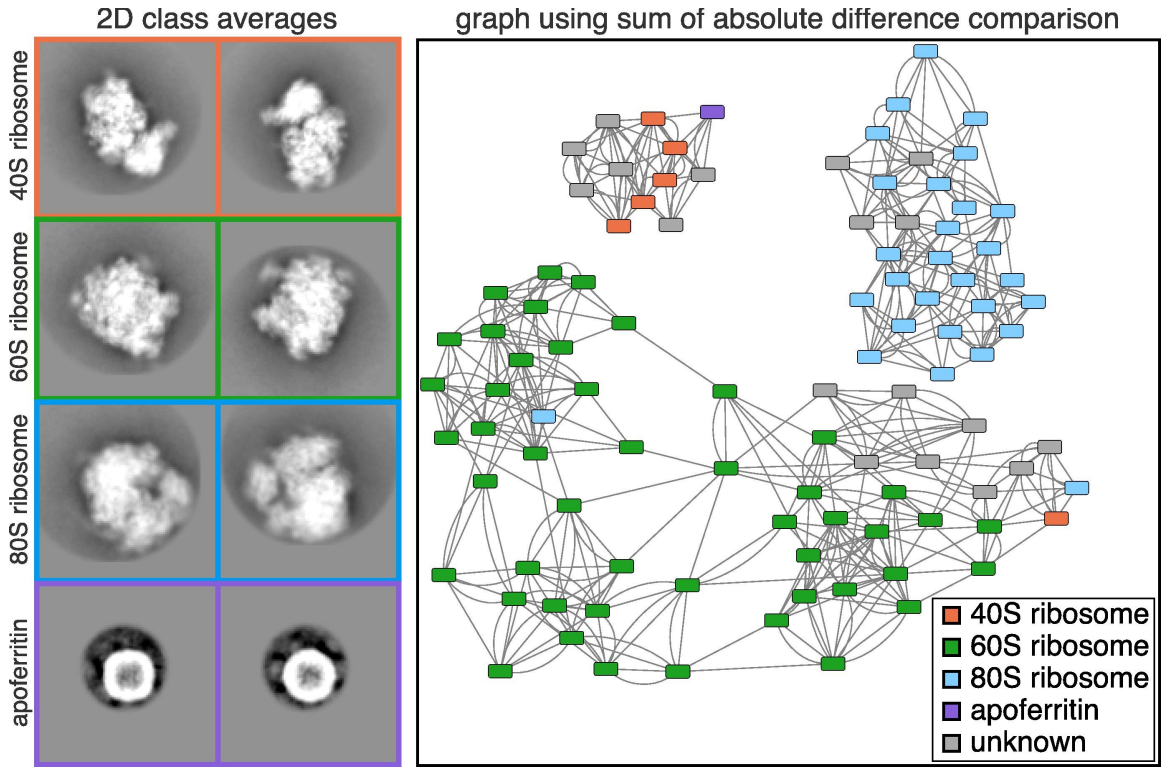

Figure 4.3. Experimental 2D class averages and resulting network.

Cryo-EM data was collected on a mixture of 5 protein and protein-nucleic acid complexes. Representative 2D class averages of the 4 complexes identified in the mixture are shown on the left. The identity of each class average was manually annotated were it could be easily identified. The class average corresponding to apoferritin was further subdivided into multiple classes for visualization. Each box corresponds to a width of 422 Å. The network displayed was generated after using SLICEM on the 100 2D class averages scored using the sum of the absolute difference metric. Nodes representing each 2D class averages are colored by their putative structural identity and are connected to their 5 most similar class averages.

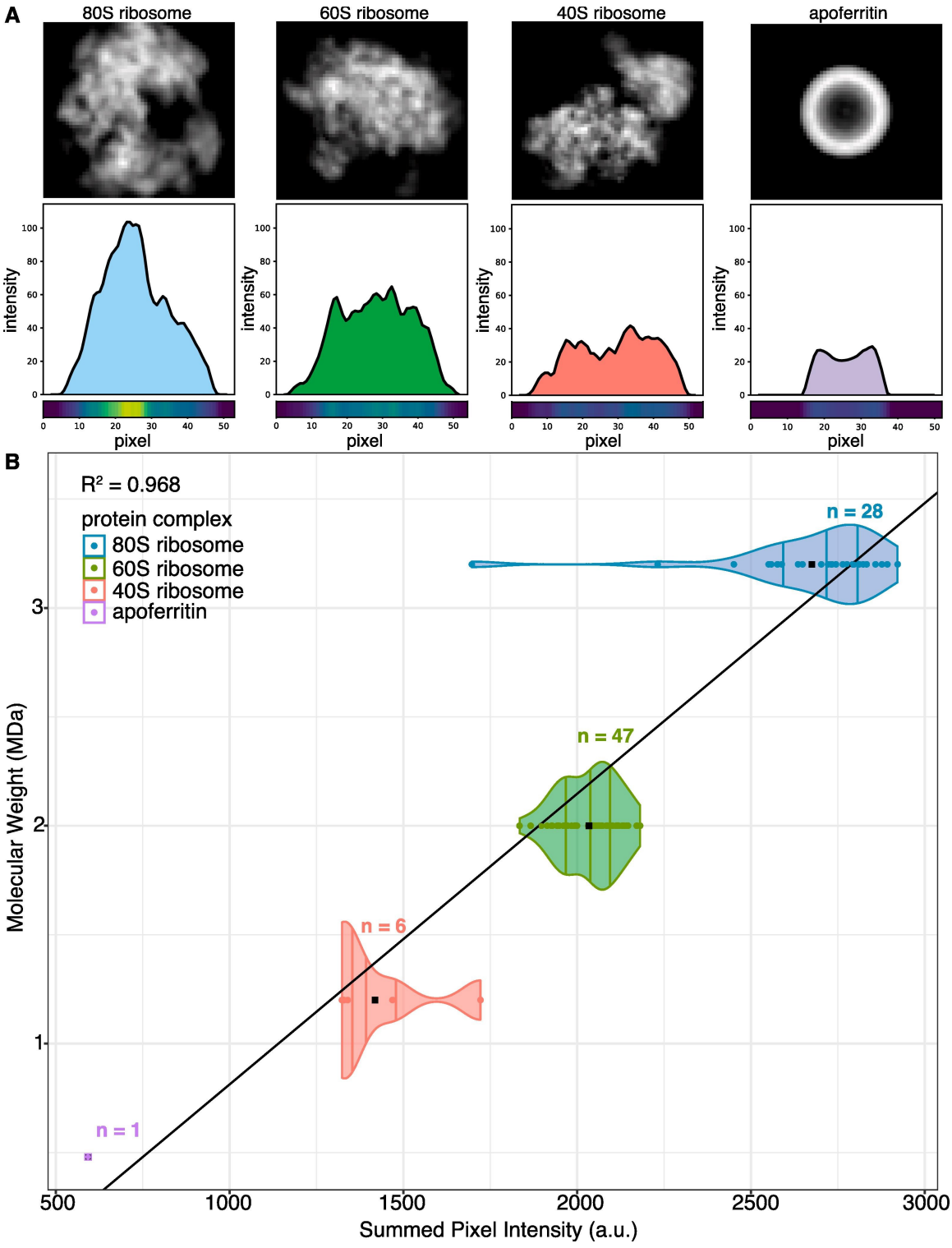

Figure 4.4. Summed pixel intensities of 2D class averages correlate to molecular weight.

(A) 2D to 1D projections (projection angle orthogonal to the x-axis) for representative 2D class averages of each structure present in the mixture. 1D projection plots show the line profile for a single 1D projection of each 2D class average. Pixel heat maps show the intensity of the line profile at each pixel. (B) Distribution of the summed pixel intensities calculated for each 2D class average. Summed pixel intensities for each manually identified 2D class average are plotted against their respective molecular weight. Black points are the mean summed pixel intensity for each structure and n indicates the number of 2D classes for each structure.

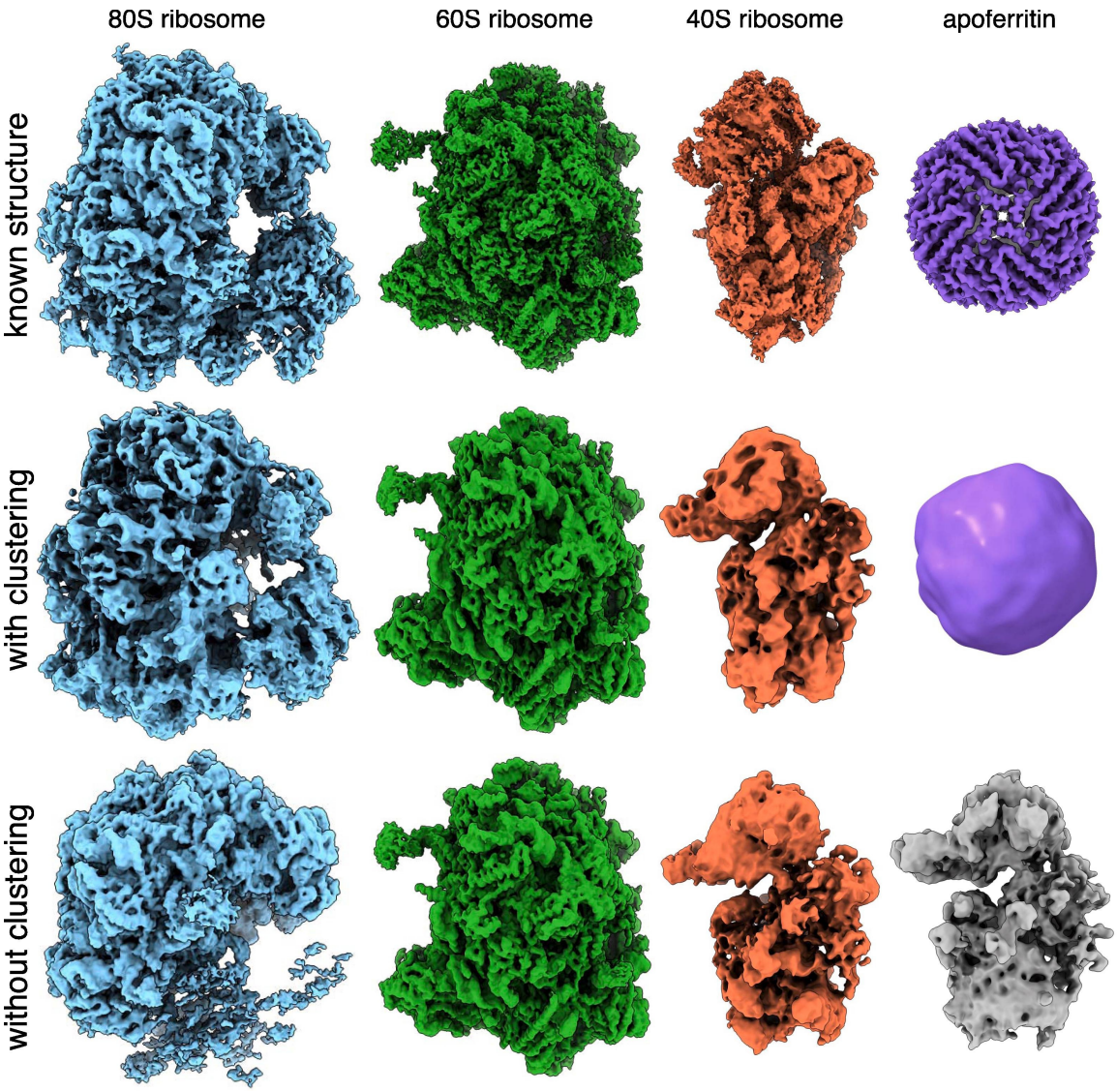

Figure 4.5. *Ab initio* structures from an experimental mixture.

(Top) High-resolution structures of the 80S ribosome EMD-2858 (Cianfrocco and Leschziner, 2015), 60S ribosome EMD-2811 (Shen et al., 2015), 40S ribosome EMD-4214 (Scaiola et al., 2018) and apoferritin EMD-2788 (Russo and Passmore, 2014). (Middle) 3D models of the 80S ribosome, 60S ribosome, 40S ribosome and apoferritin generated by sorting particles using SLICEM prior to *ab initio* 3D reconstruction in cryoSPARC. (Bottom) 3D models generated using *ab initio* reconstruction to generate 4 classes in cryoSPARC without pre-sorting particles using SLICEM.

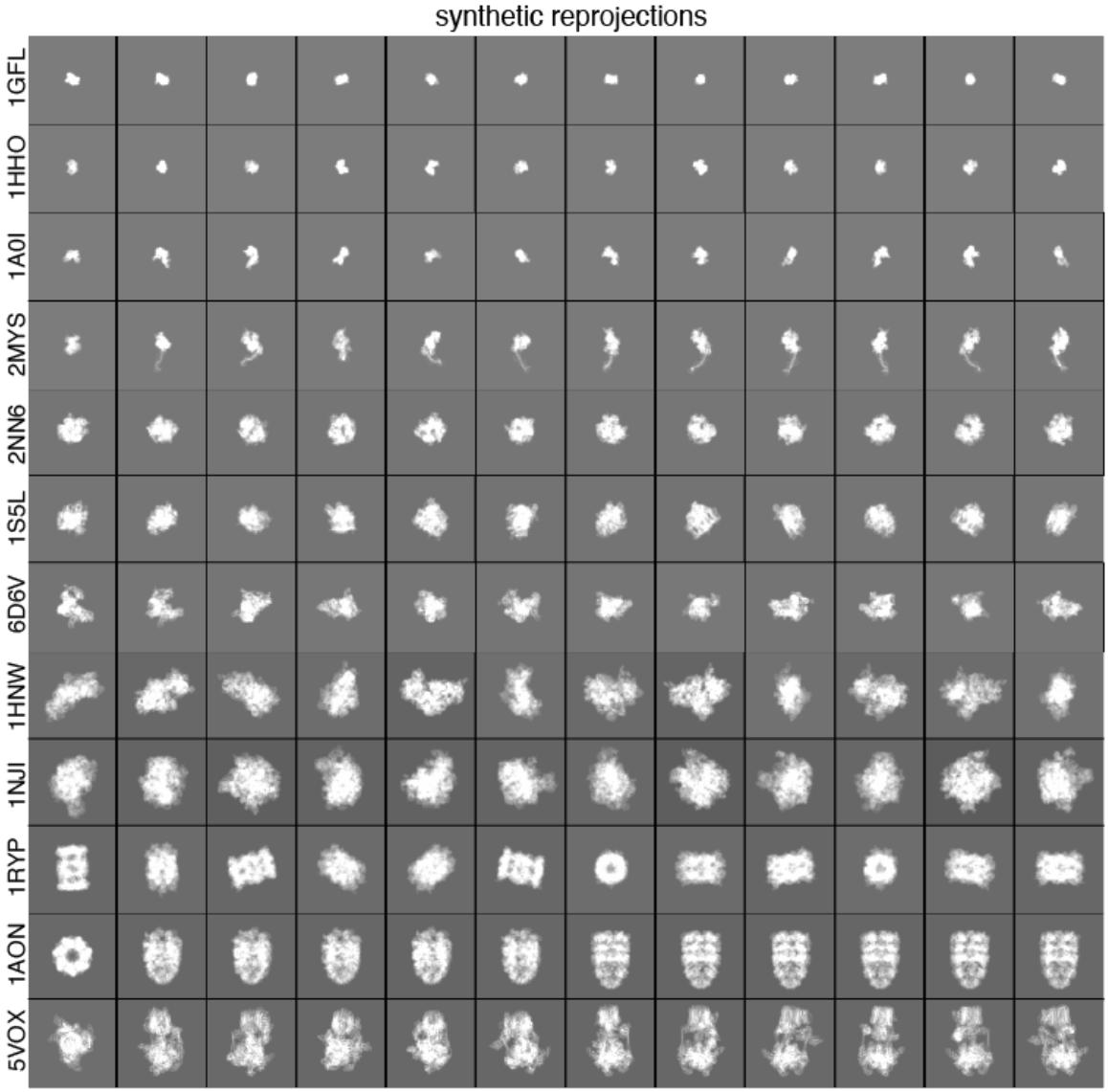

Figure 4.S1. 2D reprojections from synthetic dataset.

Subset of 2D reprojections from 12 of the 35 structures in our synthetic dataset. Box size corresponds to 300 Å.

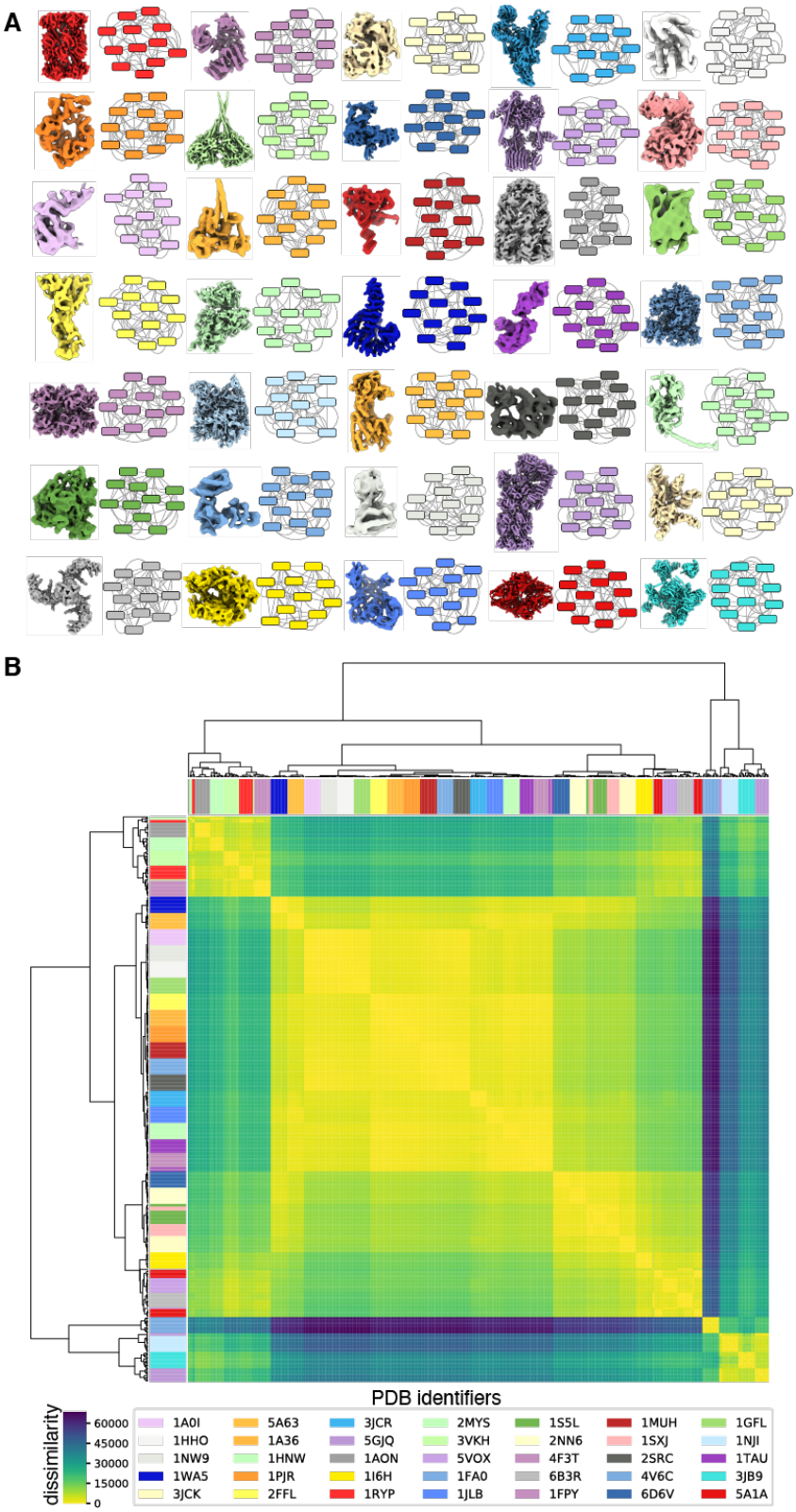
A
B
PDB identifiers
dissimilarity
60000
45000
30000
15000
0
1A0I
1HHO
1NW9
1WA5
3JCK
5A63
1A36
1HNW
1PJR
2FFL
3JCR
5GJQ
1AON
1I6H
1RYP
2MYS
3VKH
5VOX
1FA0
1JLB
1S5L
2NN6
4F3T
6B3R
1FPY
1MUH
1SXJ
2SRC
4V6C
6D6V
1GFL
1NJI
1TAU
3JB9
5A1A

Figure 4.S2. Synthetic dataset network and clustergram.

(A) Network displaying communities of 2D reprojection images determined using SLICEM. Each node represents a 2D reprojection with 5 connecting edges to the most similar reprojections as scored using Euclidean distance. The color of each node matches the structure from which it was reprojected (shown as a surface). (B) Clustergram showing the block structure of similarity between synthetic 2D reprojections based on the scoring of their 1D line projections by Euclidean distance. The distance metric used for hierarchical clustering was Euclidean and the linkage method used was "average". Row and column colors correspond to PDB structure identity and individual pixels reflect the dissimilarity between them (i.e. the lower the dissimilarity, the better the match between 1D line projections).

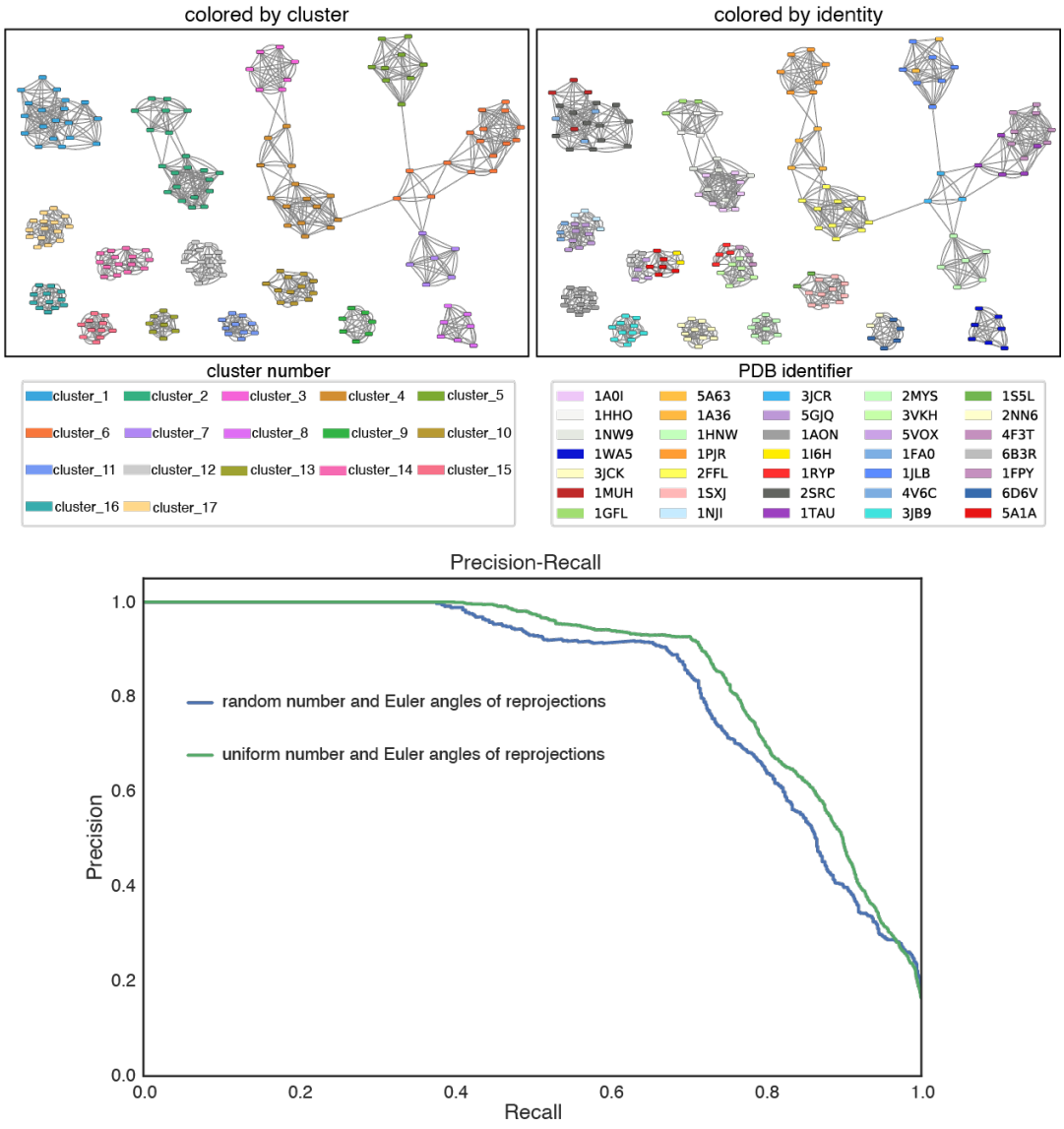

Figure 4.S3. Effect of non-uniform projection angles and number of projections.

The 35 PDB structures from Figure 2A were randomly reprojected between 2-12 times and at random Euler angles each between 0 degrees – 360 degrees. These reprojections were then subjected to our SLICEM algorithm using Euclidean distance to score 1D line projections and walktrap for clustering. The side-by-side networks show the nodes labeled after clustering and the nodes labeled by PDB identity. Projections belonging to the same PDB structure are well grouped but not always clustered correctly. Precision and recall was used to compare the non-uniform dataset to the uniform dataset and shows only a slight decrease in performance for the more realistic case of non-uniform projections.

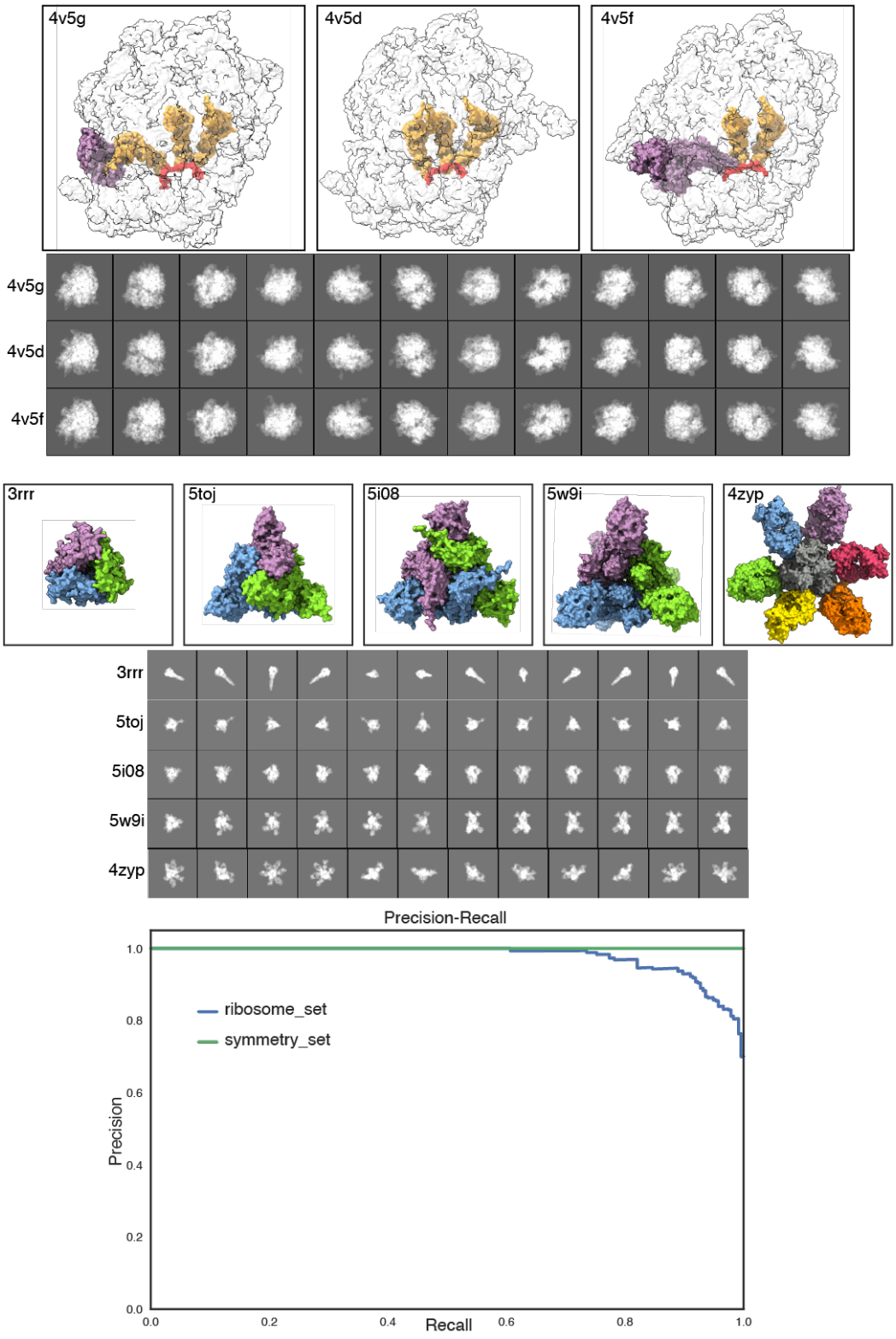
4v5g
4v5d
4v5f
4v5g
4v5d
4v5f
3rrr
5toj
5i08
5w9i
4zyp
3rrr
5toj
5i08
5w9i
4zyp
Precision-Recall
ribosome_set
symmetry_set
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Figure 4.S4. Mixtures with molecular symmetries or conformational and compositional heterogeneity.

Uniform reprojections of a ribosome, a ribosome with EF-Tu, and a ribosome with EF-G (PDB 4V5D, PDB 4V5G, PDB 4V5F) low-pass filtered to 9 Å were created to test the effect of conformational and compositional heterogeneity on our algorithm. Similarly, uniform reprojections of five C-3 symmetric structures (PDB 3RRR, PDB 5TOJ, PDB 5I08, PDB 5W9I, PDB 4ZYP) were created to test the effect of molecular symmetry on our algorithm. The box size of all reprojections is 350 Å. The similarity of 1D line projections between 2D reprojections in each set were calculated using Euclidean distance and performance was measured using precision and recall. Here, the high scoring precision-recall curves indicate that common line scores from projections of the same structure outperform scores to similar, but slightly different structures.

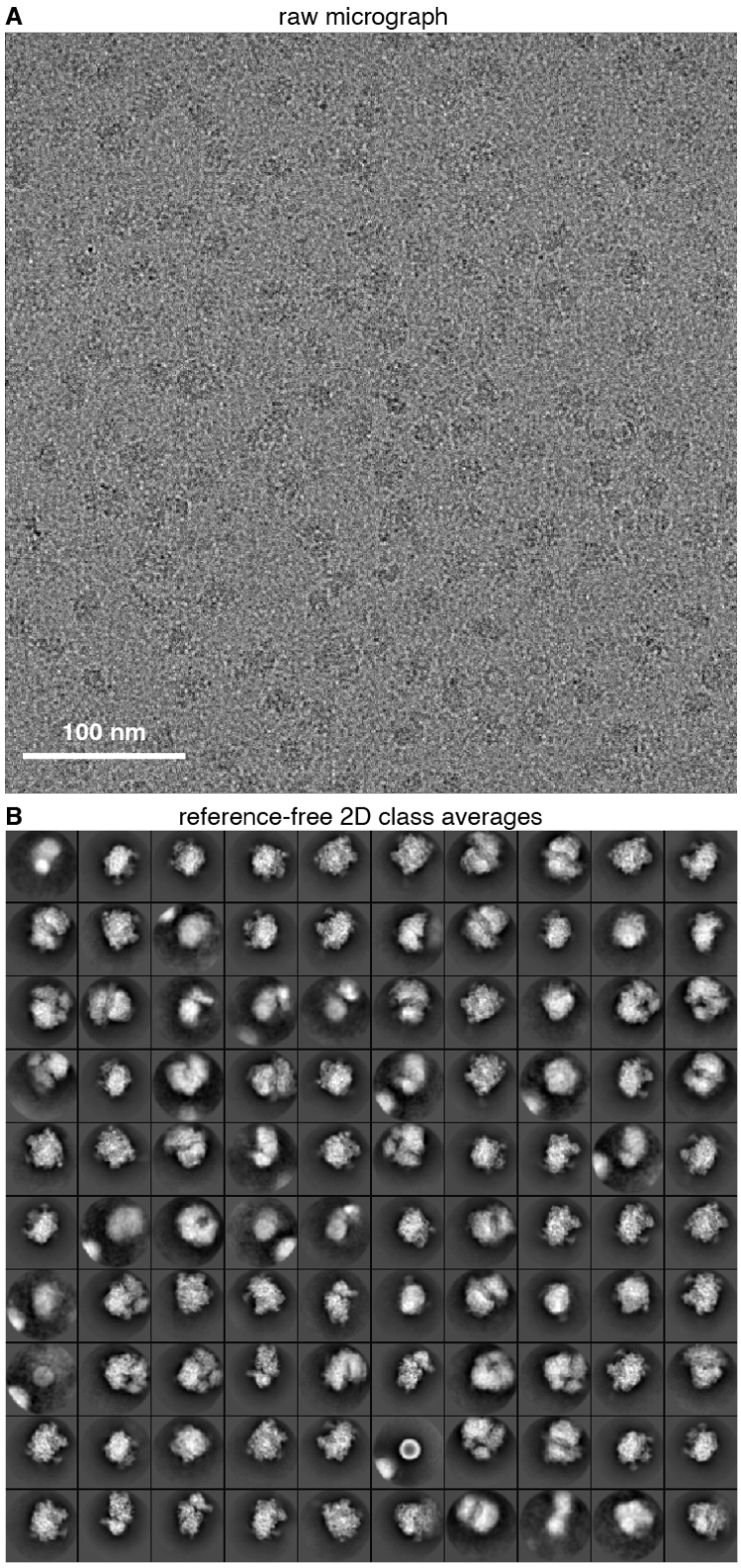
A
raw micrograph
100 nm
B
reference-free 2D class averages

Figure 4.S5. 2D classification of particles using RELION.

(A) Representative raw micrograph of a mixture containing 40S, 60S and 80S ribosomes, apoferritin and β-galactosidase. (B) Reference-free 2D class averages generated using RELION of ~203,000 template-picked particle images. Box size corresponds to 422 Å.

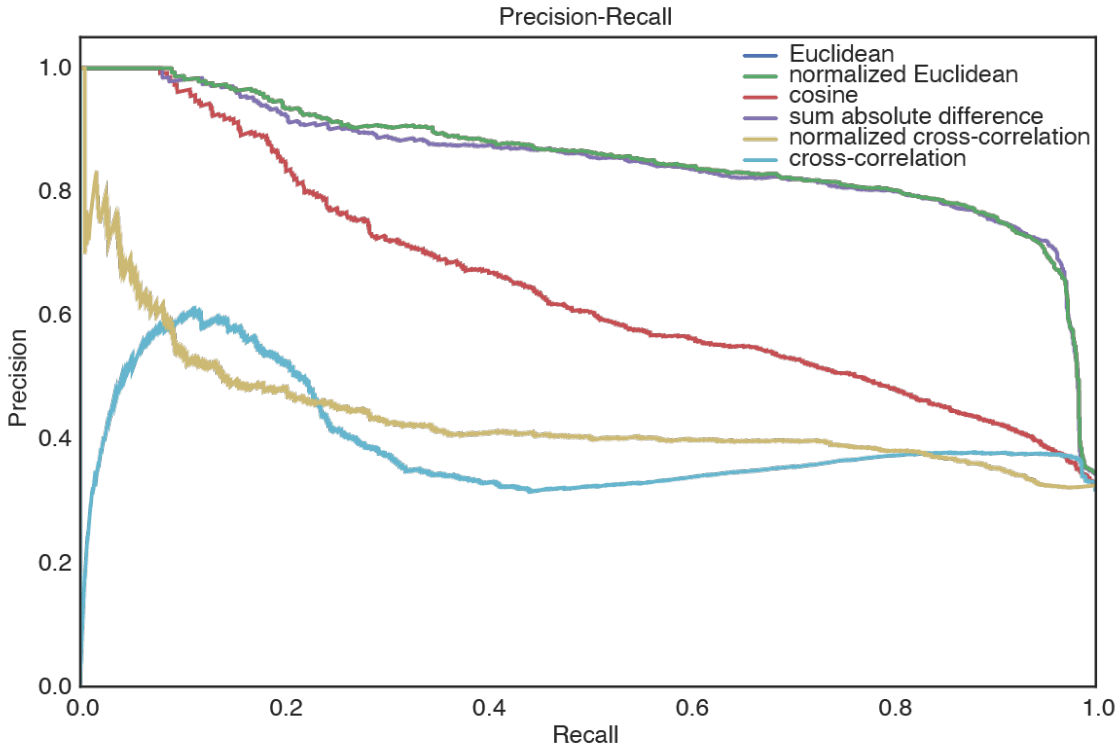

Figure 4.S6. Precision-recall curves for experimental cryo-EM data.

Precision-recall plot displaying 6 different metrics for scoring the similarity between 1D line projections from the entire set of 2D class averages. Euclidean and normalized Euclidean metrics scored identically and are overlaid.

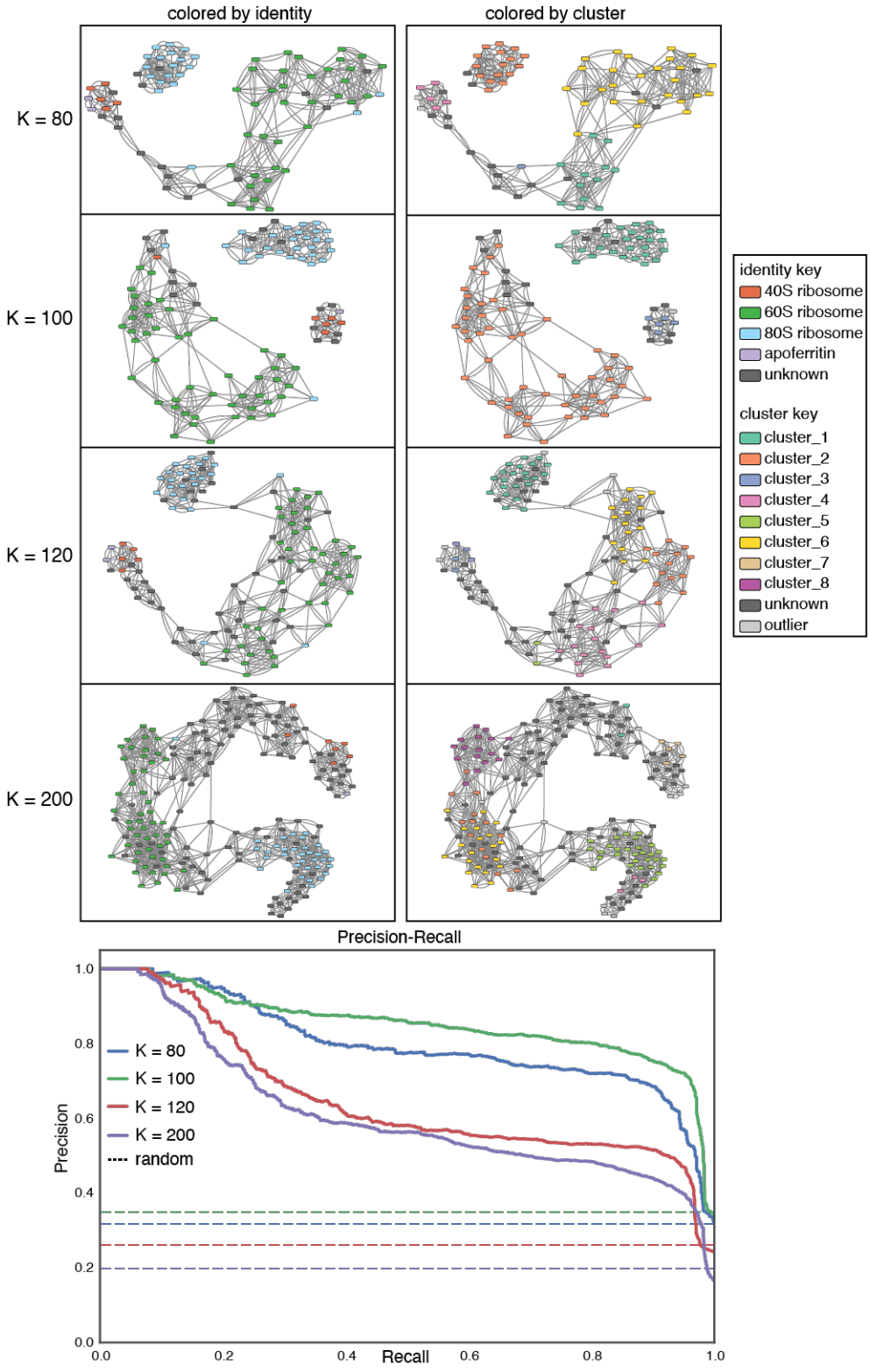
colored by identity
colored by cluster
K = 80
K = 100
K = 120
K = 200
identity key
40S ribosome
60S ribosome
80S ribosome
apoferritin
unknown
cluster key
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
unknown
outlier
Precision-Recall
Precision
Recall
K = 80
K = 100
K = 120
K = 200
random
0.0
0.2
0.4
0.6
0.8
1.0

Figure 4.S7. Effect of varying the number of 2D class averages.

Using RELION, 2D class averages with K = 80, 100, 120 and 200 were created from the final stack of ~203,000 particle images. The 2D class averages were then subjected to our SLICEM algorithm using the sum of absolute difference to score 1D line projections and edge betweenness for clustering. The side-by-side networks show the nodes labeled by manually annotated identity and by cluster. Precision and recall was used to compare the performance at various numbers of 2D classes. All precision-recall curves show substantial improvement over random assignment.

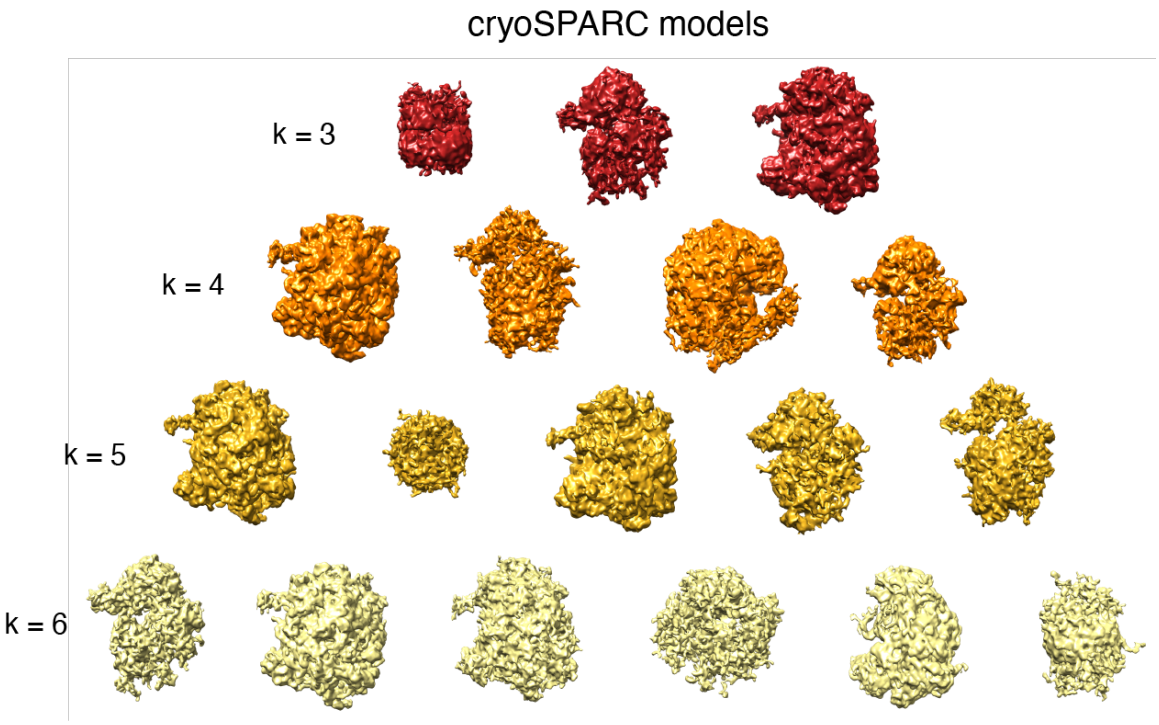

Figure 4.S8. *Ab initio* reconstructions in cryoSPARC with varying class number.

3D reconstructions using ab initio reconstruction in cryoSPARC from the entire data set with K = 3, 4, 5 and 6 classes, respectively.

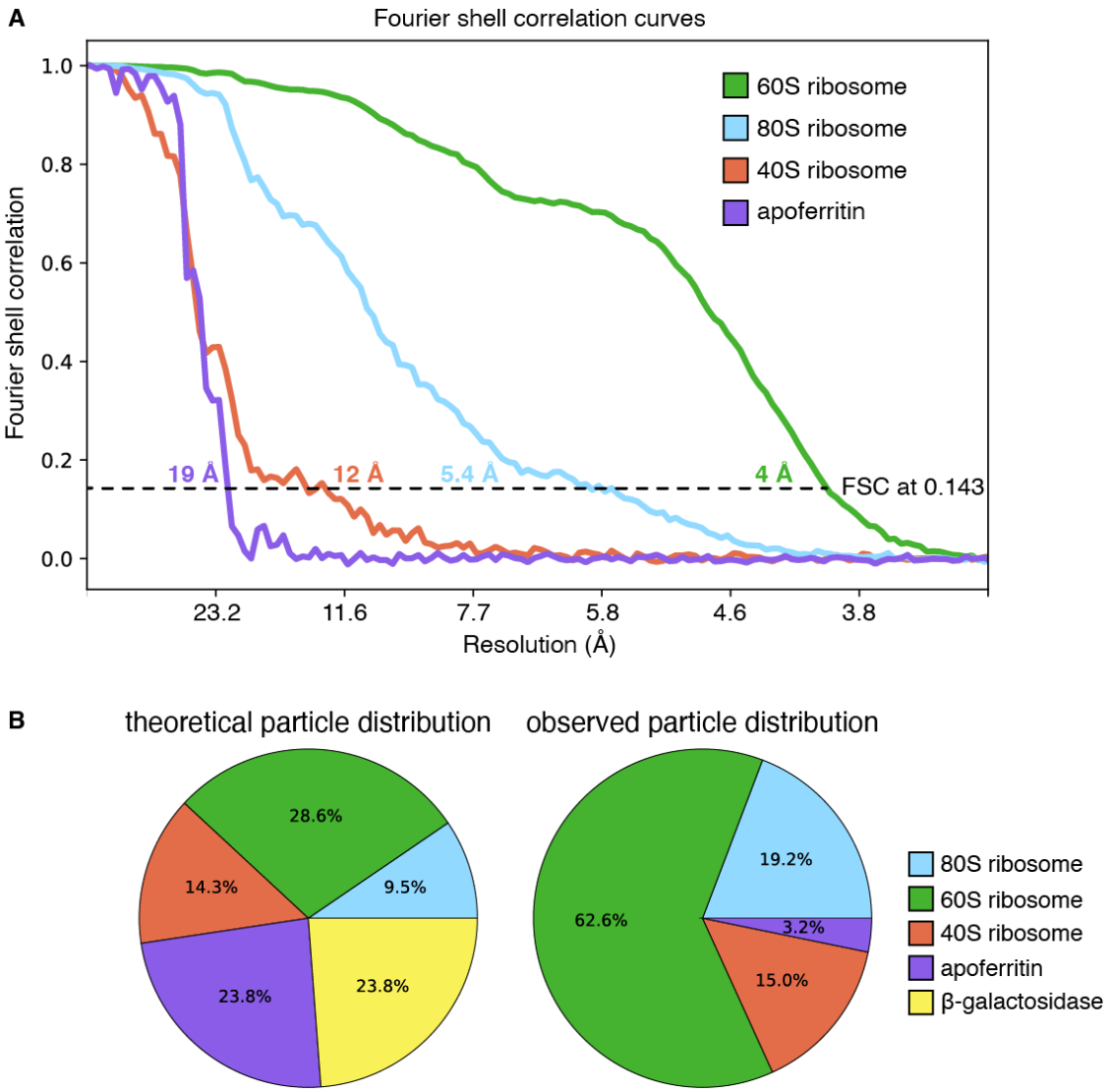

Figure 4.S9. Fourier shell correlation curves

(A) FSC curves for our clustered 80S ribosome (blue), 60S ribosome (green), 40S ribosome (red) and apoferritin (purple) shown in Figure 5. Nominal resolutions were estimated to be 5.4, 4, 12 and 19 Å, respectively, using the 0.143 gold-standard FSC criterion. β-galactosidase was not observed in our dataset and therefore no reconstruction was calculated. (B) Theoretical ratio of particles from the input concentrations compared to observed ratio of particles after clustering of the data. The input concentrations for the 80S, 60S, 40S, apoferritin and β-galactosidase were 50, 150, 75, 125 and 125 nM and the observed particles for each were 38957, 126776, 30353, 6525 and 0, respectively.

Chapter 5: Molecular architecture of the red blood cell proteome

The initial results shown in this chapter are part of an ongoing effort led by Wisath Sae-Lee to characterize the molecular architecture of the red blood cell proteome using co-fractionation mass spectrometry. As described below, red blood cells are uniquely suited to investigation by co-fractionation mass spectrometry and shotgun cryo-EM. Wisath Sae-Lee was the lead on all biochemistry aspects of the work with help from Ophelia Papoulas. Wisath Sae-Lee also performed the computational analysis and interpretation of the mass spectrometry data. I performed the electron microscopy data collection, analysis and interpretation. Some of the text in the section describing the molecular architecture of the red blood cell proteome was written by Wisath Sae-Lee.

5.1. ABSTRACT

Erythrocytes (red blood cells; RBCs) are the simplest primary human cells, lacking nuclei and all major organelles. Despite their simplicity, RBCs dynamically change cellular morphology and physiology throughout their journey in the body. These cellular dynamics are mediated by protein assemblies whose complete picture in RBCs is still unknown. While hemoglobin accounts for 98% of expressed RBC proteins, the full proteome includes >1,000 distinct proteins in which the roles of many remain elusive. In this study, we first identified a comprehensive RBC proteome of 1,202 proteins (1% FDR) using machine learning from quantitative mass spectrometry and RNA-seq on RBCs and other blood cell types. We then determined the stable protein complexes in mature RBCs, based on mass spectrometry of 1,944 native biochemical fractions of hemoglobin-depleted hemolysate and detergent solubilized membrane protein complexes,

validating large protein assemblies with electron microscopy. Our data reveal an RBC interactome dominated by protein homeostasis, redox biology, cytoskeletal dynamics, and carbon-metabolism. As the first near-complete interactome for any primary human cell type, this map of RBC protein complexes provides a better understanding of the unique constraints of RBC function and serves as a comprehensive resource for future research.

5.2. CO-FRACTIONATION MASS SPECTROMETRY OF RED BLOOD CELLS

Although powerful techniques such as affinity purification mass spectrometry (AP-MS) and proximity labeling are available to study protein-protein interactions (PPIs) in other cell types, these techniques are not amenable for RBCs because of the lack of nucleus. Therefore, we turned to another powerful technique to study protein complexes, co-fractionation mass spectrometry (CF-MS). CF-MS is a high-throughput technique that combines the use of biochemical fractionations and bioinformatics to characterize PPIs through their co-elution behavior in multiple orthogonal separations. CF-MS does not require antibodies nor transgenic epitope tagging of individual proteins, thus uniquely appropriate to RBCs. The co-elution (co-fractionation) of proteins in a separation serves as evidence for physical association (Wan et al., 2015). The power of this technique comes from the integration of the co-elution profiles from multiple orthogonal biochemical separations. This makes it possible to distinguish between real PPIs and random co-elution. In addition, the quantification of PPIs via machine learning methods allows us to have a strong control over false discovery rates of protein interactions.

We generated a large proteomics dataset of fractionated RBC proteins using various methods of biochemical fractionation (Figure 1A). Non-denatured protein extracts from hemolysate (soluble proteins) and non-ionic detergent dissolved ghosts (membrane proteins) were separated by biophysical properties such as size and. Each chromatographic fraction was analyzed by high-resolution, high-sensitivity liquid chromatography/mass spectrometry (LC/MS). In all, we collected mass spectra from 1944 individual chromatographic fractions.

Subunits of many well-known complexes co-elute with distinct patterns in different types of biochemical gradient. Although we can distinguish elution patterns from different protein complexes easily, a more rigorous computational framework is required to map PPIs in the large data set such as this (Figure 1B). We employed a supervised machine learning approach based upon observed data for known complexes. Protein-protein interactions were derived solely from the separation behavior of proteins over multiple, orthogonal biochemical fractionation experiments.

5.3. VALIDATION OF KNOWN COMPLEXES THROUGH ELECTRON MICROSCOPY

Alongside the CF-MS pipeline, fractions from HPLC size exclusion of hemolysate were analyzed using electron microscopy. First, to survey the size, shape and complexity of macromolecules across the fractionation, we used negative stain EM to visualize binned fractions (Figure 2A). Using the corresponding mass spectrometry data and prior knowledge, we were able to identify four distinct protein complexes (Figure 2B). Notably, three of the structures identified were homooligomers. TPP2, a serine protease, was the largest observed homooligomer and is known to form large 5-6 MDa

assemblies (Macpherson et al., 1987; Schönegge et al., 2012). The other homooligomers identified were ALAD (Mills-Davies et al., 2017), an ~290 kDa homooctamer with D4 symmetry and PRDX2 (Schröder et al., 2000), an ~218 kDa homodecamer with D5 symmetry.

We then chose the pooled fractions 10-21 for analysis by cryo-EM. The over-represented 20S proteasome in these fractions provided a built-in control for our ability to resolve protein complexes from fractions of hemolysate. We collected ~6,600 micrographs and used single particle analysis to obtain a reconstruction of the 20S proteasome with a nominal resolution of 3.35 Å (Figure 2C). To further investigate the proteasome assembly states in red blood cells, we collected a negative stain dataset of hemolysate after being passed through a 100 kDa filter, allowing us to see multiple states in the same image (Figure 3). We found that ~94% of the proteasomes observed through this scheme were 20S and the remaining ~6% were single-capped 26S proteasomes (Figure 3). This observation corresponds with our clustering of PPIs which shows separate clustering for 20S and 19S proteasomes.

5.4. DISCUSSION

Red blood cells are one of the most abundant cell types in humans with a primary role of oxygen transport. Despite their importance, a consensus on the complete proteome as well as protein complexes underlying phenotypes and cellular functions has not been reached. This is in part due to mature erythrocytes lacking a nuclei and major organelles, preventing affinity purification mass spectrometry, proximity labeling and other genetic

tagging methods. However, RBCs are uniquely suited to investigation by CF-MS and shotgun cryo-EM as neither technique require a genetic handle.

Using a rigorous statistical framework for our mass spectrometry data, we define a comprehensive proteome for RBCs and recover high confidence protein-protein interactions from known complexes as well as novel complexes. We additionally survey the architecture of these complexes using electron microscopy on size separated red blood cell lysate and recover several known complexes as validation. An advantage of integrating electron microscopy into the CF-MS pipeline is that we are able to quantify ratios of biochemical states by direct observation, as demonstrated with the proteasome.

In future experiments, we plan to incorporate cross-link mass spectrometry data to provide additional evidence for previously uncharacterized protein-protein interactions in RBCs. Cross-link mass spectrometry will also provide useful data for integrative modeling of protein structures produced from shotgun cryo-EM of RBCs. This work details an interaction map for the protein assemblies underlying healthy red blood cells and provides a basis for molecular mechanisms leading to blood cell disorders.

5.5. METHODS

5.5.1. Negative stain electron microscopy

4 μL of hemolysate was applied to a glow-discharged 400-mesh continuous carbon grid. After allowing the sample to adsorb for 1 min, the sample was negatively stained with five consecutive droplets of 2% (w/v) uranyl acetate solution, blotted to

remove residual stain, and air-dried in a fume hood. Grids were imaged using an FEI Talos TEM (Thermo Scientific) equipped with a Ceta 16M detector. Micrographs were collected manually using TIA v4.14 software at a nominal magnification of x73,000, corresponding to a pixel size of 2.05 Å/pixel. CTF estimation, particle picking and 2D class averaging were performed using both RELION v3 (Zivanov et al., 2018) and cryoSPARC v2.12.4 (Punjani et al., 2017). Three negative stain datasets were collected. The first dataset collected contained ~220 micrographs of pooled HPLC size exclusion fractions 1-9. ~2,500 particles were manually picked and processed in cryoSPARC to produce the TPP2 structure in Figure 2C. Two datasets were collected of hemolysate after being passed through a 100 kDa filter, one as prepared and the other at 1:100 dilution. For the diluted sample, ~400 micrographs were collected and ~42,500 particles were picked using Topaz (Bepler et al., 2019). The resulting particles were used to generate the PRDX2 structure in Figure 2C and the 2D class averages in Figure 3B. For the non-dilute sample, ~230 micrographs were collected and ~1,500 proteasome particles were manually picked followed by classification in RELION (Figure 3C).

5.5.2. Cryo-EM grid preparation and data collection

C-flat holey carbon grids (CF-1.2/1.3, Protochips Inc.) were glow-discharged for 1 min using a Solarus 950 plasma cleaner (Gatan). 2 μL of 0.2 mg/mL graphene oxide (Sigma-Aldrich) was placed onto the grids for 1 min followed by one wash with water. 3 μL of pooled and concentrated hemolysate from HPLC size exclusion fractions 10-20 was placed onto the grid, blotted for 3.5 sec with a blotting force of 0, and rapidly plunged into liquid ethane using an FEI Vitrobot MarkIV operated at 4 °C and 100% humidity. Data was acquired using an FEI Titan Krios TEM (Sauer Structural Biology

Laboratory, University of Texas at Austin) operated at 300 keV with a nominal magnification of ×22,500 (1.045 Å/pixel) and defocus ranging from –1.09 to -–2.5 μm. Dose-fractionated movies were collected using 20 frames (0.15 sec/frame) over a total of 3 sec with a dose rate of ~2.13 e-/Å2/sec and a total exposure of 42.58 e-/Å2. A total of 6,606 micrographs were automatically recorded on a K3 detector (Gatan) operated in counting mode using Leginon (Suloway et al., 2005).

5.5.3. Cryo-EM data processing

Motion correction, CTF-estimation and particle picking were performed in Warpv1.0.7 (Tegunov and Cramer, 2019). Extracted particles were imported into cryoSPARC v2.12.4 for 2D classification, 3D classification and non-uniform 3D refinement. A previously solved structure of the human 20S proteasome, PDB 6RGQ (Toste Rêgo and da Fonseca, 2019), was aligned by cross-correlation in UCSF Chimera (Pettersen et al., 2004) and docked into the model.

5.6. FIGURES

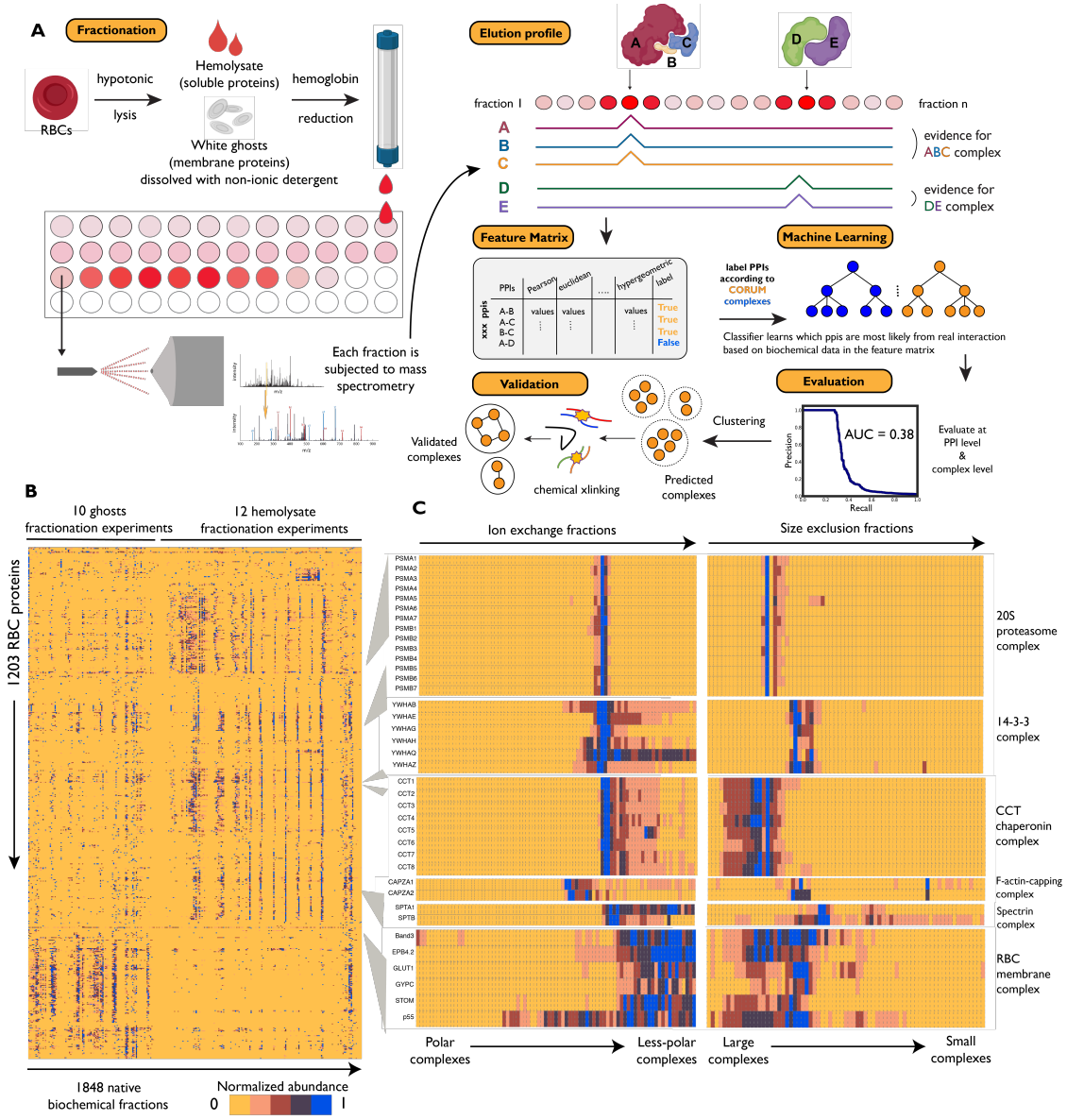

Figure 5.1. Overview of the integrative Co-Fractionation Mass Spectrometry (CF-MS) workflow used to determine stable RBC protein complexes.

(A) Hemolysate and white ghosts are chromatographically separated and the proteins in each fraction are identified by mass spectrometry. Elution profiles for each protein are represented as ridgelines across multiple separation experiments. Correlations between pairs of proteins are used to construct a feature matrix for a machine learning pipeline

which outputs a CF-MS score describing how likely an interaction between two proteins in RBCs would be. (B) Heat map of the full dataset of abundance measurements for each of the 1,202 RBC proteins across all fractionations of hemolysate and white ghosts. (C) Enlarged portions of (B) showing examples of strong co-elution observed for subunits of six well-known protein complexes in RBCs. Color intensity depicts abundances for each protein.

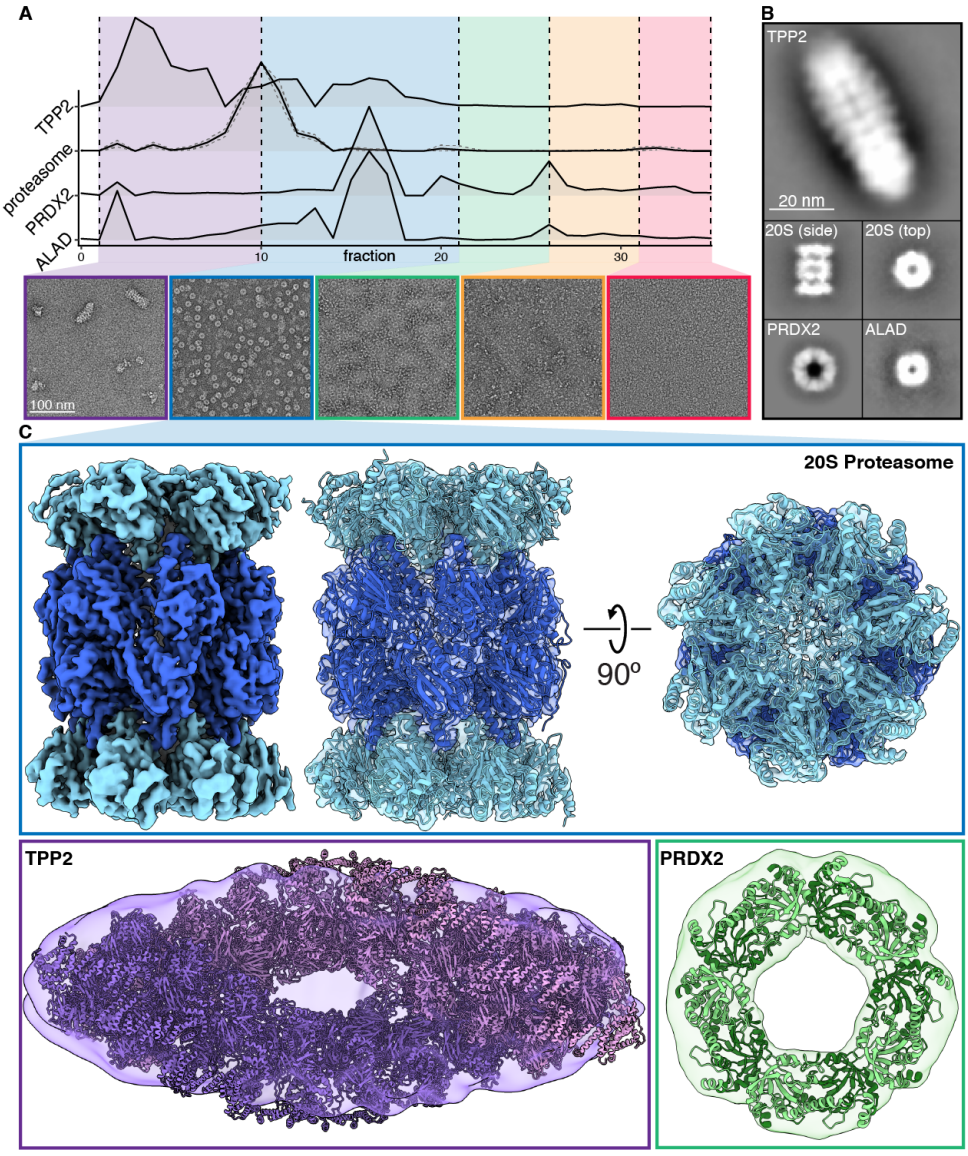

Figure 5.2. Validation of the CF-MS workflow using electron microscopy to confirm intact multi-protein complexes.

(A) Hemolysate from size exclusion chromatography was partitioned into five groups and visualized with negative stain EM. Elution profiles from corresponding mass spectrometry data were used to assist in identifying abundant protein assemblies. (B) Reference-free 2D class averages of four protein complexes spanning ~220 – 5000 kDa identified from hemolysate. (C) Cryo-EM reconstruction of the 20S proteasome and negative stain structures of TPP2 and PRDX2 along with docking of their corresponding atomic structures PDB 3LXU (Chuang et al., 2010) and PDB 1QMV (Schröder et al., 2000), respectively.

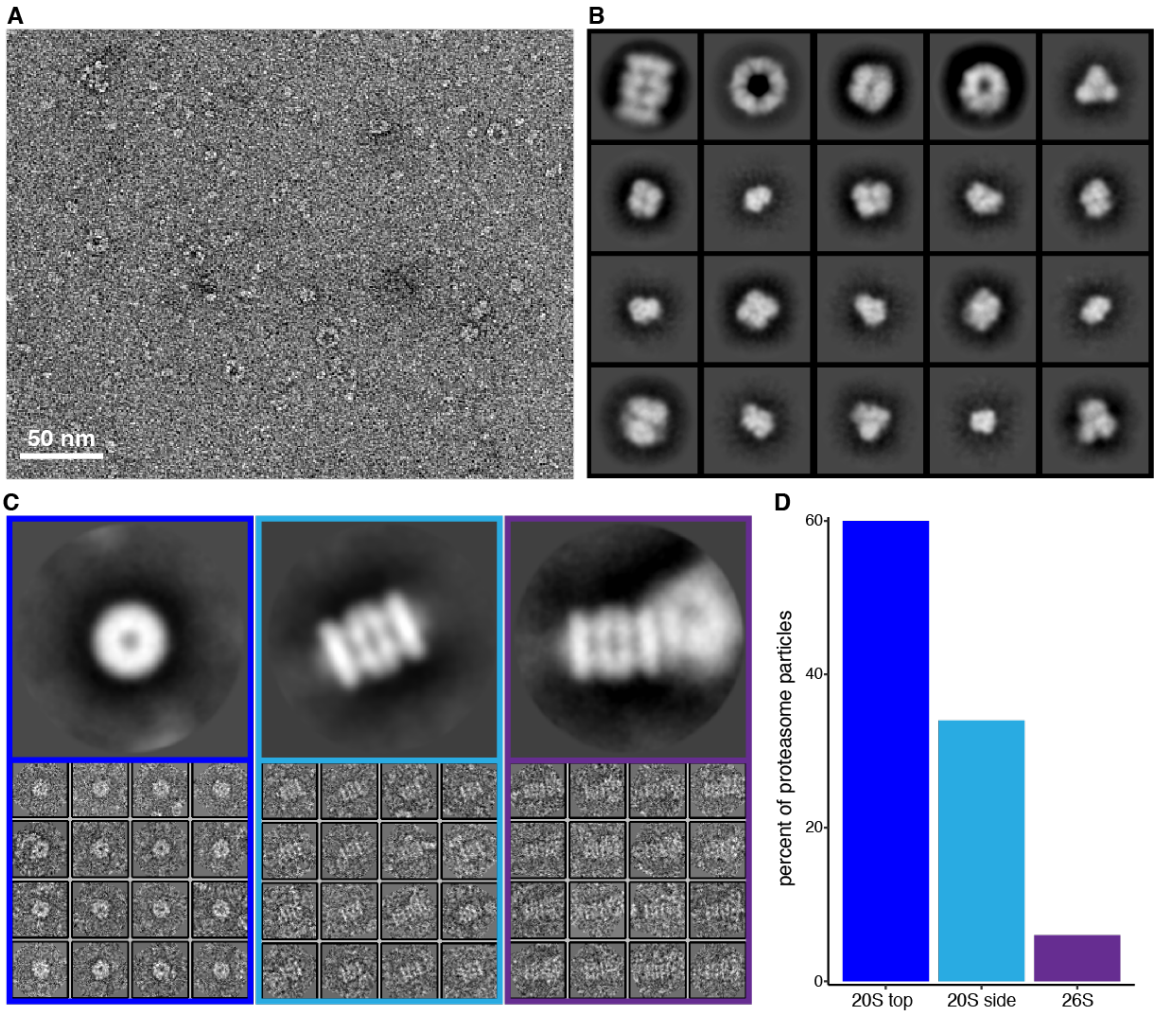

Figure 5.3. Assessment of proteasomes from negative stain electron microscopy of RBC hemolysate shows a majority in the 20S form.

(A) Example micrograph of hemolysate after being passed through a 100 kDa filter. (B) Subset of reference-free 2D class averages from filtered hemolysate showing macromolecular assemblies of distinct sizes and shapes. Box length is 254 Å. (C) Reference-free 2D class averages and aligned raw particles for 20S proteasome (top view), 20S proteasome (side view) and 26S proteasome (single-capped) from left to right. Box length is 459 Å. (D) Distribution of observed proteasome states from negative stain EM of hemolysate. The total number of proteasome particles classified was 1,510.

Chapter 6: Conclusions and Outlook

6.1. ARC OF THIS WORK

Following the adage "seeing is believing", the central theme of this dissertation has been to evaluate single particle electron microscopy as a means to visualize structures of molecular machines from cell extracts. As part of a larger quest to define the organization of the proteome into assemblies, structural studies on cell lysates can circumvent the need for purification of single targets and provide direct insight on multiple near-native protein complexes in a single experiment (Kyrilis et al., 2019; McCafferty et al., 2020). This work is made possible by many recent advances in electron microscopy (e.g. increased signal due to direct detector devices and *ab initio* 3D reconstructions), which have opened new frontiers in structural biology.

In Chapter 2 of this dissertation, we demonstrate that it is indeed possible to recover meaningful structural information using electron microscopy on fractions of size separated cell lysate containing hundreds of unique proteins. Although the results are at limited resolution and highlight only a few large, well-studied protein complexes, this proof-of-principle suggests cryo-EM need not be limited to highly purified samples. Chapter 3 is a continuation of these efforts and combines simple microfluidics with negative stain electron microscopy of single *C. elegans* embryos at differing developmental stages. This framework could provide a tractable way to study dynamics of protein structures across development. Motivated by results from the previous chapters, we then developed an algorithm designed specifically for handling single particle cryo-EM datasets containing multiple distinct structures, described in Chapter 4. Currently, the major data processing pipelines are designed for single or few related

structures. To leverage these existing software, it was necessary to design a tool for separating particles into homogenous subsets before applying conventional processing protocols. Finally, in Chapter 5, we present early results combining co-fractionation mass spectrometry and cryo-EM using red blood cells as a model system. Taken together, these chapters point towards a future of being able to combine mass spectrometry and electron microscopy to uncover meaningful and novel structures from cell lysates.

6.2. OUTLOOK FOR SHOTGUN CRYO-EM

At the beginning of this thesis work, few studies had been done explicitly using cell lysates for structural analysis of protein assemblies by electron microscopy. The first major work demonstrating the power of single particle cryo-EM and fractionation mass spectrometry was in 2017, where a 4.7 Å structure of fatty acid synthase was solved from size separated fractions of *Chaetomium thermophilum* lysate (Kastritis et al., 2017). Shortly after, we published initial progress using low-resolution negative stain EM with a focus on recovering multiple structures from cell lysate. Here, our objective was to test if cryo-EM could be integrated into the co-fractionation mass spectrometry pipeline as a way to validate and characterize known or predicted protein complexes. I anticipate that structures from these methods, in combination with electron tomography, will produce information-rich cell atlases capturing high-resolution structures of the proteome and its spatial context.

Since these initial studies, there have been a number of works published with important contributions for utilizing cryo-EM on cell extracts. One such study introduced the software cryoID, which is designed to "sequence by structure" from cryo-EM maps

(Ho et al., 2020; Terwilliger et al., 2021). This tool will be a valuable resource as more high-resolutions structures are solved from mixtures where subunits or entire assemblies could be of unknown identity. Currently, in the more likely event that structures are not solved to high-resolution, it will also be important to have methods for modelling and docking atomic structures into low- to mid-resolution maps (McCafferty et al., 2020). General machine learning approaches applied throughout the cryo-EM data processing pipeline will also increase our ability to solve structures from mixtures (Kyrilis et al., 2021a).

Samples being investigated now vary a wide range of organisms (Kim et al., 2020) and also expand beyond soluble fractions of cell lysate. So far, these include studies tackling heterogeneous mixtures from membrane fractions (Su et al., 2021), as well as nuclear extracts (Arimura et al., 2020). Other exciting avenues include targeting specific complexes by inducing cell stress (Kirykowicz and Woodward, 2020), or by combining biochemical and functional assays (Kyrilis et al., 2021b). With such rapid developments in cryo-EM technology, I am optimistic that single particle cryo-EM on native cell extracts will become an important part of uncovering many structure-function relationships.

References

Afonina, Z.A., Myasnikov, A.G., Khabibullina, N.F., Belorusova, A.Y., Menetret, J.-F., Vasiliev, V.D., Klaholz, B.P., Shirokov, V.A., and Spirin, A.S. (2013). Topology of mRNA chain in isolated eukaryotic double-row polyribosomes. Biochemistry Mosc. *78*, 445–454.

Aizenbud, Y., and Shkolnisky, Y. (2019). A max-cut approach to heterogeneity in cryo-electron microscopy. Journal of Mathematical Analysis and Applications *479*, 1004–1029.

Aloy, P. (2004). Structure-Based Assembly of Protein Complexes in Yeast. Science *303*, 2026–2029.

Arimura, Y., Shih, R.M., Froom, R., and Funabiki, H. (2020). Nucleosome structural variations in interphase and metaphase chromosomes. BioRxiv 2020.11.12.380386.

Arnold, S.A., Albiez, S., Bieri, A., Syntychaki, A., Adaixo, R., McLeod, R.A., Goldie, K.N., Stahlberg, H., and Braun, T. (2017). Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts. Journal of Structural Biology *197*, 220–226.

Asano, S., Fukuda, Y., Beck, F., Aufderheide, A., Förster, F., Danev, R., and Baumeister, W. (2015). A molecular census of 26S proteasomes in intact neurons. Science *347*, 439–442.

Baker, D. (2001). Protein Structure Prediction and Structural Genomics. Science *294*, 93–96.

Beck, M., and Baumeister, W. (2016). Cryo-Electron Tomography: Can it Reveal the Molecular Sociology of Cells in Atomic Detail? Trends in Cell Biology *26*, 825–837.

Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A.J., and Berger, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. Nat Methods *16*, 1153–1160.

Brandt, F., Etchells, S.A., Ortiz, J.O., Elcock, A.H., Hartl, F.U., and Baumeister, W. (2009). The Native 3D Organization of Bacterial Polysomes. Cell *136*, 261–271.

Chandonia, J.-M., and Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. Science *311*, 347–351.

Chuang, C.K., Rockel, B., Seyit, G., Walian, P.J., Schönegge, A.-M., Peters, J., Zwart, P.H., Baumeister, W., and Jap, B.K. (2010). Hybrid molecular structure of the giant protease tripeptidyl peptidase II. Nat Struct Mol Biol *17*, 990–996.

Cianfrocco, M.A., and Leschziner, A.E. (2015). Low cost, high performance processing of single particle cryo-electron microscopy data in the cloud. ELife *4*.

Daly, A.K. (2017). Pharmacogenetics: a general review on progress to date. British Medical Bulletin 1–15.

Danev, R., and Baumeister, W. (2016). Cryo-EM single particle analysis with the Volta phase plate. Elife *5*, e13046.

Danziger, O., Rivenzon-Segal, D., Wolf, S.G., and Horovitz, A. (2003). Conversion of the allosteric transition of GroEL from concerted to sequential by the single mutation Asp-155 -> Ala. Proceedings of the National Academy of Sciences *100*, 13797–13802.

Dickinson, A.J., Armistead, P.M., and Allbritton, N.L. (2013). Automated Capillary Electrophoresis System for Fast Single-Cell Analysis. Analytical Chemistry *85*, 4797–4804.

Dickinson, D.J., Schwager, F., Pintard, L., Gotta, M., and Goldstein, B. (2017). A Single-Cell Biochemistry Approach Reveals PAR Complex Dynamics during Cell Polarization. Developmental Cell *42*, 416-434.e11.

Doerr, A. (2018). Taking inventory with shotgun EM. Nature Methods *15*, 649–649.

Drew, K., Lee, C., Huizar, R.L., Tu, F., Borgeson, B., McWhite, C.D., Ma, Y., Wallingford, J.B., and Marcotte, E.M. (2017a). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology *13*, 932.

Drew, K., Müller, C.L., Bonneau, R., and Marcotte, E.M. (2017b). Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. PLOS Computational Biology *13*, e1005625.

Edgar, L.G., Wolf, N., and Wood, W.B. (1994). Early transcription in Caenorhabditis elegans embryos. Development *120*, 443–451.

Finley, D. (2009). Recognition and Processing of Ubiquitin-Protein Conjugates by the Proteasome. Annual Review of Biochemistry *78*, 477–513.

Flemming, D., Thierbach, K., Stelter, P., Böttcher, B., and Hurt, E. (2010). Precise mapping of subunits in multiprotein complexes by a versatile electron microscopy label. Nature Structural & Molecular Biology *17*, 775–778.

da Fonseca, P.C.A., and Morris, E.P. (2015). Cryo-EM reveals the conformation of a substrate analogue in the human 20S proteasome core. Nature Communications *6*, 7573.

Galaz-Montoya, J.G., and Ludtke, S.J. (2017). The advent of structural biology in situ by single particle cryo-electron tomography. Biophysics Reports *3*, 17–35.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., and Cruciat, C.-M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775–1787.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., et al. (2009). Transcriptome Complexity in a Genome-Reduced Bacterium. Science *326*, 1268–1271.

Harshbarger, W., Miller, C., Diedrich, C., and Sacchettini, J. (2015). Crystal Structure of the Human 20S Proteasome in Complex with Carfilzomib. Structure *23*, 418–424.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol *18*, 83.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., et al. (2012). A Census of Human Soluble Protein Complexes. Cell *150*, 1068–1081.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F., et al. (2015). A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell *163*, 712–723.

Herman, G.T., and Kalinowski, M. (2008). Classification of heterogeneous electron microscopic projections into homogeneous subsets. Ultramicroscopy *108*, 327–338.

Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Research *19*, 657–666.

Ho, C.-M., Li, X., Lai, M., Terwilliger, T.C., Beck, J.R., Wohlschlegel, J., Goldberg, D.E., Fitzpatrick, A.W.P., and Zhou, Z.H. (2020). Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. Nat Methods *17*, 79–85.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

Huang, B., Wu, H., Bhaya, D., Grossman, A., Granier, S., Kobilka, B.K., and Zare, R.N. (2007). Counting Low-Copy Number Proteins in a Single Cell. Science *315*, 81–84.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425–440.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505–509.

Jones, P.A., and Baylin, S.B. (2007). The Epigenomics of Cancer. Cell *128*, 683–692.

Joyce, A.R., and Palsson, B.Ø. (2006). The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol *7*, 198–210.

Karczewski, K.J., and Snyder, M.P. (2018). Integrative omics for health and disease. Nat Rev Genet *19*, 299–310.

Kastritis, P.L., O'Reilly, F.J., Bock, T., Li, Y., Rogon, M.Z., Buczak, K., Romanov, N., Betts, M.J., Bui, K.H., Hagen, W.J., et al. (2017). Capturing protein communities by structural proteomics in a thermophilic eukaryote. Molecular Systems Biology *13*, 936.

Katsevich, E., Katsevich, A., and Singer, A. (2015). Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem. SIAM Journal on Imaging Sciences *8*, 126–185.

Kemmerling, S., Arnold, S.A., Bircher, B.A., Sauter, N., Escobedo, C., Dernick, G., Hierlemann, A., Stahlberg, H., and Braun, T. (2013). Single-cell lysis for visual analysis by electron microscopy. Journal of Structural Biology *183*, 467–473.

Kim, S.-H. (1998). Shining a light on structural genomics. Nat Struct Mol Biol *5*, 643–645.

Kim, G., Jang, S., Lee, E., and Song, J.-J. (2020). EMPAS: Electron Microscopy Screening for Endogenous Protein Architectures. Mol Cells *43*, 804–812.

Kirykowicz, A.M., and Woodward, J.D. (2020). Shotgun EM of mycobacterial protein complexes during stationary phase stress. Current Research in Structural Biology *2*, 204–212.

Kovarik, M.L., and Allbritton, N.L. (2011). Measuring enzyme activity in single cells. Trends Biotechnol. *29*, 222–230.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nature Methods *9*, 907–909.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Kühlbrandt, W. (2014). The resolution revolution. Science *343*, 1443–1444.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome Organization in a Genome-Reduced Bacterium. Science *326*, 1235–1240.

Kumar, P., Tan, Y., and Cahan, P. (2017). Understanding development and stem cells using single cell-based analyses of gene expression. Development *144*, 17–32.

Kyrilis, F.L., Meister, A., and Kastritis, P.L. (2019). Integrative biology of native cell extracts: a new era for structural characterization of life processes. Biological Chemistry *400*, 831–846.

Kyrilis, F.L., Belapure, J., and Kastritis, P.L. (2021a). Detecting Protein Communities in Native Cell Extracts by Machine Learning: A Structural Biologist's Perspective. Front. Mol. Biosci. *8*, 660542.

Kyrilis, F.L., Semchonok, D.A., Skalidis, I., Tüting, C., Hamdi, F., O'Reilly, F.J., Rappsilber, J., and Kastritis, P.L. (2021b). Integrative structure of a 10-megadalton eukaryotic pyruvate dehydrogenase complex from native cell extracts. Cell Reports *34*, 108727.

Lander, G.C., Stagg, S.M., Voss, N.R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., et al. (2009). Appion: An integrated, database-driven pipeline to facilitate EM image processing. Journal of Structural Biology *166*, 95–102.

Lander, G.C., Estrin, E., Matyskiela, M.E., Bashore, C., Nogales, E., and Martin, A. (2012). Complete subunit architecture of the proteasome regulatory particle. Nature.

Larance, M., Kirkwood, K.J., Tinti, M., Brenes Murillo, A., Ferguson, M.A.J., and Lamond, A.I. (2016). Global Membrane Protein Interactome Analysis using *In vivo* Crosslinking and Mass Spectrometry-based Protein Correlation Profiling. Molecular & Cellular Proteomics *15*, 2476–2490.

Leitner, A., Faini, M., Stengel, F., and Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. Trends in Biochemical Sciences *41*, 20–32.

Li, Y., Cash, J.N., Tesmer, J.J.G., and Cianfrocco, M.A. (2020). High-Throughput Cryo-EM Enabled by User-Free Preprocessing Routines. Structure *28*, 858-869.e3.

Liao, H.Y., Hashem, Y., and Frank, J. (2015). Efficient Estimation of Three-Dimensional Covariance and its Application in the Analysis of Heterogeneous Samples in Cryo-Electron Microscopy. Structure *23*, 1129–1137.

Liu, F., and Heck, A.J. (2015). Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. Current Opinion in Structural Biology *35*, 100–108.

Lučić, V., Förster, F., and Baumeister, W. (2005). STRUCTURAL STUDIES BY ELECTRON TOMOGRAPHY: From Cells to Molecules. Annu. Rev. Biochem. *74*, 833–865.

Luck, K., Sheynkman, G.M., Zhang, I., and Vidal, M. (2017). Proteome-Scale Human Interactomics. Trends in Biochemical Sciences *42*, 342–354.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: Semiautomated Software for High-Resolution Single-Particle Reconstructions. Journal of Structural Biology *128*, 82–97.

Macpherson, E., Tomkinson, B., Bålöw, R.M., Höglund, S., and Zetterqvist, O. (1987). Supramolecular structure of tripeptidyl peptidase II from human erythrocytes as studied by electron microscopy, and its correlation to enzyme activity. Biochemical Journal *248*, 259–263.

Malyutin, A.G., Musalgaonkar, S., Patchett, S., Frank, J., and Johnson, A.W. (2017). Nmd3 is a structural mimic of eIF5A, and activates the cpGTPase Lsg1 during 60S ribosome biogenesis. The EMBO Journal *36*, 854–868.

Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. Journal of Structural Biology *152*, 36–51.

McCafferty, C.L., Verbeke, E.J., Marcotte, E.M., and Taylor, D.W. (2020). Structural Biology in the Multi-Omics Era. J. Chem. Inf. Model. *60*, 2424–2429.

Mills-Davies, N., Butler, D., Norton, E., Thompson, D., Sarwar, M., Guo, J., Gill, R., Azim, N., Coker, A., Wood, S.P., et al. (2017). Structural studies of substrate and product complexes of 5-aminolaevulinic acid dehydratase from humans, *Escherichia coli* and the hyperthermophile *Pyrobaculum calidifontis*. Acta Crystallogr D Struct Biol *73*, 9–21.

Mirande, M. (2017). The Aminoacyl-tRNA Synthetase Complex. In Macromolecular Protein Complexes, J.R. Harris, and J. Marles-Wright, eds. (Cham: Springer International Publishing), pp. 505–522.

Montelione, G.T. (2012). The Protein Structure Initiative: achievements and visions for the future. F1000 Biology Reports *4*.

Newman, M.E.J., and Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E *69*.

Nisemblat, S., Yaniv, O., Parnas, A., Frolow, F., and Azem, A. (2015). Crystal structure of the human mitochondrial chaperonin symmetrical football complex. Proceedings of the National Academy of Sciences *112*, 6044–6049.

Penczek, P.A., Grassucci, R.A., and Frank, J. (1994). The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. Ultramicroscopy *53*, 251–270.

Penczek, P.A., Frank, J., and Spahn, C.M.T. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. Journal of Structural Biology *154*, 184–194.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry *25*, 1605–1612.

Pons, P., and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In Computer and Information Sciences - ISCIS 2005, Pinar Yolum, T. Güngör, F. Gürgen, and C. Özturan, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 284–293.

Potter, S.S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol *14*, 479–492.

Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods *14*, 290–296.

Rappsilber, J., Siniossoglou, S., Hurt, E.C., and Mann, M. (2000). A Generic Strategy To Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry. Analytical Chemistry *72*, 267–275.

Raupach, M.J., Amann, R., Wheeler, Q.D., and Roos, C. (2016). The application of "-omics" technologies for the classification and identification of animals. Org Divers Evol *16*, 1–12.

Rickgauer, J.P., Grigorieff, N., and Denk, W. (2017). Single-protein detection in crowded molecular environments in cryo-EM images. ELife *6*.

Riekeberg, E., and Powers, R. (2017). New frontiers in metabolomics: from measurement to insight. F1000Res *6*, 1148.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. Journal of Structural Biology *192*, 216–221.

Roseman, A. (2004). FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. Journal of Structural Biology *145*, 91–99.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Research *38*, D497–D501.

Russo, C.J., and Passmore, L.A. (2014). Ultrastable gold substrates for electron cryomicroscopy. Science *346*, 1377–1380.

Scaiola, A., Peña, C., Weisser, M., Böhringer, D., Leibundgut, M., Klingauf-Nerurkar, P., Gerhardy, S., Panse, V.G., and Ban, N. (2018). Structure of a eukaryotic cytoplasmic pre-40S ribosomal subunit. The EMBO Journal 13.

Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology *180*, 519–530.

Schönegge, A.-M., Villa, E., Förster, F., Hegerl, R., Peters, J., Baumeister, W., and Rockel, B. (2012). The Structure of Human Tripeptidyl Peptidase II as Determined by a Hybrid Approach. Structure *20*, 593–603.

Schorb, M., and Briggs, J.A.G. (2014). Correlated cryo-fluorescence and cryo-electron microscopy with high spatial precision and improved sensitivity. Ultramicroscopy *143*, 24–32.

Schröder, E., Littlechil*, J.A., Lebedev, A.A., Errington, N., Vagin, A.A., and Isupov, M.N. (2000). Crystal structure of decameric 2-Cys peroxiredoxin from human erythrocytes at 1.7Å resolution. Structure *8*, 605–615.

Schweitzer, A., Aufderheide, A., Rudack, T., Beck, F., Pfeifer, G., Plitzko, J.M., Sakata, E., Schulten, K., Förster, F., and Baumeister, W. (2016). Structure of the human 26S proteasome at a resolution of 3.9 Å. Proceedings of the National Academy of Sciences *113*, 7816–7821.

Seydoux, G., and Fire, A. (1994). Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. Development *120*, 2823–2834.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for

integrated models of biomolecular interaction networks. Genome Research *13*, 2498–2504.

Shatsky, M., Hall, R.J., Nogales, E., Malik, J., and Brenner, S.E. (2010). Automated multi-model reconstruction from single-particle electron microscopy data. Journal of Structural Biology *170*, 98–108.

Shaw, D., Wang, S.M., Villaseñor, A.G., Tsing, S., Walter, D., Browner, M.F., Barnett, J., and Kuglstatter, A. (2008). The Crystal Structure of JNK2 Reveals Conformational Flexibility in the MAP Kinase Insert and Indicates Its Involvement in the Regulation of Catalytic Activity. Journal of Molecular Biology *383*, 885–893.

Shen, P.S., Park, J., Qin, Y., Li, X., Parsawar, K., Larson, M.H., Cox, J., Cheng, Y., Lambowitz, A.M., Weissman, J.S., et al. (2015). Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains. Science *347*, 75–78.

Sibanda, B.L., Chirgadze, D.Y., Ascher, D.B., and Blundell, T.L. (2017). DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science *355*, 520–524.

Sigworth, F.J. (1998). A Maximum-Likelihood Approach to Single-Particle Image Refinement. Journal of Structural Biology *122*, 328–339.

Sigworth, F.J., Doerschuk, P.C., Carazo, J.-M., and Scheres, S.H.W. (2010). An Introduction to Maximum-Likelihood Methods in Cryo-EM. In Methods in Enzymology, (Elsevier), pp. 263–294.

Silva, J.C., Gorenstein, M.V., Li, G.-Z., Vissers, J.P.C., and Geromanos, S.J. (2006). Absolute Quantification of Proteins by LCMS ^{E}: A Virtue of Parallel ms Acquisition. Molecular & Cellular Proteomics *5*, 144–156.

Skolnick, J., Fetrow, J.S., and Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. Nat Biotechnol *18*, 283–287.

Slavov, N., Semrau, S., Airoldi, E., Budnik, B., and van Oudenaarden, A. (2015). Differential Stoichiometry among Core Ribosomal Proteins. Cell Reports *13*, 865–873.

Sone, T., Saeki, Y., Toh-e, A., and Yokosawa, H. (2004). Sem1p Is a Novel Subunit of the 26 S Proteasome from *Saccharomyces cerevisiae*. Journal of Biological Chemistry *279*, 28807–28816.

Stevens, R.C. (2001). Global Efforts in Structural Genomics. Science *294*, 89–92.

Su, C.-C., Lyu, M., Morgan, C.E., Bolla, J.R., Robinson, C.V., and Yu, E.W. (2021). A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. Nat Methods *18*, 69–75.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., and Carragher, B. (2005). Automated molecular microscopy: The new Leginon system. Journal of Structural Biology *151*, 41–60.

Tegunov, D., and Cramer, P. (2019). Real-time cryo-electron microscopy data preprocessing with Warp. Nat Methods *16*, 1146–1152.

Terwilliger, T.C., Sobolev, O.V., Afonine, P.V., Adams, P.D., Ho, C.-M., Li, X., and Zhou, Z.H. (2021). Protein identification from electron cryomicroscopy maps by automated model building and side-chain matching. Acta Crystallogr D Struct Biol *77*, 457–462.

Toste Rêgo, A., and da Fonseca, P.C.A. (2019). Characterization of Fully Recombinant Human 20S and 20S-PA200 Proteasome Complexes. Molecular Cell *76*, 138-147.e5.

Vakser, I.A. (2014). Protein-Protein Docking: From Interaction to Interactome. Biophysical Journal *107*, 1785–1793.

Van Heel, M. (1987). Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy *21*, 111–123.

Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A., and Martens, L. (2011). SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. PROTEOMICS *11*, 996–999.

Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., Berven, F.S., Sickmann, A., Martens, L., and Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nature Biotechnology *33*, 22–24.

Verbeke, E.J., Mallam, A.L., Drew, K., Marcotte, E.M., and Taylor, D.W. (2018). Classification of Single Particles from Human Cell Extract Reveals Distinct Structures. Cell Reports *24*, 259-268.e3.

Voss, N.R., Yoshioka, C.K., Radermacher, M., Potter, C.S., and Carragher, B. (2009). DoG Picker and TiltPicker: Software tools to facilitate particle selection in single particle electron microscopy. Journal of Structural Biology *166*, 205–213.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., et al. (2015). Panorama of ancient metazoan macromolecular complexes. Nature *525*, 339–344.

Wang, F., Gong, H., Liu, G., Li, M., Yan, C., Xia, T., Li, X., and Zeng, J. (2016). DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. Journal of Structural Biology *195*, 325–336.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57–63.

Yi, X., Verbeke, E.J., Chang, Y., Dickinson, D.J., and Taylor, D.W. (2019). Electron microscopy snapshots of single particles from single cells. J. Biol. Chem. *294*, 1602–1608.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., Wodke, J.A.H., Güell, M., Martínez, S., Bourgeois, R., et al. (2009). Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. Science *326*, 1263–1268.

Zhan, X., Kook, S., Kaoud, T.S., Dalby, K.N., Gurevich, E.V., and Gurevich, V.V. (2015). Arrestin-3-Dependent Activation of c-Jun N-Terminal Kinases (JNKs). Curr Protoc Pharmacol *68*, 2.12.1-2.12.26.

Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat Methods *14*, 331–332.

Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. ELife *7*, e42166.